# *Linguistic Diversity Through Data*

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von M.Sc. Physik *Damián E. Blasi*
geboren am 26. November 1985 in Buenos Aires, Argentina

Die Annahme der Dissertation wurde empfohlen von:
Professor Dr. Peter Stadler (Leipzig)
Professor Dr. David Wolpert (Santa Fe, USA)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 06.03.2018 mit dem Gesamtprädikat
**summa cum laude**

# ACKNOWLEDGEMENTS

A large number of people and institutions supported or contributed in some way to my doctoral research, which reflects both the increasingly social nature of science and the characteristics of my own research, permanently in need of specialists in different aspects of the study of language. Naturally, here I restrict my acknowledgements to the period 2012-2015 (otherwise the acknowledgement list would be too long!)

First of all, the Max Planck Gesellschaft has given me unmatched conditions of research, mobility and intellectual freedom that I have never experienced before or after in any of the many universities and research institutions I have visited. I spent my time as a student between four Max Planck Institutes (Mathematics in the Sciences, Evolutionary Anthropology, Psycholinguistics and Science of Human History) and I have visited or taken courses in another two (Cognitive Neuroscience and Complex Systems) and, in spite of necessary idiosyncratic differences among them, the climate of high-end research and a preference for cutting-edge science is one and the same.

Naturally, the MPI for Mathematics in the Science takes prominence over the others, if not only due to their bold decision of accepting me as a doctoral student even after I confessed that I wanted to do research on language. Part of my puzzlement disappeared after I met Jürgen Jost. In the first conversation that we had he brought to the discussion themes and research on neurosciences, linguistics, computer sciences, human history, economy, physics and mathematics. This amplitude of interests, I learned, was also reflected in his own scientific work: from genetics to geometry and the structure of meaning, Jürgen has published about all the possible objects of scientific inquiry. No surprise, then, that the accepted my plural interests as a natural thing, as a given in other fellow scientists. Even more, he always deposited his trust in my decisions, which varied between enrolling a course in primatology, attending a school in cognitive linguistics and taking a course on information geometry for biology.

Something similar I could say about Peter Stadler. I guess many times we looked like Aristotle's pupils, walking around Leipzig while talking about how to set up a particularly tricky test or how to navigate the complex world of scientific politics. There is an anecdote that illustrates transparently his integrity very well. Within the first weeks of my arrival, he told me there was certain algorithm that "needed to be done". I understood the allusion but I was not really thrilled about the task. After a week of hesitations, I went to his office and told him that I did not want to do it. He asked me "then what would you

do?". Since then he has always been supportive of my interests and my career choices. He has been ideal as a supervisor, hands down.

Apart from my other colleagues at the MPI MIS (specially, Felix, Gerardo, Wiktor and Yuri) and my friends at Bioinf (in particular Bruno, Tomas and Lorena) I also wanted to highlight how much I owe to Antje Vandenberg, who helped so many times with the German bureaucracy — it takes ages to make understand someone from South America how important is to fill forms and keep track of papers!

The Department of Linguistics at the MPI for Evolutionary Anthropology turned out to be my second (and for some periods, my first) home in Leipzig. Bernard Comrie considered that my initial vague interests in language were sufficient for letting me hang around in the department — sometimes in the library, some other times in the common area. It was Sören Wichmann, however, who took pity of me and offered me a space in his office — and not only that: he has cooked for me in more than one occasion. Many other people at the EVA gave color to my days spent there. Susanne Michaelis and Martin Haspelmath have been extremely nice and understanding with my first steps into the study of creoles. Several conversations with Paul Heggarty, Hans-Jörg Bibiko, Heriberto Avelino and Harald Hammarström spiced up my interest on several aspects of language, usually with a Weissbier at sight.

I received quite a lot of support from other scholars during the early stages of my career, either by means of direct advise (prominently from Morten Christiansen) or by being invited to give a talk (by Gerhard Jaeger and Balthasar Bickel.) A very special spot is occupied by Michael Dunn, who hosted me for almost six months at his group ("Evolutionary processes in language and culture") at the Max Planck Institute for Psycholinguistics. It was an unique chance for me to scrutiny the daily schedule of a busy and intelligent scholar - before (and during) that period I would work at odd hours (say, from 9 pm to 6 am) and have an open-ended working schedule (*what paper/project got my attention today?*). We never got to write anything together, but I can't underestimate how important were those months for my first steps as a functional scientist. In Nijmegen I met lots of interesting people, and some of them became quite rapidly my friends: Jeremy, Hedvig, Caroline and Séan.

I spent a brief but quite productive period at the Centre for Language Evolution hosted at the University of Edinburgh. My thanks go primarily to Mónica Tamariz and Simon Kirby who made that possible, and to the (large) Edinburgh crowd that interacted with me those days. Interestingly, that visit produced in my brain the unlikely association between *haggis* and language evolution experiments.

I finished my period as a PhD student writing this thesis at the Max Planck Institute for the Science of Human History. This was possible thanks to Russell Gray, who since then has been both a mentor and

a friend to me (plus one of the few people from which you can learn about fine cuisine and wines and the theory of evolution in the same meeting.)

My non-native English was smoothed by the amazing R. Schikowski, J. Mansfield and N. Uomini. All remaining mistakes are mine.

Stefany, who underwent the same process of exile out of South America, has been the single most important person in my life in the last seven years. We coped with the same issues and hoped for a better future together.

To my friends and family in Argentina (who are way too many to enumerate - my mom alone has 7 siblings!): gracias por el aguante. Sorry for not being there at the many birthdays, parties, *asados*, trips, sad moments, visits to the doctor, etc. that I missed because of this.

To Stefany and them: I owe you much more than this degree. I do not write anymore because this is supposed to be a scholarly piece that celebrates hard work and intellectual achievement and not yet another instance where I surrender to my genes poisoned with tango and melancholy.

# CONTENTS

# Part I

# INTRODUCTION

# 1

# LANGUAGES, DATA, AND LANGUAGE DATA

## 1.1 LINGUISTIC DIVERSITY (OR NOT)

Current estimates of the number of languages spoken in the world vary between 6000 to 8000, divided in over 300 groups that derive from a common ancestor, in some cases going back as far as the end of the Neolithic [102, 137]. Some languages are natively spoken by a large number of people of radically different cultural and ethnic backgrounds — like Portuguese, spoken from Brazil to Timor-Leste and Goa — whereas others are restricted to a few individuals and seem to face an unavoidable end in the near future — such as Kusunda, a language isolate from Nepal that is spoken by a few dozen people.

Languages appear as diverse as the dimensions in which we choose to classify and describe them. Most of them make an intensive use of the sounds we can produce by passing air from our lungs through the larynx, mouth and nose, whereas others extend this repertoire to include vocalizations that originate from air trapped and released rapidly in the mouth or even clicks of the tongue against the soft palate [92]. Other languages do not make use of sounds at all, namely the many signed languages that exploit the combinations generated by hands, body and facial gestures (e.g. Stokoe [215]).

While languages serve (among other purposes) to communicate about the most diverse events and situations, the linguistic expression of tense, aspect, mood and/or source of evidence (TAME) is usually obligatory. Nevertheless, there exist exceptions (as in some Austronesian languages [88]) where this is not the case, and the burden of inferring the messages migrates from the actual linguistic signal to its context. In at least one extreme case, Indonesian Riau, not only TAME markers but also number and other grammatical features are underspecified, so a completelly grammatical sentence like *Ayam makam* might mean anything from "the chicken eats", to "someone is eating with the chicken" or "where the chicken was eating" [88].

The order in which words need to be arranged in order to produce a grammatically valid phrase is usually restricted to one or a few options, but then again there are exceptions — most notably in the languages of Australia, e.g. Jiwarli [9] — where in principle all arrangements are grammatical. Most of the times, languages develop alternative ways of marking where the relevant information is

located when word order is absent — the aforementioned Australian languages, for instance, have a complex morphology that provides sufficient information about the role of each word in the sentence [58].

The management of old in relation to new information along discourse also shows interesting cross-linguistic variation. Most of the languages posses some resource for linking related sentences, for instance as a way of expanding or specifying the information about a N(oun) P(hrase). In English, one possible way of doing this is to attach a sentence after the NP that carries the relevant information *without* the explicit reference to the relevant NP: so we have **the man** *who [...] sold the world*, **the spy** *I loved [...]*, **the state** *I am in [...]*, where [...] signals the position where the NP would have been in a regular sentence. In Hebrew, in contrast, in all but the first case speakers would have filled the gap with explicit pronouns, e.g. **the spy** *I loved him/her/it* [43].

Examples like these abound. After carefully considering these figures, it would be reasonable to ask what do they mean for language. Is the variation described above very radical or modest in comparison to other human behaviours? Are the differences guided by external pressures that do not have to do with language proper or are they better understood as historical byproducts? Can we find a common structural template underlying all languages?

Depending on the theoretical affiliation of a researcher (and the swinging pendulum of intellectual history and its fashions), these facts are used for radically opposed arguments.

On one side we have several versions of nativism, famously promoted (and reformulated many times by) Noam Chomsky in a career that spans over five decades since 1957. The core observation resides in the way humans acquire language [21, 139]. Children are able to pick up the language spoken in their environment in a effortless manner. Notably, they generalize the bits of grammar they absorb but at the same time they avoid certain structures that would follow from a simple inductive rule of a frequent pattern in the language they hear [133]. And all of this in spite of dramatic contrasts in the amount and quality of input they receive: while children in several cultures are engaged into conversations by their caregivers, the usual account of language acquisition in Samoan is that they do not regard toddlers as intentional subjects, and, in consequence, they are not spoken to directly [216]. Given this evidence, nativists proposed that humans are equipped with an innate faculty for learning languages that requires only a small amount of linguistic input to develop [139].

The argument continues. If all languages are learned by tapping the same specialized biological resources, then it is expected to find some commonalities in their inner structure. Nativists argue that this is the case, although there is no consensus on how these common-

alities should be described (or even what they are.) They emphasise that the universal aspects of language are not observable at its surface. Instead, they are general computational principles that enable the construction of complex propositions out of words (or units smaller than words) [40]. For instance, a cross-linguistically attested way of generating questions is by means of *wh-movement*, which consists of producing a sentence similar to the declarative but with a change in the order in which the thing or circumstance being asked about appears - in English, "Leonard bought **mangoes**" → "**What did** Leonard buy?". However, some types of phrases do not allow this strategy, which lead to the postulation of *syntactic islands* [28]. These restrictions are believed to be universal and to reveal some of the computational characteristics of the faculty of language.

In conclusion, in spite of the superficial variation of languages and all the dimensions in which they might differ, they are all instantiations of the same innate bias, the *universal grammar*.

Others disagree. Every aspect of grammar — from the widespread existence of synonyms to alignment systems — seem to exhibit certain degree of adjustment to a particular set of functions of mostly social and communicative nature. A classic example is that linguistic forms change according to the use they receive: infrequent words or verbal forms eventually die out, whereas common syntactic strategies get fossilized into morphology or frequent words get shorter (and thus more convenient for re-use).

Even language acquisition seems to be more naturally described by data accumulation and statistical learning, in a fashion that does not seem to require language-specific computation or inferential rules [228]. Frequency of occurrence of most linguistic behaviour appears — other things being equal— as the most reliable predictor of age of acquisition [5]. Young children are demonstrably able to latch onto different aspects of the input — from stress patterns to relative order of elements [222] — to learn morphemes, words or word classes. The way in which they utilize certain linguistic resources seems to follow a developmental path from concrete to abstract: for instance, definite and indefinite articles in English are used exclusively with the nouns they appear for the first time, with more flexible combinations occurring at a later age (Pine and Lieven [192], Pine et al. [193], but c.f. Yang [242]).

Furthermore, many linguistic features (which were previously thought to be human-specific) seem to have analogues in other species. There is evidence of compositionality in non-human primate vocalizations — combinations of individual vocalizations have a meaning distinct to just the mere juxtaposition of the atomic meanings [246]. Some species of avian vocal learners undergo a babbling stage that is (developmentally and functionally) parallel to that of humans [140]. This lends support to the idea that, rather than a phylogenetic innovation,

language seems to be composed of a combination of precursor characteristics [77], enriched and expanded in functionality due to the overall increased cognitive skills of our species (and perhaps its drive towards cooperative behavior [227]) .

Finally, close examination of the complex mosaic of grammatical structures in the world raises the question of whether anything could be said in general about the computational procedures that produce them, given the apparent lack of observational universals [71]. If there are no non-trivial universal properties defining what a language is, then the +7000 figure is a testimony of the manifold solutions for varying needs humans encountered in their historic development.

This was (and still is) the polarized scientific scenario when I started my research in the field three years ago. The discussion is heavily tainted by complex sociological and theoretical considerations that have overshadowed, in many cases, the actual data: the languages.

The main goal of this thesis is to offer a third alternative: approach the unity and diversity of the world's languages with as little theoretical commitment as possible, using state-of-art statistical modelling techniques on large typological data.

## 1.2 A SHORT BIOGRAPHICAL MOTIVATION

During the first months of my period as a PhD student I attempted to translate some of these issues into the language of physics- and biology-inspired models. Among other projects (that I do not dare to remember) I attempted deriving the existence of non-arbitrary sound-meaning associations through game theory, finding Zipf's Law as a maximum entropy distribution subject to convenient constraints, and predicting dialect formation as a percolation problem. The attentive reader should have noticed that none of these feature in the present thesis.

At first I looked into textbooks searching for explicit, widely agreed statements about (more or less) mechanistic connections between general variables I could project into mathematically interesting objects. This approach is not new, as there are plentiful examples of varying degrees of success, some of which provided relevant insights and predictions. Just to give an extremely small sample of this work, scientists have modelled the diffusion of linguistic variants with the full glory of the Fokker-Planck equation [16], analyzed the monopolising nature of prestige languages with dynamical systems that are equivalent to those found in population dynamics [189] and interpreted the process of language acquisition by appealing to random matrix theory [181].

After some time, however, my approach was proven hopeless. In contrast to the confidence and the far-reaching claims often found in textbooks, my interactions with linguists — and very particularly, ty-

pologists — ended up in a similar way: for every statement or model idea I could come up with, there were as many counter-examples as corroborating cases[1]. Turning my attention to anthropologists and cognitive scientists did not help me to escape the conundrum.

Naturally, there exists an amazing wealth of insightful and careful scholar research on language. The situation is that most of that research was conducted largely in a qualitative way, with only a minimal usage of quantitative arguments to defend or rebut competing hypotheses. Qualitative analysis is a fundamental (and irreplaceable) tool for the human sciences — the open-ended nature of human behaviour asks for a flexible coding of the information, and important impressions obtained from data might be better reflected in a verbal argument than in a list of normalized variables in a table.

However, qualitative arguments certainly have their limitations. First, they are somewhat like the proverbial snow flakes: one of a kind. In order to engage with them, one needs to become familiar not only with the empirical facts involved, but also with the particular and idiosyncratic structure of the argument itself: how much weight does the author allocate to each of the parts? How critical is the likelihood of one assumption to the overall conclusion? In contrast, normalized quantitative analyses accelerate this stage: if someone is performing growth curve analysis or a deep learning classification, say, we know immediately what to ask and where to look. This insistence on explicitness and standardization of reasoning helps to uncover usually concealed biases.

The second limitation I observe has to do with the ability of human minds to manipulate large amounts of information. A hidden assumption behind qualitative analyses is that one needs to be *personally* acquainted with the data under evaluation: until very recently, if a linguist wanted to write about the cross-linguistic preference for suffixation or some other typological pattern, they had to go over grammars and code the information themselves after a number of far from trivial judgments and decisions. With over 7000 existing languages — and perhaps half that number of detailed grammars — it appears as highly unlikely that one could develop enough insight from any comparable magnitude of languages as to deliver high-quality inferences that are faithful to all of them. Given the case, one could either opt for a science that is bound by what is achievable through qualitative means by individuals, *or* attempt to abstract the procedures or decisions one takes in order to make them applicable in new cases — potentially *all* cases.

