The ecology of human linguistic groups

by

José A. Capitán and Susanna Manrubia

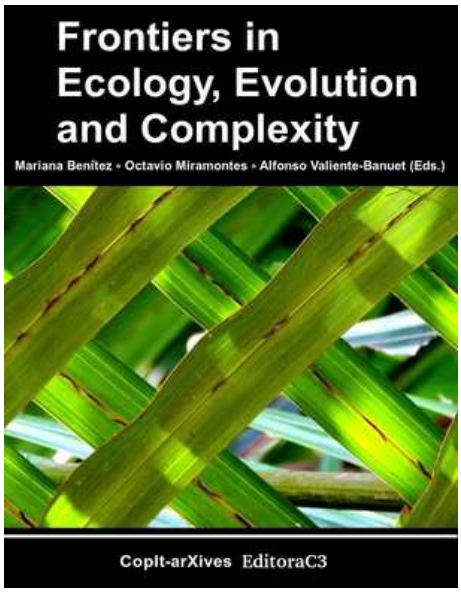
Centro de Astrobiología, CSIC, Madrid, Spain

This is a separate chapter from the open access book

Frontiers in Ecology, Evolution and Complexity

Mariana Benítez, Octavio Miramontes & Alfonso Valiente-Banuet (Editors)
CopIt-arXives, 2014
Mexico City
ISBN: 978-1-938128-05-9

http://scifunam.fisica.unam.mx/mir/copit/TS0012EN/TS0012EN.html



Contents

The ecology of human linguistic groups

J.A. Capitán and S. Manrubia, Centro de Astrobiología, CSIC-INTA, Madrid, Spain

1 Abstract

Similarities between linguistic and biological diversity were identified long ago. As research on both fields has advanced, qualitative parallelisms have turned into quantitatively comparable patterns. Remarkable examples are the statistical properties of taxonomy, the decline of diversity with latitude, or the allometric relationship between population abundaces and range sizes. Though multiple factors may underlie these remarkable patterns, the similarites uncovered between linguistic and biological diversity point to a relevant role of environment in shaping them. Eventually, the study of a human macroecology may lead to the discovery of generic mechanisms behind the evolution and interaction of populations.

2 Resumen

Hace mucho tiempo que ciertas semejanzas entre la diversidad lingüística y la diversidad biológica fueron identificadas. A medida que la investigación en ambos campos ha avanzado, lo que en principio fueron paralelismos cualitativos se han convertido en patrones cuantitativamente comparables. Algunos ejemplos destacados son las propiedades estadísticas de la taxonomía, la disminución de la diversidad con la latitud o la relación alométrica entre la abundancia y el área ocupada por una población. Aunque son múltiples los factores que subyacen a estos patrones, las semejanzas entre la diversidad lingüística y la biológica sugieren que el ambiente debe desempeñar un papel relevante en su emergencia. Finalmente, el estudio de una macroecología humana puede llevar al hallazgo de mecanismos genéricos tras la evolución e interacción de poblaciones.

3 Introduction

One of the strongest evidences for evolution is the observation of resemblances between separated entities, since similarity may speak for shared ancestry. As early as in the sixteenth century, it was independently proposed that species, as well as languages, presented intriguing commonalities that were far from trivial. At the time, the first European visitors of Asiatic regions noted similarities between Indian, Iranian, and European languages, while on the biological side systematic comparisons between the anatomy of organisms began to be carried out.

The hypothesis that linguistic similarities could be due to a common origin was put forward in the eighteenth century. In 1786, Sir William Jones, founder of the Asiatic Society of Calcutta, demonstrated the presence of fundamental similarities among Latin, Greek, Persian, Sanskrit, and, with less confidence, Celtic languages and Gothic. In his view, these similarities could only be explained if those languages arose from a common ancestor through descent with modification. Later, that ancestral language became known as Proto-Indo-European. Sir William Jones settled the basis for what is nowadays termed comparative linguistics and introduced important elements of evolution in linguistics – without natural selection, which is not applicable to languages.