Finally, the overlap on the questions qualitative and quantitative methods can answer is not complete. I readily accepted that there

1 Although here I am being extremely generous with myself — my first intuitions about language and languages were very often helplessly wrong. The fact that most humans have at least one native language lends some mysterious confidence to the speculations we are able to come up with in relation to language.

are aspects of language research that cannot be made more clear or even translated into quantitative grounds, but the mirror situation is true as well: asking for the effect of a third-order interaction between variables in a regression regarding the impact of social circumstances in the choice of different varieties of a word, computing the confidence of a multiple-sequence alignment with the purpose of unraveling deep language history or characterizing the 95% credible interval of the posterior distribution modelled upon a Bayesian word learning experiment are all good questions relevant to language, but they have nothing near a qualitative counterpart.

These considerations lead me to the niche I currently, and happily, occupy. In a nutshell, my goal is to provide a faithful translation of the discussions on linguistics about the diversity and unity of languages in the space of data science. This sometimes leads to an enrichment of the original question as a result of the available statistical technologies we possess, which unlock the possibility of posing new questions that cannot be addressed without the proper technical machinery. A second (but not less important) aspect of the marriage between diversity linguistics and data science is that we can harmonize what we know about languages with data from other disciplines. This part has become increasingly more important in my most recent thinking about language.

Because of all these reasons, it is fair to say that the spinal cord of this thesis is methodological in nature, in spite of the amounts of data discussed and analyzed.

## 1.3 THE RISE OF THE SCIENCE OF DATA

Data science is a cover term for a complex methodological melange that arises from the endeavours of two scientific communities: statistics and computer science, in particular machine learning. A brief account of their history would be warranted in order to establish precisely the source of novelty in this work.

Traditional statistics grew out of the need of making sense of patterns present in developing disciplines without the mathematical maturation of the natural sciences, like medicine or agriculture [214]. Early computational power was restricted to (error-prone) humans equipped with pencil and paper, and this purely technical restriction had an overwhelming impact on the way statisticians reasoned about their discipline. Most of the tools produced in the dawn of statistics were highly parametric, partially reflecting also the mathematical drive of the early practitioners, who sometimes came from departments of pure mathematics and ended up working on very concrete statistical problems.[2] By and large, some of those tools proved to be

2 To take a case among many: George Barnard — well known as the proposer of the likelihood principle — did graduate research work with Alonzo Church in Princeton

resistant to time and they are still part of any standard data analysis toolbox. For instance, the famous t-test developed by Guiness's employee William Gosset is still part of computationally intensive analyses as in gene expression or fMRI voxel activation analyses. Other aspects of traditional statistical theory, like the imposition of unbiasedness on estimators, have fallen out of favour.

Propelled by the dramatic increase in data availability (and production) and computational resources that has taken place since the 60s, computer engineers started to ask open-ended questions to the data in hand (e.g. *are there any interesting recurring patterns in this data?*). Exploring the geometry of data or detecting the presence of "interesting" patterns became legitimate goals, even in the absence of clear hypothesis that could canalize the analysis. What is perhaps more relevant is that even when they faced problems *telically* equivalent to the those of regression/classification, they turned their back to mathematical explicitness or tractability and went for practical yardsticks to asses the quality of their solutions. Concerns about how well a solution generalizes in the presence of noise, predictive coverage or algorithmic complexity replaced the quest for good asymptotic properties or the suitability of a particular family of models [32].

One interesting (and highly relevant) twist of fate triggered by the rise in computational power is the return of Bayesian inference as a serious contender to the frequentist school. Beyond a few relatively simple cases, the computation of posterior probabilities — the crux of Bayesian inference — was a daunting task often avoided by assumptions as dramatic as those of the frequentists. In the 1980s [86] inaugurated a brave new world with the introduction of Gibbs' sampling, which was in its turn inspired by the Metropolis-Hasting algorithm developed during the heyday of Manhattan Project [37]. With Gibbs' sampling being a household technique nowadays, even standard laptops can simulate relatively complex Bayesian models in a couple of hours.

The insistence on the efficiency of a solution in contrast to how well it could inform the (conjectured) model underlying the data goes beyond computational considerations. As put by Leo Breiman [32],

> With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that cannot be substantiated by appeal to goodness-of-fit tests and residual analysis. Usually, simple parametric models imposed on data generated by complex systems, for example, medical data, financial data, result in a

on mathematical logic, but later worked on quality standards for condom production [1].

loss of accuracy and information as compared to algorithmic models

This original tension has been resolved in favour of a more harmonic cohabitation of statistics and machine learning in the core of the data sciences. Many standard methods for the analysis of data include elements of both sources. Take for instance any of the modern versions of regression techniques that include explicit regularization strategies, such as the elastic net [83]. The problem is casted as one of traditional statistics based on ideas that go back to Tikhonov in the 1960s [226]: penalize the wiggliness of a function or some of its derivatives by means of an explicit term that depends on an ad hoc parameter, which is formally a Lagrange multiplier. The gist is that, instead of choosing the magnitude of the penalization (i.e. the value of the Lagrange multiplier) by means of pre hoc considerations as in classic statistics, these days one would choose the value that minimizes some empirical goodness-of-fit measure, like the prediction error [83].

It is not exaggerated to say that we are experiencing a golden age with respect to the development of intellectual and practical technology to understand data. Cunning and/or efficient strategies are being published in journals regularly, and reasonably operative (and fairly bug-free) packages are released in parallel with them, sometimes simultaneously. Statistical programming is becoming more statistics than programming, in the sense that the user can focus on the specifics of the analysis thanks to a wealth of efficient pre-defined functions without the need of dealing (most of the times) with limiting factors like memory allocation.

In the development of any technology, early adopters tend to be skilled individuals with an above-average understanding of the matter, and then it spreads to the rest of the population. From the first Daguerreotype users — versed in the chemistry of silver salts and the timing of light and movement — to the crowds of compulsive selfie-takers there is a clear cline of intellectual involvement with the tool. Data science methods are no exception. There are armies of graduate students applying the latest data science algorithms to whatever dataset could be found under the sun. The *modus operandi* is to calque a successful case study by finding analogies in the datasets. Even more, there are many wrappers that allow the massive application of different methods by simply stipulating the nature of the problem. For instance, if one needs to choose the subset of most efficient predictors for regression, it is possible to get with one line of code the outcomes of no less than twenty independent methods ranging from forward and backward elimination to model averaging and lasso selection. Needless to say, the variety of available methods do not correspond (only) to the vanity of the data scientists: they rest on different structural assumptions.

Now that the correlation between statistical expertise and capacity to implement data science techniques is broken, two problems grow at alarming speed. The first is that of the *consensus fallacy*: if a particular statistical claim is confirmed across methods then it is often assumed to be genuinely true; on the contrary, if only a few methods sustain the claim, then it is taken to be genuinely false (e.g. as in Roberts et al. [201]). While this makes sense in extreme situations when there are no good guesses about the structure of the problem, in my experience the most challenging cases require heterodox specifications that might go against the intuitions that hold for other, more typical, cases.

The second (and potentially more dangerous) issue is that of post-hoc argumentation. It is technically possible (and practically simple) to find a method that could yield a result in tune with a previously held bias. Scientific journals do not always enforce a thorough discussion of the context of choice of the statistical methods employed, and the celebrity of a technique seems to trump suitability in many cases.

The remedy to these issues is, unsurprisingly, cross-pollination between data specialists and those who have expert knowledge of the field.

## 1.4 DOING SCIENCE WITH LINGUISTIC DATA

The choice of the datasets used in this work — and, necessarily, the issues investigated through them — was largely opportunistic. A decade ago there was a single available dataset relevant to this thesis, the imperfect yet monumental World Atlas of Language Structures (WALS [103]). The creators of that resource, the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology, followed that direction and produced no less than a dozen other databases during their 15-year tenure [2]. The creation of new, statistically useful and scientifically sound databases requires enormous amounts of human and material resources. While the production of published results largely dictates the dynamics of scientific practice, it is my belief that all scientists should contribute to the extent of their possibilities to the generation or curation of publicly available datasets. In agreement with these principles, I am currently involved in two large-scale projects by the Max Planck Institute for the Science of Human History that aim to develop free, easily accessible and high quality data relevant for the language sciences: Glottobank [3], a large-scale inventory of genealogical, typological and lexical information of the languages of the world; and D-Place [128], a linked database that groups linguistic, anthropological and environmental data for interdisciplinary research. Also, and thanks to the generous support from the Max Planck Institute for Evolutionary Anthropology, I lead the development of a database on noun and verb preva-

lence in the languages of the world (as yet unnamed), which constitutes a personal landmark, since for the first time a research question gives place to data collection and not vice-versa.

I explored almost every available dataset that could be useful in the diversity versus unity debate — from data on affix borrowing to color terms, numeral systems and subsistence mode of the languages' populations of speakers. Not all of those results were included here. Some will require more analysis well beyond the scope of this thesis and will, hopefully, yield results in the near future. More critically, during the development of these resources it was not always envisioned the use they presently receive by practitioners of quantitative methods. These shortcomings translate most of the time into lack of statistical power, or non-overlap, or massive amounts of missing data. While sometimes this can be spotted in a superficial analysis, in some occasions this characteristic will show up only after a considerable amount of work has been invested already. This is exactly the situation that justifies one of the important aspects of this thesis: how to analyze contemporary linguistic data and, to some extent, how they differ from other kinds of data.

However, this is just one of the two intended goals of the present thesis. Beyond wrestling with partially technical problems, I hope I have been able to show that data science is more than simply a panoply of recipes and tools, an ancillary discipline that might assist the language scientist when the necessary step of statistical argumentation arrives. Under the risk of being hyperbolic, the data sciences are technologies for thinking.

## 1.5 A PARALLEL EFFORT

While the extent of this thesis covers a large number of topics, the absence of some issues and methods of prominence (for those with a statistical eye for language) might come across as curious. The reason can be inferred from the intention underlying this thesis: once a method becomes established as the unmarked answer for a problem, then either turns one's attention to fine-grained technical facets or forgets about the method per se and centers on the concrete issue being faced. The *finesse* and the argumentative nature of the papers in which a method is introduced in a field is lost with its successive improvements. In the best case, this might mean that the fundamental insight present during the conception of the method becomes a shared and understood legacy, and effective progress is made.

There are active communities that customarily use statistical tools for the study of language who succeed in establishing some common ground to their practices. With few marginal exceptions, no one would think that using mixed effects models for analyzing data from psycholinguistic experiments or Bayesian phylogenetics for the recon-

struction of language histories is odd or unsuitable nowadays. It is warranted to provide here a bird's eye-view on some of these.

The last fifteen years have seen the coming of age of methods from bioinformatics and evolution applied to the history of languages. Noblesse oblige, the reconstruction of language history has been one of the earliest fields to see a quantitative approach in linguistics [48]. The fundamental concept was that of glottochronology [220]: languages will tend to lose cognate words at a constant rate, which effectively translates the problem of inferring dates of divergence to one of assessing the fraction of joint retentions across the vocabulary. Unfortunately, the hypothesis was shown to be unsound early on [19] and was mostly abandoned. These ideas were revamped thanks to the methods of phylogenetic inference borrowed from biology [136]. While maximum likelihood methods dominated the dawn of this movement [179], Bayesian approaches have grown to become the standard in the field, in great part thanks to the development of powerful Markov Chain Monte Carlo techniques that paved the way to relatively fast calculations of posterior distributions, as discussed before [136].

Among the landmarks of this enterprise I could mention the analysis of the Pacific peopling by the Austronesian family [94], the support towards the Anatolian, agriculture-driven hypothesis of the Indoeuropean homeland [93], as well as insights on the unfolding of the Pama-Nyungan [30], Arawak [238] and Bantu [111] families in Australia, South America and sub-Saharian Africa, respectively.

Furthermore, phylogenetic inference was not the only tool borrowed from genetics that has proven extremely useful for the tasks historical linguistics deal with: the analogy between phones and phonemes and nucleic basis triggered the application of sequence alignment methods to wordlists [8, 213]. By aligning wordforms from different languages it is possible to learn what the usual phone or phoneme substitutions are; strong regularities in the replacement patterns are thus equated to hypotheses of regular sound change, namely the essential ingredient for judgements of common ancestry in a group of languages [119, 213].

The study of linguistic patterns of behavior has also reached a considerable degree of methodological maturity. Psycholinguistics has been dominated by the all-purpose statistical tool of psychology, the venerable analysis of variance (ANOVA). Simplifying a lot, experimental research in psychology aims to establish the general validity of a perceptual, cognitive or motor phenomenon by running multiple times a setup with different individuals, substantive items and conditions. ANOVA was the natural response to the need of accounting for this variability present in the design of the experiment itself. While useful, it requires several tests to be performed in all the different dimensions of variation. With better optimization tech-

niques and computational power, ANOVA has been steadily replaced by mixed effects models [10, 116]. The efficiency and relative parsimony of mixed effects models have had a long lasting impact on other language-related fields as well, in particular in the analysis of corpora [224].

It is also worth mentioning the Bayesian school of inference, which has produced some important work in the fields of language acquisition and processing [38]. One of the most interesting aspects of Bayesian statistical modelling is that it allows to explore what the suitable space of competing hypothesis might be, in contrast to uniquely revealing what the specific computational or acquisition model effectively is [80–82, 97] (c.f. Endress [69]).

Fortunately, there are more examples to add to this list, and hopefully the next years will see even more revolutions in the methods and the standards of dealing with linguistic data.

## 1.6 TWO CHALLENGES OF LANGUAGE DATA

### 1.6.1 *On comparing language structures*

Aggregating data on specific variables across languages presupposes that those variables are commensurate up to a certain extent. While folk intuition about language and languages would dictate that this should not be problematic, the reality is that there is no single linguistic category for which at least a paper exists denying its universal character, not even words [104]. Clearly this is not an issue that I could do any justice here, but I consider that it is important (and honest) to layout the basics of the discussion.

A lexical category or a grammatical feature — which are the linguistic objects we will be comparing most of the time — could be proposed on the basis of a structural diagnosis rule. One could say, for instance, that an affix is a morpheme that is bound to appear with a word (which is modified by it in some way), or that the passive construction promotes the object to subject position while relegating the subject to an oblique argument position (if expressed at all), thus effectively reducing the valency of the verb. While this approach is usually found in the context of definition of language-specific categories, some authors have produced lists where various of those diagnoses are compiled in the belief of having sufficient scope to accommodate unseen instances (e.g. [59]).

In the end, these definitions cannot escape the need of some sort of semantic and/or functional inspection. For instance, in the definition of passive construction discussed before, we need to find a verb whose valency is being reduced, but determining whether something is a verb requires access to the referent of the word. Tentatively, it seems that basing linguistic concepts on semantic properties is a

more reliable strategy since we expect to find the basic distinctions at the heart of grammatical theory, like the notions of tense, person, number and so on, to be readily available in all human languages. By appealing to these commonalities we can build useful comparative concepts that try to capture how different languages express those meanings linguistically.

However, structural and semantic definitions are customarily confounded [200]. One example is the *heaviness serialization principle* proposed by Hawkins [105], a hierarchy of linguistic categories in relation of how they order with respect to the noun: so for instance, relative clauses are "more or equal rightward positioning relative to the head noun across languages" than adjectives. The problem is that, while relative clauses are often detected by structural means, adjectives are deemed so in virtue of their capacity to modify or expand the information about a particular referent — a function that might be perfectly executed by a relative clause (e.g. in English *the car* **that has flashy red paintings on it**). So in the end, the heaviness serialization principle might convey information about entities that are not completely comparable to each other.

To sum up, I cannot guarantee that all the data used here have been built on consistent principles. I could simply step away from this conundrum and shift the weight of the responsibility to my fellow linguists on this, but it is my opinion that — at least at the level of analysis I conduct in this thesis — these issues are not critical. The intuition is that a consistent phenomenon of enough magnitude should be detectable even in the presence of some mis-specification.

### 1.6.2 *On comparing languages*

As discussed early in the introduction, for most of the languages of the world there exists evidence of some kind linking to another language or languages, which is often cued by similarities at different levels of description. Methodologically, this becomes a concern when considering the almost ubiquitous independent and identically distributed —or familiarly, IID— data requisite present in a large fraction of the statistical arsenal. Languages are not independent from each other, and we need to tap on history to estimate how strong or likely their dependencies are. This problem has been acknowledged very early in the statistical literature and it receives the name of "Galton's problem", after Galton [84].

The inference of languages' (and peoples') history is not a concern of this thesis. However, a firm grasp of the history of languages is critical for the success of the ideas presented here. Since my goal is to detect regularities and explain divergences between languages as systems of communication used by people, it is important as a first step to separate the wheat from the chaff and remove the dimension

of historical relatedness. In other words, if there are structures and processes that are overwhelmingly represented across the world's languages, it is crucial to detect whether that is due to a functional drive instead of the presence of sweet potatoes, horses or canoes.

The Glottolog classification [102] — that keeps track of published evidence about the common genealogy of languages — counts over 430 lineages regarded as independent for the time being. These lineages range from simple isolates to intricate large families with many subdivisions. The overall number of languages per lineage can be decently approximated with a power-law distribution [240], and the same holds for the number of languages that are monophyletic at different scales within lineages [244].

These figures are not a reliable mirror of the past. Successful domestication of animals and plants seems to lie at the heart of the largest linguistic families' dispersals [54]. While at the individual level early agriculturalist and pastoralist practices did not necessarily improve the quality and expected duration of life, they did boost the carrying capacity of the populations, thus giving them an edge over smaller hunter-gatherer groups. How much linguistic diversity has been lost to the spread of massive families like Bantu or Arawakan can only be estimated by the large genetic diversity observed in those areas, that lies in sharp contrast to the relative homogeneity of the languages spoken by the populations [17, 159].

A related question is how far back in time we have to go to find truly independent language structures. While the issue of the (geographical and temporal) origin of human languages received far less attention than the equivalent problem in human evolution (namely the monogenetic or polygenetic beginnings of our species, e.g. Jobling et al. [123]) it has been recently galvanized and more widely discussed. Atkinson [7] suggests that the complexity of phonological repertoires — which is roughly a measure of the number of the phonemes present in a language — seems to follow a cline from South Africa all the way to South America, following the putative routes that our ancestors followed in our exodus out of Africa. The mechanism proposed by Atkinson is formally identical to the population bottleneck phenomenon that predicts a decrease of genetic diversity in subpopulations with respect to the populations where they derive from. While the analogy is appealing and the idea is worth perusing, it is fair to say that the question of whether all languages share a common origin is still open.

This situation opens the possibility of some linguistic traits having a present day dominant distribution thanks to being in place in the inception of our verbal behaviour. The basic word order SOV has been conjectured to be the most ancestral state in which languages can order the participants of a transitive event in speech, based on diverse arguments (reviewed in Chapter 2). Some sound-meaning

associations that pop-up across unrelated languages are also believed to be fossils of an ancient language (as I discuss in Chapter 3). In particular, this argument was held about nursery words for mother "mama" and father "papa" — they are learned so early in life and the referents are so universally present that the words are not influenced by the circumstances to which the rest of the vocabulary is exposed [13] [3]

Similarity between languages transcends the strictly vertical inheritance scenario. Multilingualism is widespread [137], and the sustained employment of multiple languages leads, in some cases, to different types of borrowing between them. While the processes and conditions that lead to this are multifarious, there are regularities in the material and the direction of change of borrowings [225]. There is a gradient that goes from isolated lexical material (which is easy and frequent even between languages with no active multilingual speakers) to structural or syntactic patterns that require multilingual speakers with high competence in the relevant languages.

Naturally, similarities due to borrowing are more likely between languages spoken in spatial proximity. The varying degrees of interaction patterns between populations — dependent on characteristics such as commercial ties, common cultural features and political domination — have yielded regions where borrowing shaped multiple languages in a homogeneous manner, producing commonalities that transcend genealogical affiliation. In some cases, the number of these features is such that linguists talk about *convergence areas* [153].

One well-known case is that of Mainland South East Asia, conformed by languages from different families, some of which came into coexistence due to the pressure of the Chinese Empire in the north [152]. The common characteristics of these languages are simple syllable structure, fairly complex tonal and consonantal systems, isolating morphology and the usage of numeral classifiers. Another popular case is the development of a Balkan *Sprachbund* under the domain of the Ottoman empire, which resulted in similar verbal and case systems [229].

Dominance of local interactions notwithstanding, researchers have put forward a number of areal phenomena of continent-wide scale. Perhaps the most far-reaching of such hypotheses concerns the Pacific Rim, which covers the Americas, Oceania and the territories and islands closer to the Pacific coast of Asia. The fundamental idea is that a few common features — like numeral classifiers, head marking and the inclusive/exclusive distinction in pronouns, among others — are regarded as the by-product of successive migration and contact

3 The classic (and still widely accepted) explanation for the similarity of these words across languages is still Jakobson's: the articulation of the stop and nasal consonants are the simplest to produce, and as such the parents, in a rather egocentric way, associate these first words to themselves [121].

patterns that have taken place in the area at least since the Neolithic [25].

This is a complex matter but, in a nutshell, I deal with non independence either by specifying appropriate covariances according to genealogical or areal groups (as in Chapter 5 by means of random effects) or by weighting the evidence so that the results are representative of the languages of the world (for instance through resampling techniques, Chapter 5).

Critically, linguistic genealogies and areas and their time depth and composition are objects of debate, so there are several competing hypotheses about what the appropriate groupings are. In this thesis, I will privilege consistently some of these classification over others: specifically, I would use Glottolog [102], AUTOTYP [180] or eventually WALS [103] for genealogical information, at the expense of Ethnologue [137] or classifications that assumed macro-families as well-established. Glottolog and AUTOTYP complement every genealogical statement with published evidence, something that the Ethnologue fails to deliver [100]. I use WALS only in contexts in which that classification is contrasted with some other (as in Chapter 3), or when I make use of genealogical units of roughly comparable time depth (like in Chapter 2), which WALS provides under the name of "genera". For areas, only AUTOTYP provides precise areal information at a level smaller than continents; for those, I usually use the classification by Hammarström and Donohue [101] or that one by Dryer [62]. Differences in how they group Australia, South-East Asia and the Pacific notwithstanding, they coincide on the linguistic soundness of North America, South America, Eurasia and Africa as areas of ancient contact.

## 1.7 STRUCTURE OF THE THESIS

This thesis is naturally divided in two parts, both in terms of the methodological approaches and the aspect of linguistic diversity being in focus. In what follows I present a summary of these sections and their corresponding studies.

The first part concentrates on unraveling common patterns across the languages of the world. While the choice of the data and the relevant dimensions of observation are inspired by a wealth of linguistic theorizing, the prevalent stance in these investigations is that of the detection of regularities with a robust statistical underpinning.

- Robust sound-meaning associations. The arbitrary relation between sound and meaning is hailed as a design principle of language, with exceptions regarded as marginal and anecdotal. Using the largest lexical database to date, it is shown that there exist consistent biases that make the presence of certain segments more likely to appear in the words associated with

certain concepts, and that this affects a considerable fraction of the lexical concepts tested. This phenomenon cannot be explained away by appealing to areal or phylogenetic causes that bind languages together, nor to other detectable confounding factors. Further explorations on the areal and diachronic features of these regularities are discussed.

- The internal structure of word order patterns. Historically and conceptually, the study of the relative order in which words are arranged to form sentences is key to our understanding of linguistic diversity. There are languages that are almost specular images of others in this sense — for instance, while in varieties of Aymara (spoken mainly in the Andean portion of Bolivia) the adjective, demonstrative, numeral and genitive classes all precede the noun they refer to, in the Berber languages (which belong to the Afroasiatic family and are found in the Sahara) they all come after it. In spite of this, only a handful of combinations are attested, which triggered an almost century-old discussion on what the latent causes of this asymmetry could be. Cognitive, diachronic and communicative factors have been proposed as explanations, but only a few produce predictions that can be corroborated empirically. We found that the observed word order patterns are compatible with some of these.

The second part of the thesis is focused on the evaluation of models that could contribute to the understanding of linguistic diversity. The latent philosophy is that potentially not all the variation is simply due to unsystematic and contingent forces acting upon languages, and that a careful scrutiny of the extra-linguistic factors influencing languages might deliver a better understanding of the source and the dimensions of plasticity.

- Creole as a typological group. Creole languages are special languages in that they underwent a peculiar process that contrasts with the classic tree-thinking paradigm: they emerge from extreme contact situations (i.e. circumstances of strong interaction between speakers of different languages with no common language among them, like in multinational commercial outposts). Strikingly, creole languages have been observed to have similar typological properties - they seem to prefer certain word order configurations and they all have very regular nominal and verbal patterns, just to name a few. This curious parallelism opens up an interesting research question: can we infer typological properties from the kind of history the languages have had, and can we infer something about the genesis of a language just by looking at the features of its grammar? I provide to this (heavily polarized) debate a classification analysis that captures these

questions. I show that sampling decisions have a dramatic effect on the results, but that in principle the common patterns found so far can be readily explained by means of a common European ancestry of many of these languages.

- Ecological pressures on speech. Human language is a (prominent, complex, perhaps unique and unrivaled) animal communication system. While the unity of biological computations underlying macaque calls, Passerine's songs and human language is a hotly debated topic, it is clear that the anatomical support and the physical medium for vocalizations for speech in humans can be readily put in relation to that of other animals. In contrast to other animal communication systems, however, most of speech occurs at short distance, without major metabolic investments and in relatively noise-free environments — all factors that have been studied as shaping forces on the nature of communicative signals from Anurans to insects. What happens when humans communicate in adverse environments, when the phonation becomes costly or diffuse or when they are bound to communicate across long distances? I find that human languages can be shown to be sensitive to at least some of these factors.

Due to the fact that this thesis is intended to be read by specialists with potentially non-overlapping backgrounds, I sometimes provide detailed discussions on facts or methods that are usually given only a cursory treatment at this level. In general, I begin each chapter with the the proper background on the issues at stake and a personal appreciation of the situation, mostly in terms of language sciences. A final point of divergence between this thesis and others is that, instead of having individual sections with completely independent discussion of the statistical or machine learning techniques to be featured, there is usually a two-way dialogue between the technical side and the idiosyncrasy of the data being analyzed.

Most of the text in this thesis has been written exclusively for its purpose. In some cases, excerpts from the relevant papers have been included. For obvious reasons, I have not included any piece of text written mainly by some of my co-authors.

## 1.8 SCIENTIFIC OUTPUT ASSOCIATED WITH THIS THESIS

### 1.8.1 *Peer-reviewed journals*

- **D. E. Blasi**, S. Michaelis and M. Haspelmath, *Grammars are robustly transmitted even in extreme situations: the emergence of creole languages*, (to appear in Nature Human Behaviour).

- **D. E. Blasi**, M. Christiansen, S. Wichmannm, H. Hammarström and P. Stadler, *Universal sound-meaning associations across the languages*. Proceedings of the National Academy of Sciences (2016)
- **D. E. Blasi** and S. Roberts, *Beyond binary dependencies in the structure of the world's languages* In N. Enfield (ed), *Dependendencies in Language*, Language Sciences Press (2017)
- C. Everett, **D. E. Blasi** and S. Roberts, *Language evolution and climate: the case of desiccation and tone*. Journal of Language Evolution (2016)
- C. Everett, **D. E. Blasi** and S. Roberts, *Response: Climate and language: has the discourse shifted?*. Journal of Language Evolution (2016)
- M. Dingemanse, **D. E. Blasi**, G. Lupyan, P. Monaghan and M. H. Christiansen *Arbitrariness, iconicity and systematicity in language*. Trends in Cognitive Sciences (2015)
- C. Everett, **D. E. Blasi** and S. Roberts, *Climate, vocal folds and tonal languages: connecting the physiological and geographical dots*. Proceedings of the National Academy of Sciences (2015)
- S. Moran and **D. E. Blasi**, Cross-linguistic comparison of complexity measures in phonological systems. In F. E. Newmeyer and L. B. Preston (eds), *Measuring Linguistic Complexity*. Oxford University Press (2014)

### 1.8.2 *Peer-reviewed conference presentations*

- **D. E. Blasi**, *Do ecozones affect the development of phonological systems?*, Phonetics and Phonology 9, University of Zürich, Switzerland (2013)
- **D. E. Blasi,** *New methods for causal inference in the language sciences*, New Developments in the Quantitative Study of Languages, University of Helsinki, Finland (2015)
- **D. E. Blasi**, S. Roberts and C. Everett, *How climate affects the evolution of languages*, 18th International Conference in Phonetic Sciences, Glasgow, UK (2015)
- S. Wichmann, D. Dediu, M. J. Dunn, H. Hammarström, G. Jäger, **D. E. Blasi**, T. Zakharko, and B. Bickel, *Simulating language, family, and feature evolution: a review of the state of the art*. Association for Linguistic Typology Meeting 10, University of Leipzig, Germany (2013)

# Part II

# UNRAVELING UNITY IN THE WORLD'S LANGUAGES

# 2

# DEPENDENCIES IN WORD ORDER PATTERNS

## 2.1 COMBINATORICS WITH WORDS

The way in which languages arrange words to conform grammatical patterns has been, and still is, a prominent topic in the sciences of language. The classic approach to typological word order studies begins by endorsing certain lexical or phrasal categories (e.g. "adjective" or "relative clause") to then determining what their relative positions are within a coherent unit. To illustrate this with a concrete example: the demonstrative can appear before the noun it refers to[1]:

(1) *má* ntamá — Fore, Trans-New Guinea
**this house**
"this house"

or the other way around,

(2) *mašuk* wak kulak ti — Maba, Nilo-Saharan
**man** **this** tall be.3SG
"this man is tall"

or even both orders could be acceptable, maybe within different constructions

(3) *il-qaraar* haadha — Gulf Arabic, Afroasiatic
**the-decision this**
"this decision"

(4) *haadha il-qaraar* — Gulf Arabic, Afroasiatic
**this** **the-decision**
"this decision"

and finally it is also possible for the noun to be flanked by two (not necessarily identical) demonstratives

(5) *yo miu yo* — Milang, Tibeto-Burman
**this** **boy this**
"this boy"

1 All examples are taken from the WALS chapter "Order of Demonstrative and Noun" by [63]; references to the glosses below can be found there.

In contrast to other structural properties of language, some word orders have the practical virtue of being *relatively* simple to elicit. Relatively, because (unlike English) most languages allow more than one particular way of arranging words in valid ways, and sometimes it becomes hard (or directly impossible) to determine a unique word order. For instance, a common linguistic strategy consists in changing the order of words to channel attention over a particular component of the sentence, usually "fronting" the relevant material [35]. For those reasons language descriptions usually report the canonical order of words, which would correspond to the most frequent and/or the one with the least number of pragmatic implications [209]. While I do not claim that the notion of canonical word order is uncontroversial (see e.g. [131]), I embrace the principle suggested in the introduction: if a phenomenon is robust and recurrent, it should be able to be detected in spite of noise or misspecification.

To the researcher unfamiliar with the field, word order might appear as a rather uninteresting feature — at least when compared with the luxurious variation exhibited in other domains, some of which I have discussed before. The importance of word order studies lies, ultimately, in the puzzling observation that only a few of the many configurations are enough to account for the large majority of attested languages [96]. The question of why some word order arrangements are so frequent (while others are vanishingly rare) and what is the reason behind the coincidence of so many different word orders into regular associations are a micro-cosmos of the intense theoretical and methodological debates in the language sciences. Where some propose that the observations follow from deep cognitive or processing pressures, others suggest that they derive from historical processes, potentially unique and unrepeatable. In contrast to those defending the idea that these patterns found their explanation within the domain of grammar, yet another group argues that we should look instead to the brain, to the structure of the events as they unfold in the world, or to the degree of political complexity attained by the population of speakers. The far-reaching implications of these questions justify, hence, the fundamental status of word order in the study of language.