Comparative linguistics had its biological counterpart in comparative anatomy, a discipline that, after the pioneering work of Edward Tyson on mammals, became established also in the eighteenth century. Studies carried out by anatomists like George Cuvier, Richard Owen or Thomas Henry Huxley represented a breakthrough in our understanding of the relatedness among vertebrates. Comparative anatomy and embryology have been the major tools to understand phylogeny until quite recently, when they have been complemented and even displaced by genomic knowledge. Though techniques other than comparative linguistics are currently used to establish the relatedness of languages, a revolution tantamount to that brought by sequencing techniques has not been produced in linguistics.

Studies on the origin of languages were severely impeded shortly after the publication of Darwin's book *On the Origin of Species by Natural Selection*, at a time when evolution was becoming a most fashionable concept. Actually, most absurd theories on the origin of language and on the nature of the "primitive language" were sprouting like weeds to the point that, in 1866, the Linguistic Society of Paris included in its founding statutes the following statement: "The Society does not accept papers on either the origin of language or the invention of a universal language". This scholarly disapproval continued well into the twentieth century, when advances in human evolution and comparisons between human and animal communication systems turned the origin of language into a respectful topic [1].

At present, the analogies between biological and linguistic evolution are much deeper than previously suspected [2], and relevant to the point that some models of evolution are applicable to both systems. An interesting advance has been to realize that many

features of the distribution of biological populations can arise in neutral scenarios, where selection plays an insignificant role. In many respects, a human language is equivalent to a biological species, and this similarity applies to qualitative as well as quantitative aspects. The comparison of those two evolutionary systems has been mediated by the ever increasing amount of data describing both biodiversity and languages. Ecology has a long tradition of cataloging species, their locations and their interactions. Nowadays, information can be easily downloaded by any interested user from databases such as The Global Biodiversity Information Facility[1] or the Web of Life[2]. As for languages, detailed information can be obtained from The Linguasphere Register[3] or from The Ethnologue[4] which is the most comprehensive catalog to date, with information on over 7,000 living languages. All these databases are being continuously amended and enlarged, and their reliability depends on the work and criteria of expert ecologists and linguists. Though these extensive datasets might contain errors that should affect predictions at the level of specific species or languages, the overall, statistical patterns that we are going to discuss should not be qualitatively affected by present mistakes or future improvements.

4 Linguistic and biological taxonomy

The Indoeuropean family of languages is formed by several hundred related languages, about half of them now extinct. It was the first linguistic family to be recognized and accepted, at the beginning of the nineteenth century. The identification of other major families was more difficult and not devoid of controversy. Between 1940 and 1960 Joseph Greenberg made significant progress when he convincingly demonstrated that about two thousand aboriginal African languages could be grouped into only four families. Towards the end of the 1980's, Merritt Ruhlen, one of his disciples, suggested that all human languages can be grouped together, a claim that implied the existence of an ancestral language from which seventeen families, in his classification, should have branched [3].

The current classification of languages into families is congruent with knowledge gathered from anthropology and genomics. That is to say, when two populations are close from a genomic viewpoint, they tend to speak languages belonging to the same family. The tree that compares linguistic families and genetic similarity is coherent in this respect with three exceptions: Lapps, Ethiopians, and Tibetans [4]. Congruence of the two data sets, but differences as well contribute to disentangle the patterns of divergence and dispersion of human populations. Interestingly, new independent data are continuously added to those studies, as in an investigation where linguistic phylogeny was complemented with the genetic analysis of human gastric bacterial parasites, leading to a reliable

[1] http://www.gbif.org
[2] http://www.web-of-life.es
[3] http://www.linguasphere.info
[4] https://www.ethnologue.com