## 2.2 THE PLACE OF THE SUBJECT, THE OBJECT AND THE VERB

Joseph Greenberg, the father of the 20th century linguistic typology, researched, wrote and speculated profusely on the order of the components of simple transitive clauses: subject (S), object (O) and verb (V) [96]. While there are as many word order patterns as combination of lexical classes and grammatical categories, understanding *basic word order* is a key to the main discussions and competing theories in the field.

| Word order pattern | Number of languages | Sample fraction |
|---|---|---|
| SOV | 565 | 0.41 |
| SVO | 488 | 0.35 |
| VSO | 95 | 0.07 |
| VOS | 25 | 0.02 |
| OVS | 11 | 0.007 |
| OSV | 4 | 0.002 |
| No basic word order | 189 | 0.13 |

Table 1: Basic word order counts based on Dryer [64]

As of today, different sampling and counting schemes converge to the claim that three orders out of six (SVO, SOV and VSO) monopolize the large majority of the occurrences. On the other extreme, earlier in the 20th century there were no attested cases of OVS and OSV languages, but Amazonia (and to a lesser extent Australia and Sub-Saharian Africa) delivered the missing data points — see Table 2.2

SOV is by and large regarded as the primitive basic word order of one or multiple prehistoric languages [85, 155, 176]. Gell-Mann and Ruhlen [85], backed by a large collection of historical word order transition events (and linguistic genealogies) found that, with the exception of diffusion processes, the large majority correspond to a change *from* SOV[2]. Coming from a different angle, Maurits and Griffiths [155] studied the ancestral states of basic word order of seven large linguistic families with Bayesian phylogenetic methods, based on trees reconstructed from cognacy judgements. Using uniform priors over the word order patterns, they found that four of the families — Afro-Asiatic, Indo-European, Sino-Tibetan, and Trans-New Guinea — have SOV as the order with the largest posterior probability at 10,000 ybp, whereas the other three — Austronesian, Nilo-Saharan and Niger-Congo — have instead VSO, VSO/SOV and SVO respectively. However, when all these families are rooted together, coalescence before about 50,000 ybp leads to SOV as the most likely ancestral state.

Far-reaching reconstructions of this kind are not unwarranted given the relative stability of word order patterns through time and their resilience to be replaced by borrowing. In the classic scale by Thomason and Kaufman [225], word order borrowing only appears in situations of "intense or strong cultural contact" between populations. Need-

2 Their claim partially depends on the existence of controversial linguistic *macro-families*. For instance, they mention that the Amazonian OVS and OSV languages descend directly from SOV by assuming that they belong to the Amerind macro-family, a large group with extensive coverage in the Americas ranging from Selkńam in Tierra del Fuego to Blackfoot in the Northern plains of the US and Canada. The evidence for these groupings and other macro-families — lookalikes in the pronoun paradigms — is, in the best of the cases, weak.

less to say, there are well attested cases in which they are induced by contact, as for instance with Austronesian languages spoken on the East coast of Papua New Guinea that switched from SVO to SOV (in tune with the languages spoken in the island) [31].

Furthermore, SOV appears in experimental conditions where subjects are forced to convey a message on a transitive event without using speech. In what came to be a contemporary classic, Goldin-Meadow and collaborators [90] asked speakers of languages with different basic word orders — English, Turkish, Spanish, and Chinese — to either describe a picture with gestures or to stack a number of transparencies (with objects or actions), all depicting basic transitive event such as a man playing a guitar. Regardless of their language, individuals overwhelmingly preferred the agent-patient-action sequence both in gestures and in the arrangement of the transparencies. Equating these semantic categories to S, O and V might not be straightforward, though, and others have provided similar evidence for gestural preference for SVO [87].

Nevertheless, the development of a few recent sign languages suggests that, rather than being *only* an experimental effect, languages might latch to this SOV preference in some circumstances. The most famed case comes from Al-Sayyid Bedouin Sign Language [203]. It was documented across three generations of an endogamous population that carries genetically recessive deafness in the Negev region of Israel, and its signers converged into a regular (S)OV language. While they interact regularly with (hearing and deaf) Hebrew and Arabic speakers, the gist is that neither language could be the source of its basic word order — Hebrew and the colloquial and classic varieties of Arabic do not allow SOV, and Israeli Sign Language uses SOV very rarely. More in general, it has been recently pointed out that in all attested sign languages SOV is always grammatical [173].

SVO is, in contrast, the state to which any other word order is more likely to transition, which on the long term implies an increase in the overall number of SVO languages [85]. This observation sparked a number of studies aiming to determine what could possibly be the advantage of SVO over the other possible word arrangements.

In transitive scenarios, the symbol S usually represents an agent, and O, a patient. It has therefore been suggested that SVO mimics the unfolding of events through time: the Agent (S) performs an action (V) that affects the state of the Patient (O) [233] [3].

3 Curiously the same kind of argument was used by XVIII philosopher de Condillac to suggest that OVS — incidentally, the rarest basic word order — is the most natural account of events:

> [...] first the noun indicating the object one was talking about, then the verb indicating the operation one intended to carry out on that object: for example, fruit want; the subject of the verb came at the end of the whole series: for example, fruit want Peter.

Dependency length minimization is a reasonable candidate for a pressure that shapes grammars [75, 89]. Across the languages of the world, agreement between the S and O arguments and V is considerably more frequent than S and O agreement between them. This produces a dependency between V, on the one hand, and S and O on the other — somehow the language processor needs to access the verb when determining the shape of O and S. The farther they are from V in the speech stream, the more difficult the operation becomes. By placing V between S and O the overall dependency distance would be minimized, and as such it turns out to be the most efficient basic word order in this respect [74].

Research on gestures and charades has also been used to argue for the functional value of SVO. Gibson et al. [87] have shown a preference for the agent-action-patient order when the patient could potentially serve as an agent as well (as in "Dante loves Beatrice" in contrast to "Dante loves wine"). By separating the two arguments, they argue, the patient and agent roles are harder to confound with each other. However, further research has shown that individuals do not seem to have any problems parsing expressions with two adjacent animate arguments, and they all seem to adjudicate agent role to the first argument [99]. This agent-first bias has been shown to be robust, popping out even in languages where the expectation would be biased towards finding non-agents in the first position [26].

Animacy and agenthood are not the only reasons that have been invoked to switch preference from SOV to SVO. Schouwstra and de Swart [206] reviews previous experimental setups and conclude that the verbs that occur on those are of the *extensional* kind, where a change in the world is implied — someone moves/cuts/breaks/eats/ . . . something, typically. When similar pantomine experiments are performed with *intensional* verbs (like *think*, *regard* or *appreciate*) subjects show a strong tendency for SVO instead.

Some researchers deem SVO as more likely in the inception of language than SOV. Bickerton argued that the fact that most creoles — languages that emerge from extreme contact situations —exhibit SVO reflects the *default* basic word order that would be in place before other factors start to shape grammar (Bickerton [27]; see Chapter 2 for a thorough discussion).

Finally, while engaging with the terminological apparatus of formal syntax is not warranted here, it is worth mentioning the very influential theory by Kayne on the universal relative order Specifier-Head-Complement (which most of the times would be analog to S-V-O [127]). The fact that languages display other "surface" basic word orders is due to ulterior transformations on this fundamental state.

## 2.3 THE SYSTEM OF WORD ORDERS

Language is a system, thus it should not come as a surprise that the word order of major constituents predicts very well (and is predicted by) other linguistic features — specially other word orders. The uncanny alignment of languages in only a few very frequent word order combinations asks for a reconsideration of the scope of some of the ideas presented before.

Some of the initial ideas on word order patterns suggested that they might simply come into existence as an agglomerate of overlapping functions or histories between pairs of lexical categories (as in Fig 1. (b)), possibly due to the cross-linguistic tendency of some lexical categories to be marked or used in the same way as others [45, 126], or because they grammaticalize from a particular category while preserving the same relative position in the clause as their source [6].

1.

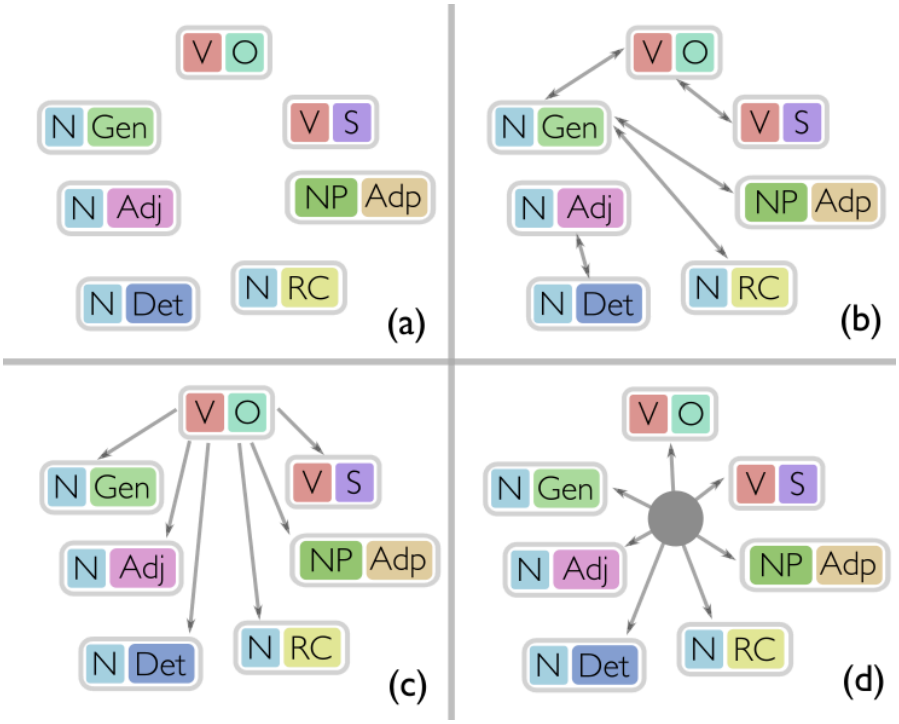


Figure 1: **Theories of word order**

Schematic representation of four classes of theories on word order patterns. (a) they are lineage dependent with no universal cognitive or functional value, (b) they are affected by each other in an unequal manner and each could be the locus of independent forces (as grammaticalization), (c) the adjacency between verb and object explains the patterns (d) a unique external cause affects some or all of them, as for the instance the preference for consistent head ordering.

Crucially, it should be noted that the discussion of the previous section can be perfectly framed as an opposition between OV and VO languages. The S-initial position is irrepressible — Table 2.2 shows that the ratio between S-initial and non-S-initial is about 8:1. I have mentioned that the reason behind this seems to be the widespread preference for interpreting the first noun or NP in a clause as S. This warped expectation is so strong that it might bend the interpretation of passives (in which the usual relative order of S and O is inverted), which leads young infants to interpret them as active sentences and it might even trigger a whole language to develop an ergative alignment system. On top of that, S is regularly a dispensable element — it can be omitted in many languages [134].

Following the discovery of the utility of the OV-VO opposition, Lehmann [135] and Vennemann [235] proposed a number of generalizations. While a few researchers have argued against the OV-VO typology [177], the range of typological variables it bundles together seems to outfox any available alternative [65].

Lehmann [135] suggests that the adjacency of O and V is the principal building block of the other word order patterns, what he calls the *Fundamental Principle of Placement*. Thus, if a category that modifies or complements V (like negation or adverbs) have to be allocated in a OV language, they will come to the right of the verb, and to the left in VO languages. Similarly, words or phrases that act upon O (like genitives or adjectives) will show the reverse pattern. This stems from the fact that, from a syntactic point of view, O and V are strongly and symmetrically dependent on each other [106] to the point that some languages directly forbid any intervening elements between the two [230]. Thus, the OV-VO disctinction organizes the rest of the relations (see panel (c) in Figure 1)

Vennemann [235] points out to structural analogies in the patterning. Roughly speaking, word order relations usually involve two categories with unequal degrees of dependency — recall our discussion on dependency length minimization in the previous section. The *dependent* category extends or makes more precise the meaning of the *head*; thus we have the dependent-head pairs O-V, Adj-N, Det-N, Num-N, among others. Vennemann proposed under the name of *Principle of Natural Serialization* that the consistency of word order is due to the harmonic arrangement of words into two classes: head-dependent and dependent-head, which correspond to VO and OV.

While elegant, this account does not fare very well against data. In an attempt to salvage the idea, Hawkins [106] suggested that the preference for head-dependent or dependent-head languages is, rather than categorical, a gradient force — languages can handle a mixture of both patterns, but the less harmonic they are, the less likely to be become fixed due to the extra strain. In both Hawkins and Vennemann's approaches, the OV-VO distinction is just one out of the

many predicted by the Principle of Natural Serialization without any special status (see panel (d) in Figure 1).

Finally, a more general criticism to the validity of synchronic word order patterns as cognitive or functional phenomena got its most recent exemplar in work by Dunn et al. [66]. If there exist real pressures pushing languages towards specific combinations of word order, this should appear most clearly in the diachronic dimension: a change OV→VO should be followed by its putative associated variable, Pos→Pre. They perform a standard Bayesian model comparison between two evolutionary models: one in which each word order pattern is free to change independently from the other, and another where they are correlated. They did not find evidence for universally valid correlated changes, which lead them to conclude that no word order correlations exist beyond the scope of individual lineages. This lead them to conclude that

> [...] systematic linkages of traits are likely to be the rare exception rather than the rule. Linguistic diversity does not seem to be tightly constrained by universal cognitive factors specialized for language. Instead, it is the product of cultural evolution, canalized by the systems that have evolved during diversification, so that future states lie in an evolutionary landscape with channels and basins of attraction that are specific to linguistic lineages.

Thus, no fundamental word order correlations exist beyond the accidental and contingent events of human history, as reflected by panel (a) in Figure 1.

## 2.4 WORD ORDER CORRELATIONS: FACTS, CHANCE OR STATISTICAL CHIMERA?

In a nutshell, we are a far cry from having any statistically conclusive evidence on any of the four theories presented so far.

First, word order patterns are transmitted vertically (i.e. from the previous generations of the related population of speakers) and they are susceptible to change due to contact with other languages, which results in apparent large areas with rather homogenous word order features [62, 225]. Because of this, it is a possibility that universal tendencies are the residual of particularly prolific lineages or ancient large-scale contact events. Making the case for universal cross-linguistic tendencies requires the assessment of as much independent evidence as possible, but the actual distribution of languages is not balanced across the putatively independent genealogical and areal units. When the data for individual units is scarce, it might lead to either the rejection of truly universal tendencies or the adoption of spurious associations.

For instance, Dryer's procedure for establishing a word order correlation consists in counting the proportion of genera within six large linguistic areas — if all of them share the same bias, then a new correlation is established. When analysing the distribution of NGen in SVO languages, Dryer counts the evidence for Africa (10 Gen(itive) N(oun) and 34 N(oun) Gen(itive)) as equivalent and opposite to that of North America (2 GenN and 5 NGen) [65]. A simple (two-tailed) binomial test reveals that, while the first bias is supported at conventional values ($p < .001$) the second effect is indistinguishable from the no-bias situation ($p = .45$), so definitive conclusions cannot be drawn.

Similarly, while Dunn et al.'s approach reveals important aspects of the independent historical development of individual linguistic families, it is limited by the absence of enough variation in word order patterns within them [46]. In addition, the fact that four linguistic families — no matter how big — fail to conform to a global statistical tendency is not enough to disqualify any strong informative generalization about the languages of the world [23]. Conversely, when the pattern being considered comprises several possible levels, unattested instances cannot be deemed to stand for impossible languages since the sample size required for ruling out them with confidence exceeds the amount of data available. For instance, the arrangement of Adj(ective), Num(eral), Dem(onstrative) and N, admits 24 logically possible combinations [41]: even in the unrealistic case in which all of the combinations have the same a priori chance of being attested, and assuming a sample of 50 completely independent languages, we would still expect to miss about 3 combinations.

Some contemporary methods can accommodate some phylogenetic information while being statistically sound and generalizable to a large number of language families and isolates. Bickel and collaborators developed a method known as "Family Bias" [24]. Binomial tests are performed within linguistic families —and with some adjustments, to isolates— and those that succeed in showing a consistent bias are marked as such. The method can accommodate specific information about the phylogeny of the involved languages as well, and it constitutes a dramatic improvement with respect to the preceding tradition. Other regression-based approaches are mixed effects models (where areal and genealogical information can be structured as random effects [117]) and (if more precise genealogical information is available) phylogenetic generalized least squares, where branch length is mapped into the covariance matrix via a model of neutral evolution (cf. [169]). Finally, a fully Bayesian approach to the issue has been attempted as well [49].

Second, since these proposals were all formulated on the basis of similar empirical observations, they coincide to a large extent on the expected pairs of associated word orders [6]. Furthermore, some of

the more technical notions involved in those theories — like "phrasal head" or "fully-recursive category" — provide extra degrees of freedom to accommodate discrepancies [209], and even clear violations are justified in ways that are post hoc if correct. For instance, the absence of patterning between OV/VO and Num, Dem and Adj (that conflicts with the head ordering processing theory) can be argued to be the result of the later categories composing characteristically shorter phrases — more concretely, Num, Dem and Adj are argued to be usually single words — which would not impose a large enough burden to the parser as to force a word order change according to the general preference [106].

Third, most research focuses on pairwise dependencies between patterns. The only explicit testing of the relevance of higher-order dependencies was attempted by Justeson and Stephens [126]. Based on 147 languages, they performed a log-linear modelling of the joint probabilities of six word order patterns, and they equated the specification of different interaction terms to the explicit theories of Hawkins, Greenberg and Lehmann. The conclusion they reach is that Greenberg's take, although probably wrong about the specifics of the correlation pairs, was essentially right about the fact that there do not seem to be reasons to posit higher-order dimensional dependencies beyond the pairwise levels. While ahead of its time, Justeson and Stephens paper faces two important limitations. First, the number of languages is too small to yield reliable estimates on the higher-order coefficients. Second, their model selection procedure is (presumably) based on a hierarchical partition of the interaction terms, which is arbitrary in principle and can lead to different best models depending on it. Finally, their calculations are based on likelihood ratio tests assuming the $\chi^2$ asymptotic distribution as in Wilk's theorem, which might be inadequate due to poor sample size or the violation of the assumptions of the theorem.

If this statistical approach seems to favor only low-order interactions, on the contrary, "universal" trends are easier to find when more variables are taken into account. Consider for instance the *postpositional noun modifier hierarchy* [105], which states that (in Hawkins' notation)

$$\text{Post} \supset ((\text{AdjN} \cup \text{RCN} \supset \text{DemN\&NumN}) \& (\text{DemN} \cup \text{NumN} \supset \text{GenN}))$$

The problem with these universal claims is that the number of conditioning variables diminishes dramatically the subsets for which the statement is relevant. In this case, the postpositional noun modifier makes a maximal statement about languages that have postpositions, that are AdjN, RCN and DemN. The result is usually a dramatic reduction of scope and increased concerns about statistical validity.

Ultimately, while word order patterns occupy a central role in our understanding of linguistic behaviour, these issues hold us from de-

ciding which of the competing theories of word order patterns (if any) explains the available data better.

In spite of all these limitations, it is possible to build a comprehensive statistical evaluation that 1) models the complex dependencies between word orders while 2) taking into account the individual genealogical and areal histories of languages using 3) fleshed out predictions that bind synchronic patterns with diachronic processes. I provide such an evaluation in the following sections.

## 2.5 MATERIALS

Word order data from 853 languoids was collected from WALS,[4] covering 742 unique languages, all continents, 161 Glottolog families, 275 WALS genera and 24 AUTOTYP areas. As it is usually the case, the relevant groupings are unbalanced: 55% of the genera have only one language, and 15% only two. A total of eight word order variables were registered: order of verb and object (OV-VO), subject and verb (SV-VS), genitive and noun (NGen-GenN), adposition and noun phrase (pre-post), noun and adjective (AdjN-NAdj), demonstrative and noun (DemN-NDem), numeral and noun (NumN-NNum) and noun and relative clause (NRel-RelN). The sample contains only languages marked with one of the two canonical orders; those with other values (like simultaneously having both NRel and RelN or inpositions instead of pre- or post-positions) where excluded beforehand. While there are other sources for word order patterns —- even WALS have over two dozen of such variables — the variables chosen for the analysis are the most widely discussed in the literature, plus they all have a decent coverage and enough observed variability, which is a sine qua non for any ulterior inference we might want to attempt.

The recorded variables vary in their coverage (see table 2), ranging from an almost perfect coverage of OV-VO (96%) to a rather poor 63% of NRel-RelN with a mean of 87%.

A map displaying the geographical distribution of the languages can be observed in Figure 2

## 2.6 THE INFERENCE OF DEPENDENCIES

### 2.6.1 *Directed Acyclic Graphs*

The first fundamental step in the direction of understanding how word order patterns associate with each other is to choose a flexible yet informative model in which the data could be represented. Ideally, such a model will permit me to approach the question of which are

4 All of the WALS chapters used here were authored by Matthew Dryer.

| Word order pattern | Coverage | WALS code |
|---|---|---|
| Object and verb | 0.97 | X83A |
| Subject and verb | 0.96 | X82A |
| Adjective and noun | 0.94 | X87A |
| Genitive and noun | 0.88 | X86A |
| Numeral and noun | 0.87 | X89A |
| Demonstrative and noun | 0.86 | X88A |
| Adposition and noun phrase | 0.81 | X85A |
| Relative clause and noun | 0.63 | X90A |

Table 2: Data coverage and WALS code of the word order patterns used in this study

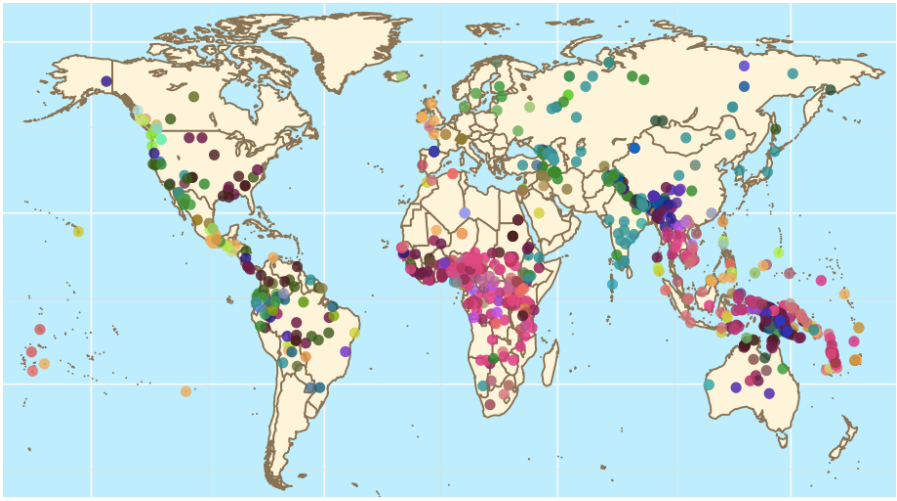

Figure 2: **Map of languages with word order information**

Geographical distribution of 853 languages for which word order patterns information was available. The colouring correspond to the projection of the first three principal components of their word order variables into RGB coordinates.

the causal dependencies among the variables $Y$, which will amount to estimate the objects

$$\Pr(Y_k = y_k || Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}, Y_{k+1} = y_{k+1}, \ldots, Y_N = y_N) \quad (1)$$

The elements on the left of the $||$ glyph need to be understood as indicating that the values those variables express have been achieved by means of an *intervention*. This stands in opposition to the analog observational object, simply the conditional dependency relation

$$\Pr(Y_k = y_k | Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}, Y_{k+1} = y_{k+1}, \ldots, Y_N = y_N) \quad (2)$$

The difference between the two is rather subtle: while in principle both are statements about random variables, in the second case the

variables stand for observations in the world, whereas the first captures the probabilistic effect of of supernaturally changing the variables on the conditioning side to specific values. To illustrate this in our case,

$$\Pr(\text{OV}|\text{NGen}, \text{postpositions})$$

could be approximated by the ratio of the observed languages in the world with NGen and postpositions that are OV as well. In contrast,

$$\Pr(\text{OV}||\text{NGen}, \text{postpositions})$$

would be estimated by forcing languages to adopt NGen and postpositions and then (perhaps after waiting for them to become stable languages) calculating the fraction of those that happen to be OV.

The model I will employ here is a *Directed Acyclic Graph* (DAG), a graph structure with enough flexibility to accommodate all plausible dependency relations which at the same time provides powerful statistical ways to cross (under appropriate conditions) the line from observational to interventional statements [210].

Let us assume that underlying the data exists a causal model for which the exact causal dependencies can be defined. Variables are displayed as nodes and causal connections as directed edges — so an edge going from variable $A$ to $B$ should be read as "$A$ causes $B$". In this case, $A$ would be the *direct* cause of $B$, but any variable that is connected via a series of directed edges to $B$ would be also regarded as a cause of $B$ — an *indirect* cause (see Figure 3) All the variables that are direct causes of $A$ are its *parents*, $\text{PA}(A)$, and the collection of all the variables that are either direct or indirect causes of $A$ are its *ancestors*, $\text{AN}(A)$. Conversely (and following with the kinship metaphors) all the variables for which $A$ is their direct cause are referred as its *children*, and in general all those variables that have $A$ as their direct or indirect cause are its *descendants*. Naturally, all those variables that are not the descendants of $A$ are its *non-descendents*. Figure 3 illustrates the nomenclature and the basic concepts introduced here.

We will let aside directed cycles from our representation, since they imply that causality could be reflexive[5]. The class of graphs that reflect these properties are called Directed Acyclic Graphs (DAGs).

So far we have not discussed in which way causal relations are expressed explicitly. For that we need the *Causal Markov Condition* [190, 210], which simply states that a variable is *independent* from its non descendants by conditioning on its parents. In other words, if we fix in some way the causes of a variable then the variation of the values of that variable can only affect those which are its descendants. This is an explicit bridge from a causal theory to (statistical) independence.

5 Models that involve reflexive causal relations exist but they are far less understood (from an algorithmic and formal point of view) than models without that property.

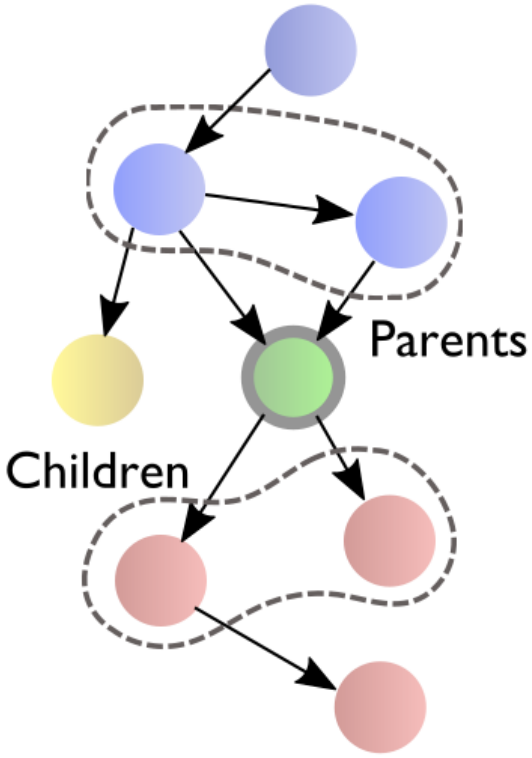


Figure 3: **Basics of a DAG**

Scheme of a directed acyclic graph (DAG). In the perspective of the green node, nodes that have ongoing / ingoing edges to / from it are called parents / children. The set of all the nodes to which there exists a directed path to / from the green node are referred as ancestors (in blue) / descendants (in green).

The Causal Markov Condition then determines the causal relations that can be read off the causal graph. Explicitly, it implies that the joint probability distribution of all the variables in the graph is simply

$$\Pr(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{N} \Pr(X_i | \textsc{Pa}(X_i)) \tag{3}$$

Which in other words means that the set of causes of a variable determines its probability distribution. Needless to say, this general framework can suit any probabilistic functional dependency between an effect and its causes. A variable $X$ could be a simple linear function of its parents with normally distributed noise (which is perhaps one of the most common modelling choices) but any mixture of distributions depending on any kind of function can be plugged in as well. Naturally, a purely deterministic function can be modelled in this manner as well.

More in general, conditional dependency relations between any set of variables can be easily deduced from the graph thanks to the

Markov Causal Condition by means of the graph-theoretic concept of *d-separation*, [190]. In this context, if a set Z of variables d-separates (d-connects) another, then the latter group of variables is conditionally independent (dependent) by conditioning on Z. The final relevant terminological ingredients are the *v-structures* (node triads following the scheme $A \rightarrow B \leftarrow C$, where the central node is referred as a *collider*) and *unblocked path*, a path that does not contain any v-structures. D-separation is composed by a few simple rules, which I will express by considering the simple case of two nodes X and Y:

- If there is an unblocked directed path between X and Y in the DAG, X and Y are d-connected.
- X and Y are d-connected by a set Z if no blocked path connecting them contains a v-structure that traverses a member of Z. If there are no blocked paths between X and Y, then Z d-separates X and Y.
- If a collider or a descendant of it is in Z then all the paths it traces should be regarded as unblocked

We will see that the inference of the quantity expressed in equation 1 is hard, but that we will be able to secure some information about the causal dependencies, namely which variables affect which others (this is, without a precise estimate on the magnitude of the influence) by assuming that the variables we used achieve *causal sufficiency* (see below). Even when this does not hold, DAGs still express the best statistical guess about the causal dependencies between variables: their conditional dependencies.

### 2.6.2 *From observational data to causal graphs*

We are already equipped with a natural model for causal relations. The challenge is now to infer causality out of purely observational data. The link comes from merging the graphical representation discussed before with statements about conditional independence between variables as they are inferred from data.

As we have seen, the causal graph model entails a series of conditional dependencies. Crucially, given proper conditions it is possible to invert the logic and go from a series of judgements about the statistical independence of a series of variables — which is in principle obtainable from observational data — to the actual causal relations in our framework. This last inference is possible once we assume the *Causal Faithfulness Condition*, which states that all of the independence relations that can be read from the data are represented in the graph and vice versa[6].

6 Concretely, it has been proved that in the space of parameters of multinomial distributions on a graph, non-faithful distributions have measure zero — in other words,

A naïve strategy would be to test, for each pair of variables, whether they are dependent after conditioning on every other possible set of variables. If two variables that are dependent remain so after these tests, then by the structure of the causal graphs the only alternative is for them to be connected in the graph. A smarter strategy is implemented in the PC algorithm (which we adopt in our analysis; Spirtes et al. [210]), which essentially determines the same facts by performing a smaller number of tests.

The roundabout strategy of the PC algorithm consists in considering sets of neighbors of larger cardinality sequentially. It starts with a complete graph. First, marginal conditional independence tests are performed over each set of variables. For each pair of nodes that turn out to be independent, their edge is removed. Then, for each of the pairs that are still connected, independence is tested again but this time conditioning on a third variable. Instead of trying with each variable, the PC algorithm considers only those nodes that are still connected to the target nodes. The reason is simple: if there is any set of nodes that d-separates both nodes, it has to include at least one direct neighbor of each. The process thus continues with subsequent conditioning sets[7].

The output of any of these procedures is called the "skeleton" of the causal graph, which is an undirected graph [210]. Given a conditional independence oracle, the PC algorithm is guaranteed to retrieve the true underlying skeleton and a class of compatible directed edges, as I discuss later.