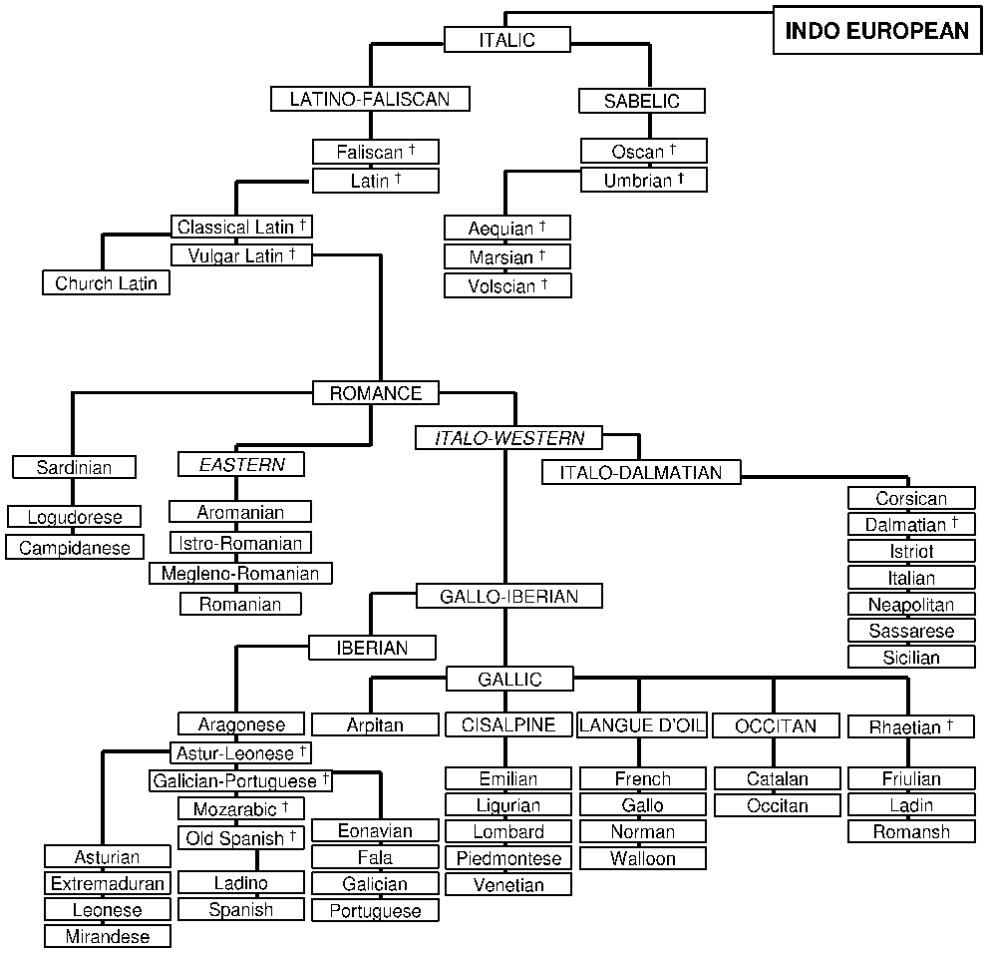


Figure 1: Languages can be hierarchically grouped in taxonomic levels, as it is done for species. In this example we observe several taxonomic levels that link present languages (leaves of the tree) in the Indo-European family to a hypothetical ancestor through a significant number of now extinct languages (indicated with a † sign). This is a partial tree that only represents the Italic group. The complete Indo-European family can be found in open places such as the Wikipedia.

reconstruction of Pacific population history [5]. This is an extreme (and rare to date) example of how similar biological and linguistic phylogenies might be.

The classification of languages includes a variable number of taxonomic levels in addition to that of family. Particularly rich linguistic groups, as that of Bantu, in Africa, may entail up to seventeen hierarchical levels. These levels are conceptually similar to biological taxa in that new similarities among groups of related languages appear every time we go down one level in the taxonomy (languages occupy the lowest level). As it happens with hierarchical groupings of species, the branching of languages in their reconstructed phylogenies is highly uneven: most groups are small, while a few are composed of many languages, the pattern repeating as one climbs up taxonomic levels. This was one of the first observations regarding the quantitative properties of biological phylogeny. The processes behind such regularities are thought to be of multiplicative nature, analogous to branching processes.