Once the skeleton of the causal graph has been obtained, the algorithm proceeds on orienting the edges. Given three nodes connected by two edges ($X - Z - Y$) we can have 4 oriented graphs, as shown in Figure 4.

Two of those graphs correspond to a directed path from one extreme to the other: $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \leftarrow Y$. This structure correspond to the case in which all the influence from $X$ to $Y$ (or the other way around) is mediated by $Z$. Another possible graph is the so-called fork, $X \leftarrow Z \rightarrow Y$, which corresponds to $Z$ being a common cause of both $X$ and Y. In all of these cases, if we intervene in the world and remove $Z$ from the system, then $X$ and $Y$ cease to be connected causally. In particular, if we condition on $Z$ we obtain that $X$ and $Y$ are independent. If we actually find that this is the case (namely that $X$ and $Y$ are independent given $Z$) then we cannot tell which of the three cases represent the data we observe. The last possible directed graph based on three nodes and two edges is $X \rightarrow Z \leftarrow Y$, usually referred as a v-structure. Here $X$ and $Y$ are in-

there is a zero chance of picking randomly any distribution consistent with a graph and finding that is not faithful [157]

7 It should be noted that this entails that the variables to be tested depend on the order on which the analysis is performed. Current implementations of the PC algorithm solve this issue [42].

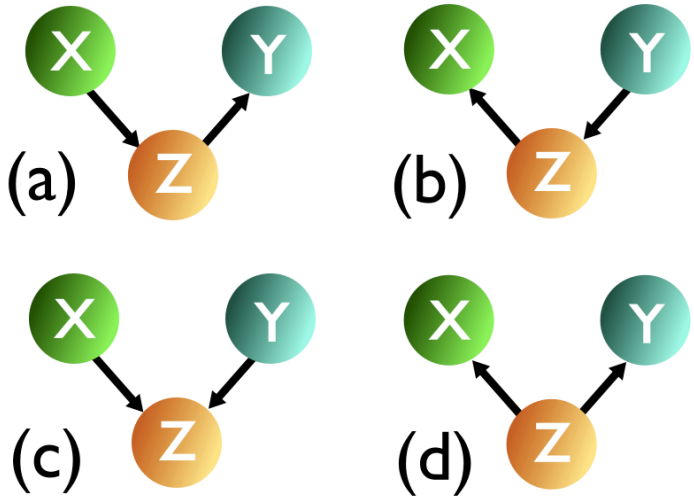


Figure 4: **Node triads**

All possible four orientations of two edges and three nodes. In the context of DAGs, (a), (b) and (d) imply the conditional independence of X and Y with respect to Z. In contrast, in (c) X and Y are independent but become dependent by conditioning on the levels of Z. This different behavior is key for linking conditional dependency statements and (unobserved) structural properties of the DAG.

dependent but contribute to the variable $Z$. Now, if we condition on $Z$, then $X$ and $Y$ become dependent. If we do find this phenomenon, we are in presence of a v-structure, and hence we can induce the orientation of the edges.

In this way, orienting v-structures can help orienting adjacent non v-structures. For example, imagine that there is a collider detected $X \rightarrow Z \leftarrow Y$, and the skeleton connects another variable $A$ to $Z$. If there is no collider detected, $A \rightarrow Z \leftarrow Y$ or $A \rightarrow Z \leftarrow X$, then by exclusion the orientation must be $Z \rightarrow A$. Once these orientations have been established, further directed edges can be determined, for instance when one of the possible orientations leads to a cycle — which we recall are excluded from our representation. Resolving all orientations in this way may not be possible for every skeleton. In other words, the independence statements usually underdetermine the causal graph. The final product of the algorithm is not, then, a simple causal graph but a collection of possible DAGs compatible with the data, with both oriented and non-oriented edges [210].

The DAG generated with the PC algorithm can be proven to retrieve the real underlying causal graph if we are equipped with the real dependency relations and, critically, if the set of variables are *causally sufficient*, which means that all the causes have been captured in our set of variables. Critically, DAGs are not closed under marginalization — DAGs inferred after removing or including one

variable do not hold any trivial relation with the DAG based on the original data. Furthermore, selection variables (which can be equated to an unbalanced sampling of data in the space of their covariates) can alter the dependencies.

Under appropriate yet stringent conditions, there are ways of inferring relations even in the presence of an arbitrary number of selection and hidden variables. However, for my research question this is irrelevant. As we have seen, three out four theories of word order patterns lead to contrasting DAGs even with a limited number of variables — following the panels of Figure 1, (a) would imply no dependencies, (b) diverse pairwise dependencies and (c) dependencies involving mostly OV-VO. I will make use of a different strategy for the assessment of (d), that explicitly appeal to factors outside the observable word order patterns.

### 2.6.3 *Results*

In order to account for genealogical dependencies, I repeat $n = 5000$ times the DAG inference procedure by sampling randomly one language per WALS genus — which are defined so to have comparable time depths. Each random sample will have some missing values. Three different approaches were used for missing data imputation. In order of complexity, they are:

- Random imputation. Missing values are assigned randomly to either of the two levels of the variable,

$$X_{\text{miss}} \sim \text{Bernoulli}(0.5) \tag{4}$$

  This scheme treats missing values as effective noise. While missing potentially relevant information that might conduct towards a more efficient imputation, it also serves as a baseline condition: any pattern found under this procedure can be considered as firmly established.

- Bernoulli imputation. Similar to the previous scheme, but with the parameter of the Bernoulli trial estimated from the observed values

$$X_{\text{miss}} \sim \text{Bernoulli}(\hat{p}_{\text{obs}}) \tag{5}$$

  The observed parameter of the Bernoulli distribution was estimated using a simple Laplace rule of succession.

- Gibbs sampling imputation. This is a Bayesian approach that estimates the cells of the contingency table by means of a Dirichlet process mixture of products of multinomial distributions [207]. The prior has full support on the space of possible distributions

consistent with the observed data, and in the case of finite mixtures the problem can be solved via maximum-likelihood - indeed, that will be the strategy I will be follow in Chapter 2 under the name of *latent class analysis*. The infinite mixture model I use here, while complex, it is nonetheless a classic Bayesian approach — uninformative priors, Gibbs sampling via MCMC, etc. — and it has the purpose being as flexible as possible.

The sequence of imputations schemes also follows the degree on dependency on the rest of the data: Bernoulli imputation does not depend on the data at all, Bootstrap imputation depends on the distribution of values of the imputed variable and Gibbs imputation depends on the information present in the joint probability distribution of all variables.

As independence test I use a simple $G^2$ test. It involves the calculation of the statistic

$$G^2 = 2 \sum_{x,y,S} \Pr(x;y;S) \ln \left( \frac{\Pr(x;y|S)}{\Pr(x|S)\Pr(y|S)} \right)$$

where $S$ is any set of variables other than $x$ and $y$. This quantity has an asymptotic $\chi^2$ distribution in the independent case, with appropriate degrees of freedom. As it was mentioned before, we have to make sure that the amount of data is sufficient for such an approximation; a common rule of thumb asks for no less than $10 \cdot 2^{|S|}$ data points, where $|S|$ is the size of the conditioning set. This implies that we will be able to condition on no more than 4 variables at the time — conditioning on 5 variables requires a nominal number of 320 data points, slightly off the 275 independent points from each genera. However, in the procedure it was never encountered the need of conditioning on sets larger than 4.

Given the asymptotic $\chi^2$ distribution, we proceed to test the competing hypothesis of our observed $G^2$ not being a sampled value from it at a fixed $\alpha$ value. While formally a level for significance, in this case it should be regarded more as a regularization parameter. In general, it is expected that the set of edges inferred from smaller values of $\alpha$ will be contained in the DAGs for smaller values. In this case, the fact that multiple random samples were integrated allows us to have a natural estimate of the robustness of the inferred edges. Heuristically, it can be seen that changing $\alpha$ changes the support for a particular in a more or less continuous manner. $\alpha = 0.001$ was chosen since it does not yield more than a couple of saturated edges — this is, edges that are inferred in *all* the samples — while being able to capture even very weak links at the same time. In the end, the discussions will be based on the relative support across the random samples.

The output of the PC algorithm assuming a cutoff of 0.001 is presented in Figure 5.

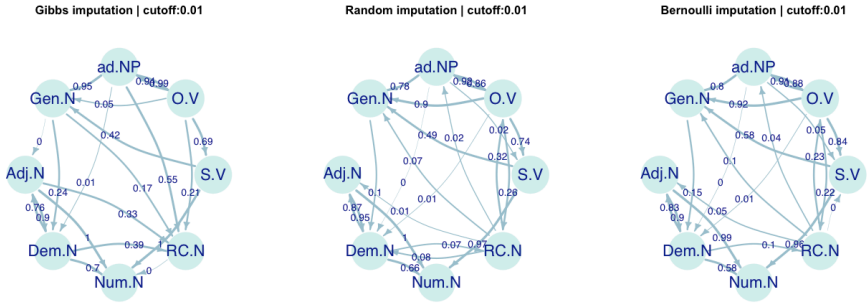


Figure 5: **Word order DAGs with different imputation methods**

DAGs from the PC algorithm applied to 5000 random samples based on three imputation methods (discussed in the main text). The numbers attached to the edges represent the fraction of samples in which the edge was inferred.

The graphs turn out to be quite dense, although a considerable fraction seems to be supported only by a small number of samples. The graphs involving edges appearing in at least 2/5 of the samples is simpler to interpret, as it can be seen in Figure 6

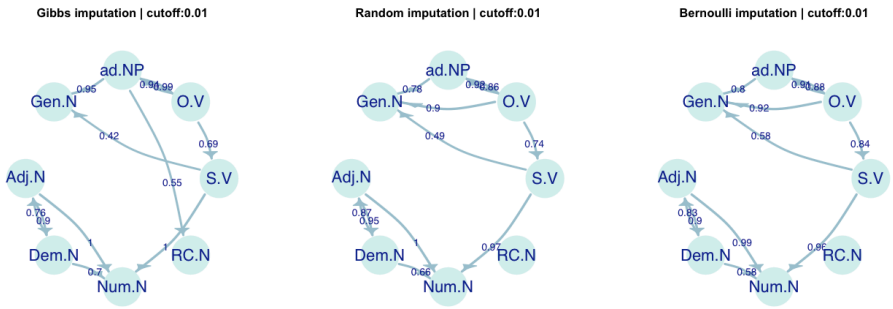


Figure 6: **Word order DAGs with different imputation methods (only well-supported edges)**

DAGs from the PC algorithm applied to 500 random samples based on three imputation methods (discussed in the main text). The numbers attached to the edges represent the fraction of samples in which the edge was inferred. Only edges with a value over 2/5 of the total number of supporting samples are displayed.

The output of the three methods are almost identical, as it can be seen. The random and Bernoulli strategies show the same edges (and directions), whereas the Gibbs' imputation method differs with respect to them in two respects: it has a directed edge from adNP to RCN and it lacks the directed edge binding OV to GenN. This is not entirely surprising since the Gibbs' sampler exploits the structure between the variables, thus inflating the dependencies found in

the fraction of observed data. Unsurprisingly, the changes imply the three variables with the poorest coverage in the data.

From now on, we will focus on the analysis of the DAG inferred via the Bernoulli method, which makes use of the information present *within* variables without building in the observed dependencies *between* variables. The edges with the largest representation with this method can be observed in Figure 7

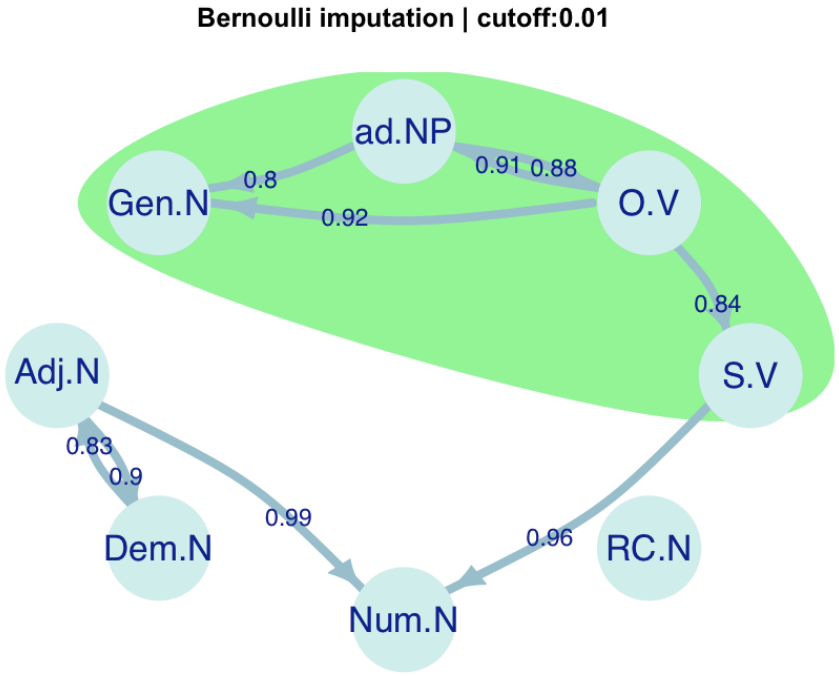


Figure 7: **Close-up of Bernoulli-imputed word order DAG**

The DAG composed by the bootstrap imputation method is shown. The numbers attached to the edges represent the fraction of samples in which the edge was inferred. Only edges with a value over 2/3 are displayed.

In agreement with the literature, we find that a cluster of word order patterns seems to accommodate to OV-VO to some extent: SV-VS, GenN-NGen and pre-post. Relative clauses might be related as well, but the presence of links involving it have a low support (most likely due to missing data). The other observed edges involve a reciprocal relation between DemN-NDem and AdjN-NAdj, and NumN-NNum with SV-VS and AdjN-NAdj. These results strongly argue against a scenario where OV-VO determines the overall structure of word order patterns, and it is compatible with the Greenbergian "pairwise only" stance. As for the case where no real dependencies exist in the data, it would be still possible to argue that beyond genealogical control, the observed dependencies come to be as a result of areal diffusion. Let us flesh out more precisely how this argument could

be articulated here: word order patterns that appear as dependent in our analysis could be conjectured to be simply the result of a few bundles of properties being particularly frequent in some linguistic areas. This would mean that beyond those areas there will be no support for such dependencies.

As a manner of testing precisely this, I repeated the inference of the DAG but this time randomizing the word order pattern values within languages belonging to the same area (according with the AUTOTYP classification). If the hypothesis to evaluate is that the patterns are due to areal spread as described above, then this randomization test should yield results similar to those of the original analysis. Randomizing areas where no such bias exist should not be relevant to the overall dependency, and the same process over areas where a particular set of features has spread over should reflect a degree of homogeneity that would be unchanged under this procedure.

It should be noticed that this procedure is likely to overestimate the areal influence since, for many linguistic areas, there will be no effective change at all — for some of them in the Americas, Australia and Asia there are no more than five languages. Furthermore, if the regularities I found are really areal, then the randomization process should weaken the influence of the covariates, thus enhancing the chance of finding the relation.

I ran the randomization process 100 times and compared the fraction of samples in the output against the empirical values (without randomization). The results can be observed in Figure 8

A number of telling conclusions can be drawn from these results. All in all, it is clear that (apart from the link ad.NP $\rightarrow$ O.V) the per area permutations yield support values comparable or larger to those in the original data in less than 5% of the times. Under the hypothesis that there is no cause behind word order patterns apart from the areal spread of certain features, we would expect much less extreme values in general. Using corrected P-values and the Benjamini-Hochberg method -both conservative choices- the expected FDR turns out to be 0.18 for all of the links and less than 0.034 when ad.NP $\rightarrow$ O.V is removed. What this implies is that, even after the sequence of decisions that inflate the number of false negatives, we would have expected less than two out of nine links found in the DAG as the result of a purely areal effect. We can, thus, safely reject contact as the sole explanation.