In this kind of processes, the essential mechanism is the branching of a variable number of subtaxa from a given taxon, independently of the taxonomic level. In simple representations of the process, the probability that the taxon has no subtaxa, or 1, 2, 3 or more "daughter" branches is assumed to be independent of the "parental" taxon. Figure 1 contains several cases of branching. For instance, Gallo-Iberian is the ancestor of a single taxon one level below, Iberian, from which Aragonese, Astur-Leonese, Galician-Portuguese, Mozarabic, and Old Spanish, branched. In its turn, Old Spanish splitted into Ladino and Spanish. The first model of this kind aimed at explaining the structure of biological taxonomy was proposed in 1924 [6]. Much later, the statistical properties of the classification of human languages were analyzed [7] to reveal that the distribution of the number of subtaxa within a given taxon follows a power-law distribution, with an exponent that increases in absolute value with the taxonomic level. This scaling is fully analogous to the self-similarity that had been described about ten years earlier for biological taxonomy [8]. The invariance of the functional form describing both systems supports its robustness against different possible classification schemes that coherently assign subtaxa to the taxon from where they originated, and plausibly establishes its emergence from an underlying stochastic branching process.

Linguistic phylogenies reflect highly contingent historical processes of language change, diversification, and extinction. Several such processes are known, though the time scales involved and the depth of the modifications caused are not easy to quantify. Words modify their prevalence in a population through time, change their meaning, are borrowed from other languages, or disappear when speakers stop using them. It suffices to pay attention to different regions where the mother tongue of any of us is spoken to realize how often names of plants or food change, and how particular idioms characterize subpopulations of speakers. These modifications at the local scale do not alter languages in any major way and resemble minor, neutral mutations in genotypes. More severe changes have occurred historically and can be identified in languages with a written record. An example is English, which incorporated a huge amount of lexicon and some grammar

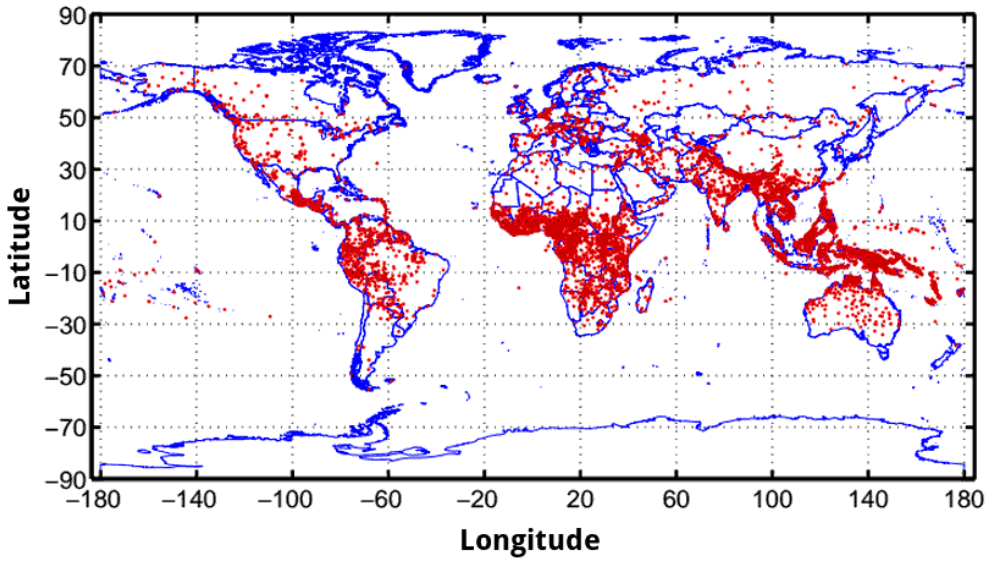


Figure 2: Distribution of linguistic diversity as compiled in the Ethnologue. Each point represents the centroid of the area covered by each language, as reported in the database. As it happens with biodiversity, linguistic diversity diminishes with increasing latitude.

from different languages in successive waves, as from Norman French or later from Latin. These processes are reminiscent of what is known as horizontal gene transfer in biology. A more dramatic influence of one language over another is the case of Creoles, full fledged natural languages that emerge from two parent languages in a time as short as two generations. Haitian Creole has been described as a West African language with French words, since it took the grammar from the former and the lexicon from the latter [9]. Cases as this one are, now metaphorically speaking, evocative of hybridization or genomic admixtures, where the two "parent" languages contribute in similar amounts to the emerging language, or of symbiotic associations, where one language provides the structure for interactions (e.g. the grammar) and another one the molecular elements (the lexicon).

5 Diversity and latitude

The spatial distribution of species over the Earth's surface develops several regularities that are far from trivial. Among them, the most prominent pattern relates biological diversity and latitude. Ecological communities in the tropics are fundamentally more diverse, and biodiversity declines as latitude increases. Though this observation was already known at Darwin's time, we still lack a convincing explanation of why ecological communities are more diverse near the equator [10].

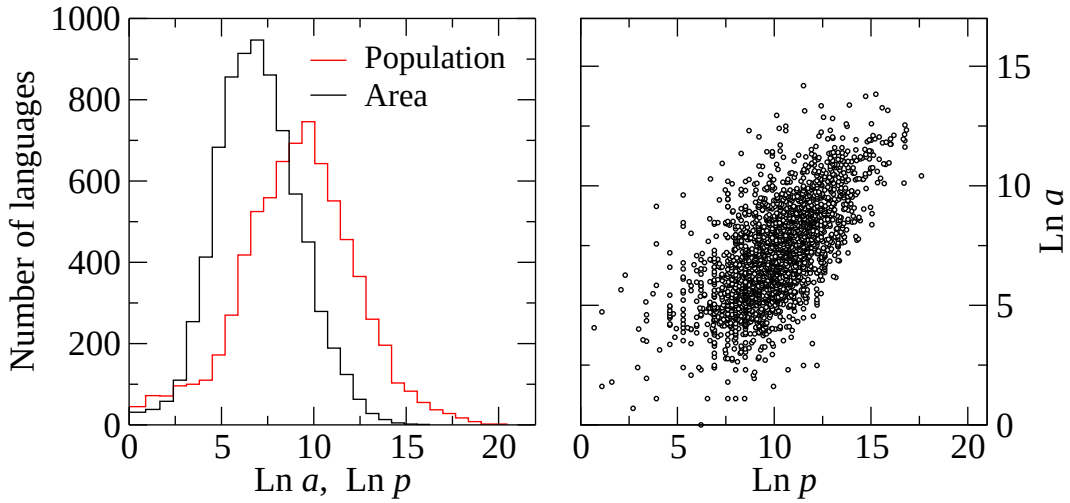


Figure 3: Left: Histograms of the number of linguistic groups occupying an area a or formed by p speakers. Note the Gaussian shape of the distributions, which can be well fitted by log-normal functions since the $x-$axis is the logarithm of the relevant variable. Right: Correlation between the number of speakers of a language (population) and the area over which they spread. The plot contains 2314 African languages. In this case, $z = 0.94$. Modified from [15].

In studying mammals, Eduardo Rapoport observed that home ranges, that is, the area spanned by a given species, were generally smaller at lower latitudes [11]. One may conclude that narrower ranges at lower latitudes would facilitate the coexistence of a larger number of species, and this may provide a partial explanation for this pattern. But it was later shown that there are many exceptions to this rule, which seems to be applicable only to high latitudes and for a subset of the species living there. Thus, it has been argued that the rule simply describes a local phenomenon, and that it can not be used to explain the latitudinal decline of biodiversity [12]. It has been put forward [13] that the latitudinal pattern of biodiversity could be a simple, statistical consequence of the wide distribution of species ranges via the so-called mid-range effect, which means that if species within a bounded geographical domain were randomly shuffled, their ranges would overlap towards the center of the domain. Another hypothesis stresses that ecological phenomena, such as climatic variability, act as selection pressures driving species to acquire high climatic tolerances, thus favoring adaptation to wider latitudinal ranges [14].

Setting aside the multiple mechanisms devised for explaining the decline of diversity with latitude, the same pattern is found in the distribution of human linguistic domains [16]. Figure 2 shows the geographical distribution of linguistic diversity. Apparently, language richness concentrates in a latitudinal band at both sides of the equator.