The links involving OV-VO are the ones for which the random permutations yield a small yet significant fraction of cases with comparable support. This situation can be opposed to what is found with links involving the order of demonstratives and numerals. It is unclear why is this the case.

Finally, the link between OV-VO and SV-VS is the one with the poorest presence in the permuted samples. This seems to be a conse-

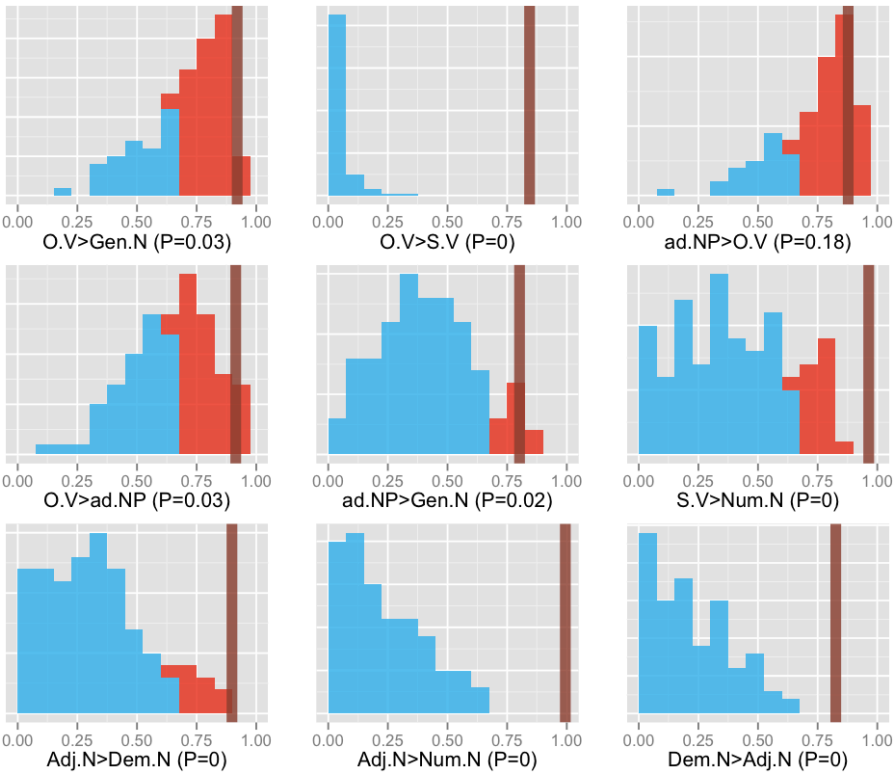


Figure 8: **Support for edges in DAG with randomizations per area**

Each panel displays the distribution of the support — the fraction of random samples — based on 100 random permutations of the word order patterns within linguistic areas.The fraction in the distributions in blue and red correspond to random permutations that yield less and more than 2/3 of support. The vertical line correspond to the empirical value. One-tailed P-values are included in the label of each graph.

quence of the subject preference: the presence of VS forces the object to come after the verb and the subject, conditions whose violation leads to some of the rarest word order patterns, VOS and OVS.

## 2.7 WORD ORDER ARCHETYPES

### 2.7.1 *Higher-order dependencies*

The causal graph of word order properties allowed us to retrieve the structure of pairwise dependencies, and we found evidence for the existence of non-trivial relations and a relevant yet not overwhelming role of OV-VO as an centralizing feature. We have practically ruled out the extreme scenarios represented by panel (a) and (c) in Figure 1.

The accumulation of pairwise dependencies is not transparent with respect to higher-order groupings of levels among the variables. Let us take the case of a particular value of the variable OV-VO: OV. OV determines adpN, GenN and, at a lesser extent, SV. This does not entail, however, that we expect to find two word order "archetypes" coinciding with each of the two possible orderings of the object and the verb. Each level can have a different predictive value, although overall OV has to be informative of the other variables *in toto*. In information theoretic terms, while the mutual information between two variables can be positive and of considerable magnitude, individual pointwise mutual information terms can perfectly be zero or negative.

The challenge is to find statistical evidence for word order archetypes in data. While for the analysis of two variables we count with several measures that let us inquire with precision potential dependencies, the panorama is more complicated with the inclusion of more dimensions. For example, in the "simple" case of three variables, given the joint probability distribution $\Pr(X_1, X_2, X_3)$ we can ask whether it is statistically distinguishable from any of the 4 possible factorizations of the variables that imply some independence: $\Pr(X_1, X_2)\Pr(X_3)$, $\Pr(X_1)\Pr(X_2, X_3)$, $\Pr(X_1, X_3)\Pr(X_3)$, $\Pr(X_1)\Pr(X_2)\Pr(X_3)$.

The inferred DAGs provide answers for at least the connected variables in the graph. Given any two nodes or variables $X_1$ and $X_2$ connected in the DAG and any of their neighbors $X_3$, by construction we know that $\Pr(X_1, X_2|X_3) \neq \Pr(X_1|X_3)\Pr(X_2|X_3)$ via the $G^2$ test. From this it follows that, in the same circumstances, $\Pr(X_1, X_2, X_3) \neq \Pr(X_1)\Pr(X_2|X_3)$. Naturally, given any third variable (namely, not necessarily a neighbor of $X_1$ or $X_2$), it is true as well that $\Pr(X_1, X_2, X_3) \neq \Pr(X_1)\Pr(X_2)\Pr(X_3)$

A more interesting question is how much do the relative orders of dependencies count in the overall joint probability distribution. In the framework of information theory, Amari has suggested to compare joint probability distributions against the maximum entropy distribution resulting from fixing its n-order marginals to the ones of the target distribution. Thus, for instance, $\Pr(X_1, X_2, X_3)$ could be compared against a distribution $\Omega(X_1, X_2, X_3)$ such that $\sum_k \Omega(X_i, X_j, X_k) = \Pr(X_i, X_j)$ for each assignment of the indexes. In this manner, if both the empirical and the maximum entropy distributions are statistically indistinguishable, one could determine that there is no information encoded in the joint probability distribution that it is not present in any of its (pairwise) marginals. These ideas, while promising, yet awaits for a better statistically grounded treatment.

### 2.7.2 *Latent class modelling*

The question of higher-order dependencies will be approached here by means of clustering analysis. While a large number of categorical clustering methods consist on the assignment of instances to classes, this is only of secondary interest in our case because the focus is on unraveling the archetypes themselves, independently of their particular instances. Furthermore, many of those methods use metrics between instances that are hard to motivate theoretically, and they cannot be assessed with model comparison techniques.

A worthwhile strategy is to model the data by means of a number of latent (unobserved) classes with the aid of a statistical suite of tools referred as Latent Class Analysis (LCA). In the most basic setting (and within the binary variables scenario we have investigated so far) this is equivalent to propose that the underlying generative model of the observed data is given by a a simple Bernoulli process independently for each variable ($W_i$) given the class $c_j$:

$$\Pr(W_1 = w_1, W_2 = w_2, \ldots, W_N = w_n | C = c_j) = \prod_{i=1}^{N} \pi_{c_j,i}^{I_i}(1 - \pi_{c_j,i}^{I_i}) \quad (6)$$

where $\pi_{c_j,i}$ is the probability of variable $i$ to have one particular level under class $c_j$ and $I_i$ is an indicator variable that equals one for the level with probabilities $\pi_{\cdot,i}$. Given a number of classes $K$, there are $K \cdot (N-1) + K - 1$ free parameters to be estimated - $K \cdot (N-1)$ variable probabilities and the $K-1$ values associated with the probability distribution of the classes, $\Pr(C)$. Thus, the instances can be thought as coming from a finite mixture model,[8]

$$\Pr(W_1 = w_1, W_2 = w_2, \ldots, W_N = w_n; C = c_j) = \Pr(C = c_j) \prod_{i=1}^{N} \pi_{c_j,i}^{I_i}(1 - \pi_{c_j,i}^{I_i}) \quad (7)$$

If the distribution of the classes $C$ was available, then the estimation of the binomial parameters for each class reduces to a simple maximum likelihood calculation. On the other hand, if we knew beforehand the parameters attached to each class then we could infer the mixture of classes in the data. Unfortunately, finding the maximum of both model and class parameters cannot be done analytically in the non-trivial case of $|C| \geq 2$.

Our data are not particularly intractable with brute search over a grid of sufficient precision over the parameter space, but the problem of missing data would still be present - this is not minor since, we recall, 13% of the data are absent from our database. One alternative

8 There exist infinite mixture models in which LCA can be casted into. Because these models bear no utility to the issues studied here they will not be considered.

would be to do multiple imputations as in the causal graph case, but a much better alternative exists that solves at the same time the problem of the efficient calculation of the maximum of the likelihood function and the missing data. The solution is the famous expectation-maximization (EM) algorithm[9].

The gist of the method resides in the convexity of the logarithm and a convergence theorem. Consider the log-likelihood function,

$$\ell(w;\theta) = \log(\sum_c \Pr(c)\Pr(w|c,\theta))$$

where $\theta$ summarizes all the binomial parameters of all classes. By Jensen's inequality,

$$\ell(w;\theta) \geq \sum_c \Pr(c)\log(\Pr(w|c,\theta)) = g(w;\theta) \tag{8}$$

So $g(w;\theta)$ is a lower bound for the likelihood function. On the other hand, given $\theta$ and $w$ we get

$$\Pr(c|w;\theta) = \frac{\Pr(w|c;\theta)\Pr(c|\theta)}{\ell(w;\theta)} \tag{9}$$

Starting from a (probably random) estimate of the parameters subject to the usual normalization constraints, the EM iterates between estimating $\Pr(c|w;\theta)$ and plugging it into equation 8 so to get an approximation of $\ell(w;\theta)$ by maximizing $g(w;\theta)$. The $\theta$ that maximize the previous expression are used again in equation 9, starting again the process. While in practice convergence is determined by tracking the differences in likelihood (or $\Pr(c|w;\theta)$), a theorem guarantees the convergence towards a local maximum of $\Pr(\theta|w)$.

### 2.7.3 *Determining the number of classes*

We have assumed throughout the discussion that the number of classes $|C| = k$ is fixed, but this is exactly one of the main aspects of the analysis - strong evidence for the existence of archetypes would be expressed as a few classes capable of satisfactorily capturing the variation in the data, whereas an "everything goes" situation should break down the data into a myriad of combinations with modest coverage. Fortunately, we can appeal to the residual likelihood functions estimated before and perform model selection across the dimension of number of classes.

Naturally, the larger the number of classes the better the goodness-of-fit metrics. However, if data is effectively explained by a few archetypes, the improvement in goodness-of-fit beyond their number

9 By the time I write this, the paper that formalized the technique — since it was invented a few times in different contexts— accumulates a whooping 42000+ citations. EM is the textbook successful case of marriage between frequentist statistics and computation.

would be marginal. Importantly, the number of classes can feature explicitly in the model selection. Two popular techniques for this use as a comparison the log-likelihood function with an additive penalization term: for the Akaike Information Criterion (AIC) this is simply $k$, and in the Bayesian Information Criterion (BIC) this is $k \log N$, where N is the sample size. Both have interesting properties and (perhaps in contrast to their similar functional form) they depend on markedly different takes on model selection. Because we expect a rather simple model out of the data, BIC appears as the reasonable choice [237].[10]

Let us take an Empirical Bayes approach to model selection as a way of motivating BIC, which is explicitly defined as

$$\text{BIC} = -2\ell(w;\theta) + 2\log N$$

Recall that the Bayes Factor (BF) is defined as

$$\text{BF} = \frac{\Pr(x|\mathcal{M}_1)\Pr(\mathcal{M}_1)}{\Pr(x|\mathcal{M}_2)\Pr(\mathcal{M}_2)}$$

which determines how much more likely a model $\mathcal{M}_1$ is given the data and priors with respect to an alternative model $\mathcal{M}_2$. Ideally, we would like to have access to the real model $\mathcal{M}_{\text{real}}$ so we could choose among our candidates models which is the one that minimizes its relative BF. In practice, we are left only with the likelihood functions of individual models, and we can hope the priors to be not extremely relevant in the case of large $N$.

Suppose now that the priors on the model parameters are modelled as a multivariate normal distribution,

$$\theta|M_k \sim \mathcal{N}(\hat{\theta}_{\text{ML}}, \hat{\Sigma}_{\text{ML}})$$

where $\hat{\theta}_{\text{ML}}$ and $\hat{\Sigma}_{\text{ML}}$ are the maximum-likelihood estimates of the mean and the covariance of the parameters, respectively. This is a fairly typical parametrization of uncertainty.

Given the following conditions,

- The true model is under consideration

10 On the BIC/AIC divide, Vrieze [237] says:

> In our experience applied articles rarely defend their use of a particular model selection criterion, but instead give a reference or two to seminal (and too often mathematically inaccessible) articles about it. The articles leave to the reader to determine why one criterion was used and not another. In an extreme example, we have recently seen an article that used four criteria (two of which were linear functions of each other), and stated that the model for which three of the four criteria agreed would be selected as the best model. This approach does not reconcile the differences between selection criteria (except in the case where all criteria agree). Reconciling the differences between AIC and BIC appears to be very diffcult, if not impossible.

- The dimension of the model is independent of the sample size $N$
- The number of parameters is finite

the BIC will asymptotically approach the $BF$ between the target model and the real model with error $\mathcal{O}(N^{-1})$.

I evaluated the BIC values associated to $C$ from 1 to 10 in 500 random genealogical samples of the data. The results can be observed in Figure 9

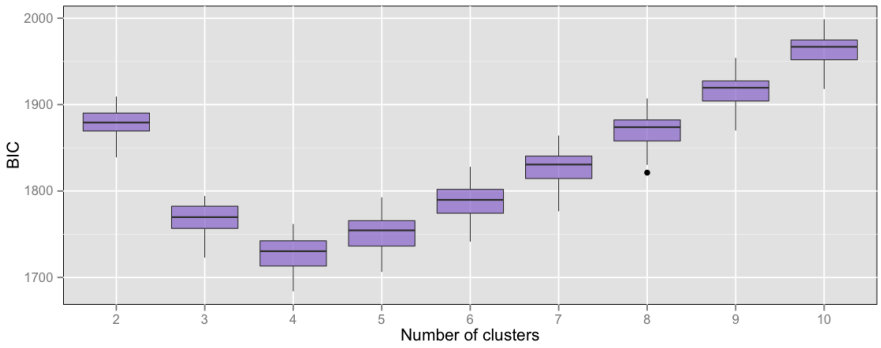


Figure 9: **BIC by number of clusters**

Boxplot of BIC values associated with random genealogical samples according to the number of clusters posited by the LDA. The BIC value associated to only one cluster is considerably larger than for any other value (∼2400).

$C$=4 appears as the best candidate, although the difference with $C$=5 and $C$=3 is not sufficient to embrace the idea that these four clusters are a "natural" description of the data. It is important to bear in mind that each genealogically balanced sample yields a different class profile. Although the distribution of the goodness-of-fit and penalized likelihood measures suggested that the overwhelming majority of them can be adequately described by roughly four latent classes, this does not guarantee the inferred classes to be comparable across samples. Firstly, even in the case of perfect equivalence, the stochastic nature of the EM algorithm might yield different label orderings to classes with strictly the same parametrization — a phenomenon referred as *label switching*.

Second, and more importantly, it might be that some (or all) classes are not robust at all. The last step in the evaluation of the existence (and composition) of word order archetypes is to gather evidence on the presence of four robust clusters of classes.

A superficial inspection of the data by comparing the Euclidean distances between the aggregated collection of four vectors of binomial parameters (for each of the 500 random samples) yields the heatmap in Figure 10.