Is this pattern caused by the same phenomena that determine the geographical distribution of species richness? We do not yet have a fully conclusive answer to this puzzling question, though we might guess that the dominant processes should not rely solely on strictly ecological or cultural factors: whatever determines the diversity-latitude pattern should be affecting species and human linguistic groups in a similar manner. The role of the physical environment in determining this pattern might be essential, despite the unsolved controversy on its precise origin. It has been demonstrated that up to 80% of the linguistic diversity measured in relatively small regions of 200×200 km^2 can be explained on the basis of few environmental variables, among which river density and landscape roughness are those with the higher explanatory power [17]. It is likely that a better understanding of the commonalities between biodiversity and linguistic diversity can discriminate between candidate mechanisms to explain the latitude-diversity pattern.

6 Population abundance and range sizes

There are two important quantities revealing the nature of population dynamics. These are the home range and the abundance of individuals within a given group. The distribution of these two quantities, and their mutual dependence –leading to what can be called a population-area relationship– are essential signatures to unveil relevant mechanisms shaping large-scale diversity patterns.

The probability that a language is spoken by a certain number of humans follows a log-normal distribution [18]. This pattern can be easily explained on the basis of demographic dynamics. An important assumption in this context is that linguistic change can be essentially discarded [19], since demography is the dominant process on historically short time scales (several centuries). A log-normal distribution of abundances has been also reported for some biological groups, as birds and insects [20], though this pattern is not universal in biology. Regarding the abundance distributions of species and languages, we may confidently state that the same pattern speaks for the same process, which in this case is the intrinsic dynamics of population growth. If demography can be represented as the result of a certain (variable) growth rate at each year or generation – that is, the population one step later is the original one times its growth rate that year or that generation–, then the abundance distribution takes a log-normal shape. When other processes affect demography (for instance shortage of resources or space limitation), the distribution might change.

The number of individuals a species hosts has been correlated with the size of the range it spans, yielding an allometric (power-law) relationship in which the exponent varies along different taxa and habitats [21]. Several mechanisms have been proposed to explain this scaling, such as self-similarity in the spatial distribution of individuals [22] or stochasticity [23]. As of today there is no agreement on the mechanism that explain the species abundance-species range relationship. Interestingly enough, a similar functional

dependence has been described for human linguistic groups [15]. Since both variables, language size p and area a, are log-normally distributed, it is natural to assume a mutual dependence with the same functional form as observed for species abundances and ranges, $a \sim p^z$. This relationship is indeed observed for all languages reported in the Ethnologue database, as well as for the languages spoken in the five largest continental landmasses separately (Europe, Asia, North and South America, and Africa). Different continental regions are characterized by different values of the exponent z, a fact that suggests that particular historical processes may have quantitatively influenced the current distributions of areas and populations.

Are the abundances of linguistic groups and the areas they span shaped by the same processes? Actually, though the distribution of domain areas is also log-normal, a multiplicative mechanism analogous to demographical dynamics –which explains the distribution of the number of speakers per language– does not appear as natural for the case of the areas. However, the strong correlations observed between both variables support the existence of a process that couples demographic growth to area occupied. It seems reasonable to assume that growing populations tend to expand their ranges, and that neighboring populations might clash if they both grow and thus compete for the same territory. Following this idea, it has been suggested [15] that the addition of a form of conflict between neighboring human groups might be the ingredient explaining the variations in the value of z. It remains to be seen whether a similar competitive scenario could be translated to the case of species.

7 Prospects

The multiple quantitative patters shared by biodiversity and human linguistic groups pose a number of questions related to their origin, causes, development, interaction, and fate. While some models shed light on the mechanisms behind some of the observations and occasionally reproduce them, others, such as the decrease of diversity with latitude, remain puzzling. It may well be that some patterns result from multiple causes, and in this sense be intrinsically more difficult to explain. Also, we cannot discard that some others which are apparently repeated in species and human groups happen to be due to chance or trivially result from external factors –as the two-dimensional space where they are bound to happen. The use and integration of independent data sets (from genetics, archeology, or history) will permit further advances in the characterization and eventual understanding of the ecological processes behind human cultural diversity, and likely of the relevant differences, if any, between human groups and biological species at the large scale.

8 Bibliography

[1] J. Aitchinson, *The Seeds of Speech. Language Origin and Evolution*. Cambridge: Cambridge Univ. Press, 2000.

[2] J. McWorther, *The Power of Babel*. New York: Times Books, 2001.

[3] M. Ruhlen, *The Origin of Language*. New York: John Wiley & Sons, 1994.

[4] L. L. Cavalli-Sforza, "Genes, peoples, and languages," *Proceedings of the National Academy of Sciences USA*, vol. 94, pp. 7719–7724, 1997.

[5] R. D. Gray, A. J. Drummond, and S. J. Greenhill, "Language phylogenies reveal expansion pulses and pauses in pacific settlement," *Science*, vol. 323, pp. 479–483, 2009.

[6] G. U. Yule, "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis," *Proc. R. Soc. London Ser. B*, vol. 213, pp. 21–87, 1924.

[7] D. H. Zanette, "Self-similarity in the taxonomic classification of human languages," *Adv. Compl. Syst.*, vol. 4, pp. 281–286, 2001.

[8] B. Burlando, "The fractal dimension of taxonomic systems," *J. Theor. Biol.*, vol. 146, pp. 99–114, 1990.

[9] D. Bickerton, *Bastard Tongues*. New York: Hill and Wang, 2008.

[10] S. L. Pimm and J. H. Brown, "Domains of diversity," *Science*, vol. 304, pp. 831–833, 2004.

[11] E. H. Rapoport, *Areography. Geographical strategies of species. Trad. B. Drausal*. Oxford: Pergamon Press, 1982.

[12] K. Rohde, "Rapoport's rule is a local phenomenon and cannot explainlatitudinal gradients in species diversity," *Biodoversity Letters*, vol. 3, pp. 10–13, 1996.

[13] R. K. Colwell, C. Rahbek, and N. Gotelli, "The mid-domain effect and species richness patterns: What have we learned so far?" *American Naturalist*, vol. 163, pp. E1–E23, 2004.

[14] M. H. Fernandez and E. S. Vrba, "Rapoport effect and biomic specialization in african mammals: revisiting the climatic variability hypothesis," *Journal of Biogeography*, vol. 32, pp. 903–918, 2005.

[15] S. C. Manrubia, J. B. Axelsen, and D. H. Zanette, "Role of demographic dynamics and conflict in the population-area relationship for human languages," *PLoS ONE*, vol. 7, p. e40137, 2012.

[16] R. Mace and M. Pagel, "A latitudinal gradient in the density of human languages in north america," *Proc. R. Soc. Lond. B*, vol. 261, pp. 117–121, 1995.

[17] J. B. Axelsen and S. Manrubia, "River density and landscape roughness are universal determinants of linguistic diversity," *Proceedings of the Royal Society of London B*, 2014.

[18] W. J. Sutherland, "Parallel extinction risk and global distribution of languages and species," *Nature*, vol. 423, pp. 276–279, 2003.

[19] D. H. Zanette, "Demographic growth and the distribution of language sizes," *Int. J. Mod. Phys. C*, vol. 19, pp. 237–247, 2008.

[20] F. W. Preston, "The commonness, and rarity, of species," *Ecology*, vol. 29, pp. 254–283, 1948.

[21] T. M. Blackburn, K. J. Gaston, R. M. Quinn, H. Arnold, and R. D. Gregory, "Of mice and wrens: The relation between abundance and geographic range size in british mammals and birds," *Philosophical Transactions of the Royal Society of London B, Biological Sciences*, vol. 352, pp. 419–427, 1997.

[22] J. Harte, T. Blackburn, and A. Ostling, "Self-similarity and the relationship between abundance and range size," *American Naturalist*, vol. 157, pp. 374–386, 2001.

[23] K. J. Gaston and F. He, "The distribution of species range size: a stochastic process," *Proc. R. Soc. Lond. B*, vol. 269, pp. 1079–1086, 2002.