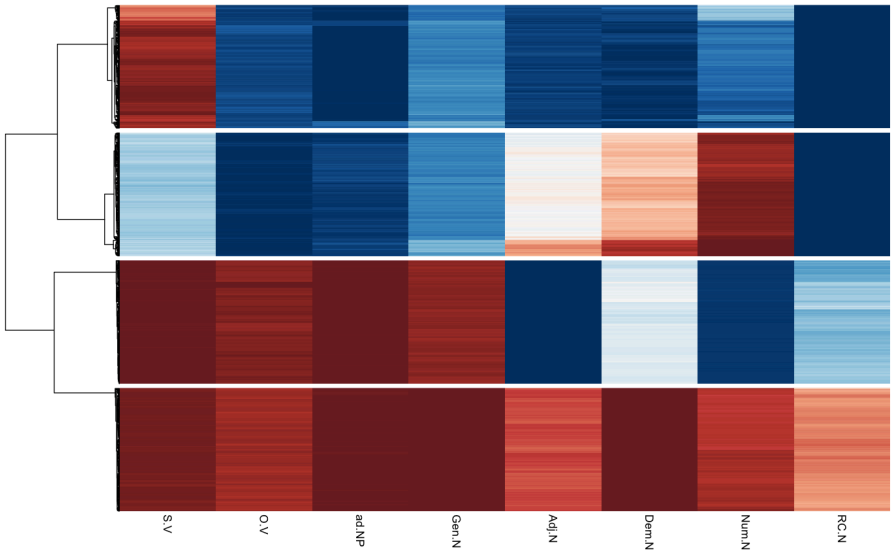


Figure 10: **Archetypes' clusters**

Heatmap based on the average clustering of archetypes' instances. More intense tones of blue imply larger chance of N or NP last (as in VO, prepositions, Dem N) and the contrary is true for red.

| Class | S/V | O/V | adp | Gen/N | Adj/N | Dem/N | Num/N | RC/N |
|---|---|---|---|---|---|---|---|---|
| 1 | SV | VO | pre | NGen | NAdj | NDem | NNum | NRC |
| 2 | VS | VO | pre | NGen | AdjN | DemN | NumN | NRC |
| 3 | SV | OV | post | GenN | NAdj | NDem | NNum | NRC |
| 4 | SV | OV | post | GenN | AdjN | DemN | NumN | RCN |

Table 3: Word order of the four inferred archetypes. In all of the cases but one - AdjN in class 2 - the probability of any of the archetypal word order patterns is such that the 95 % CI lies above 0.5.

There seems to be four coherent blocks of roughly equal size. If we chose to characterize the inferred classes in terms of the most likely word order patterns exhibited, we obtain Table 2.7.3. Interestingly, there are two sets of three variables that tend to change harmonically: one that integrates O/V, adp and Gen/N and a second that comprises N and Adj, Dem and Num.

More rigorously, we can characterize the statistical properties of the clusters. In principle, following the philosophy discussed early in this chapter, it seems honest to model the clusters with some degree of parametric specification. This seems appealing because, in fairly general cases, the maximum likelihood estimators are asymptotically normal. Because probabilities are doubly censored data — in the sense that they are contained in the closed interval $[0, 1]$ — and it is not unreasonable to expect some estimated parameters as being

strictly 1 or 0, appealing to simple Gaussian mixture models is not warranted unless we are willing to introduce these restrictions. I will explore other alternatives here.

A popular non-parametric evaluation metric for the number of clusters is the silhouette statistic. Given a fixed assignment of datapoints – latent classes in this case – to clusters, we can define the average distance of datapoint $i$ to the rest of the member of its cluster $C_i$ as

$$m_i = \frac{1}{|C_i|} \sum_{j \in C_i} d_{ij} \tag{10}$$

Similarly, define $n_i$ as the minimum average distance of a datapoint to the members of *another* cluster,

$$n_i = \min_{C_h \neq i} \left\{ \frac{1}{|C_h|} \sum_{j \in C_h} d_{ij} \right\} \tag{11}$$

then the ratio

$$s_i = \frac{n_i - m_i}{\max\{n_i, m_i\}} \tag{12}$$

$s_i$ ranges between 1 and -1, and it can be interpreted as how suitable the assignment of $i$ to its cluster is. The average of $s_i$ across all datapoints, $S$, is the silhouette statistics and it can be used to choose the most appropriate number of cluster in the data given a clustering method and a distance. With k-means and four clusters appears as optimal when compared to alternative $k$ clusters ranging from 2 to 10, yielding $S = .940$ with Euclidean and $S = .947$ with Manhattan distance, respectively.

In conclusion, we can capture a considerable amount of variation of the word order data by means of four word order archetypes. They are represented by different fractions of the population in each sample, as evinced in Figure 11

## 2.8 CONCLUSION

In the preceding sections, I have discussed in detail the analysis of word order patterns data in the light of four competing theories about their mutual dependencies: (1) there are no consistent relations between them beyond contingent associations due to cultural evolution (2) only pairwise relations can be accounted for, without any further structure (3) the adjacency of the verb and the object is the main structuring force and (4) a common external factor (like the consistent placement of the heads in relation to their dependents) harmonizes word orders. The analysis of pairwise dependencies via DAGs showed the existence of reliable links involving at least 9 (directed) pairs out of a total of 56. Those dependencies did not disappear even after conditioning on other neighbouring variable or variables, and

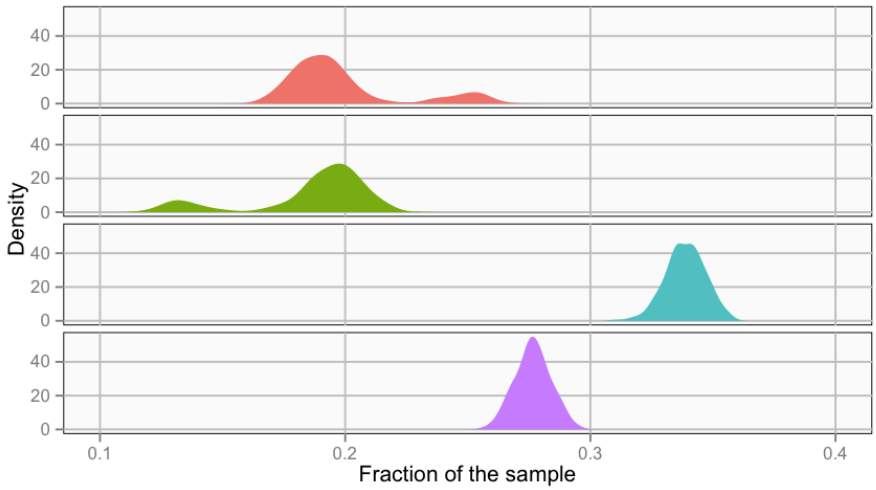


Figure 11: **Archetypes' populations**

Distribution of the populations of the four word order archetypes based on 500 genealogically balanced samples of languages.

the strength of their presence could not be explained by areal effects only. The pairwise dependencies might be part of higher-order dependencies; in particular, we find two sets of three variables that tend to be aligned, and four wide word order alignments (in contrast to the 64 logically possible combinations licensed by the 8 binary features). The hierarchical structure among those relations beyond pairwise dependencies does not specially highlight the role of OV/VO, nor it is strong enough as to imply a single latent variable behind them.

The resounding message is that, in agreement with claims from traditional synchronic typology, robust word order patterns exist beyond the effects of historical contingencies. The next natural question is: why? Probably there is no single factor determining these arrangements, and the synchronic developments leading to them might be as diverse as the actual attested cases. While the literature does not lack facile attempts at explanations based on loose processing or cognitive advantages, the fact that certain patterns are seen time and again lends itself to an explanation based on a common bias expressed by different mechanisms under different circumstances.

An apt analogy can be made with the ideas surrounding convergent evolution in biology [143]. Unrelated species might converge to the same traits or behaviours across the tree of life due to a number of causes, for instance as a consequence of common constraints that direct their evolution, limitations on the production of varieties or common selective pressures. A paradigmatic example of the later case is the development of wings: the (fairly distant) species of bats, birds, insects and Pterosaurus all evolved wings at some momentous point of their phylogenies, and they did it by means of sharply different

genetic, developmental and structural pathways — so for instance insect wings were exapted from membranes whose function was to cool down body temperature, whereas in bats they are the homologues of mammal legs. Still, in all of those species, wings allow for metabolically cheap transportation, a trait with a clear functional value that seem to have been the product of natural selection. It is perhaps too soon to declare that the explanation we are seeking is of a functional or adaptive flavour — the existence of 'non-harmonic' word order patterns clearly shows that languages do not require them in order to be suitable for their purposes. Nevertheless, given the saliency and the central role word order patterns occupy in the way speech is structured, they are among the strongest candidates for serious scientific inquire in this direction.

# 3

# NON-ARBITRARY SOUND-MEANING ASSOCIATIONS

## 3.1 INTRODUCTION

Although there is substantial debate in the language sciences over how to best characterize the features of spoken language, there is nonetheless a general consensus that the relationship between sound and meaning is largely arbitrary [110, 194, 204]. Plenty of exceptions exist, however, within individual languages. For instance, ideophones—a class of words found in many languages—convey a communicative function (or meaning) through the depiction of sensory imagery [55]. In the Mel language Kisi Kisi (spoken in Sierra Leone) *hábá* means "(human) wobbly, clumsy movement", and *hábá-hábá-hábá* "(human) prolonged, extreme wobbling"; here repetition serves as a way to convey the meaning of intensity. More generally, the resemblance between certain aspects of the acoustic basis of speech and their referents, *iconicity*, is the most researched and well-known case of non-arbitrary associations between sound and meaning [57]. *Systemacity*, in contrast, refers to (statistical) regularities that are common to particular set of words, created by historical contingencies and analogical processes [57]. For example, word-initial *gl-* in English evokes the idea of a visual phenomenon (as in *glare*, *glance*, *glimmer*) [18]. At a larger scale, there is evidence that the phonological properties of whole morphosyntactic classes of words (like verbs and nouns) are distinct in several languages [161].

The evidence of recurring regularities in sound-meaning mappings across multiple languages is considerably more modest, despite its potential importance for fundamental questions about language evolution and the role of basic perceptual biases in cognition. For example, certain shape-sound associations—known as the *bouba-kiki* effect [129, 154, 197]—are believed to rely on the ability that humans (and perhaps also other primate species [144]) have for associating stimuli across different modalities [47]. Other plausible sources of cross-linguistic associations include, for instance, the relationship across many animal species between vocalization frequency and animal size [109], the mimicry of referents via unconscious mouth gesturing [20],

and the persistence of vestiges of a conjectured early human language [114].

Experimental studies support the hypothesis that humans are indeed sensitive to such associations. It has been demonstrated several times that paticipants perform above-chance when asked to pair up words with opposite meanings (antonyms) in languages unknown to them [186], and that English speakers might even be able to decide on the concreteness of words from languages to which they have not been exposed [198]. However, this evidence for non-arbitrary sound-meaning associations pertains only to narrow pockets of the vocabulary, making it unclear whether a more general pressure towards arbitrariness may overpower such potential biases when considering a more semantically diverse selection of the vocabulary [110, 162].

A further issue with current studies of non-arbitrariness in sound-meaning correspondences is that, save for a single exception [239], cross-linguistic corpus studies of non-arbitrary associations have tended to rely on a small number of languages (maximally 200) and focusing on small semantically-restricted sets of words, ranging from phonation-related organs [232] to South American animals [20], to spatial orientation (demonstratives) [109, 124], repair initiators (like *huh?* in English) [56] and the conceptualization of magnitude in Australian languages [108]. These studies involve confirmatory analyses, aiming to test specific hypotheses regarding sound-meaning correspondences; as a consequence, they are guided by a priori intuitions or indirectly by findings from other disciplines. These limitations may help explain, at least in part, why language scientists typically consider non-arbitrary associations to be marginal phenomena that may only apply to small, strictly circumscribed regions of the vocabulary [194]. In this paper, we therefore conduct a comprehensive set of analyses involving a semantically diverse set of words from close to a two-thirds of the world's languages.

## 3.2 TESTING ASSOCIATIONS ON A GLOBAL SCALE

The availability of a large collection of word lists allows us to search for statistically robust associations in an unsupervised, theory-neutral manner. This collection is the version 16 of the ASJP database [70]. ASJP comprises 6895 word lists from around 62% of the world's languages, covering 85% of families, isolates, and unclassified languages (using the Ethnologue [138] for these statistics). To summarize genealogical relatedness, we introduce the notion of *lineage*: a maximal set of languages that can be shown to have a common ancestor. Such a set may have only one member (an isolate) or multiple members (a family). After removing artificial languages, pidgins and creoles and varieties whose ISO-639-3 code cannot be confirmed, the number goes down to 6447 word lists, corresponding to 4298 different lan-

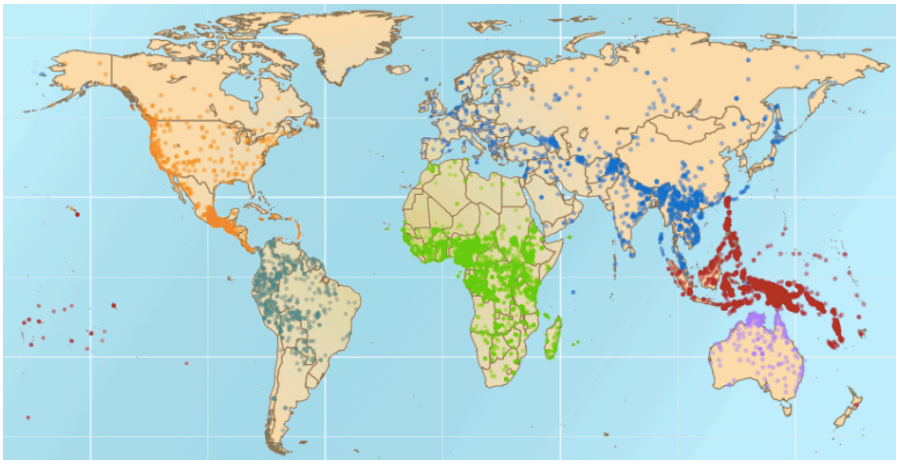

Figure 12: **Geographic coverage of ASJP**

Geographic distribution of the 6452 word lists from the ASJP database [70]. Colors distinguish different linguistic macro-areas, regions with relatively little or no contact between them (but with much internal contact between their populations). These are North America (orange), South America (dark green), Eurasia (blue), Africa (green), Papua New Guinea and the Pacific Islands (red) and Australia (fuchsia).

guages and 359 lineages. In terms of lineages, then, the data covers about 85% of of the totality of them (see Fig. 1).

The database was not constructed for the specific purpose of studying sound-meaning associations, but rather for identifying genealogical relations among languages. For this reason, it generally consists of the 40-item subset of the 100-item so-called Swadesh list [219] that are assumed to remain stable as languages diverge into different lineages over time [112]. Of these word lists, 328 additionally contain the remaining 60 Swadesh lists items.

Words are rendered in a unified transcription system, which facilitates cross-linguistic comparison but also ignores phonetic details such as vowel length, nasalization, tones, and retroflexation. Vowel quality distinctions are merged into seven categories (high front, mid front, low front, high-mid central, low central, high back, mid-low back) (see [34] for a discussion of the system). The ASJP symbol scheme and its phonetic counterparts can be found in Table 4.

Each 40-item word list provides translational equivalents, when available, for the following items: *blood*, *bone*, *breast*, *come*, *die*, *dog*, *drink*, *ear*, *eye*, *fire*, *fish*, *full*, *hand*, *hear*, *horn*, *I*, *knee*, *leaf*, *liver*, *louse*, *mountain*, *name*, *new*, *night*, *nose*, *one*, *path*, *person*, *see*, *skin*, *star*, *stone*, *sun*, *tongue*, *tooth*, *tree*, *two*, *water*, *we*, *you (sg)*. The additional Swadesh list items contained in some of the word lists are: *all*, *ash*, *bark*, *belly*, *big*, *bird*, *bite*, *black*, *burn*, *claw*, *cloud*, *cold*, *dry*, *earth*, *eat*, *egg*, *feather*, *flesh*, *fly*, *foot*, *give*, *good*, *grease*, *green*, *hair*, *head*, *heart*, *hot*, *kill*, *know*, *lie*,

Table 4: ASJP symbols and their description. IPA equivalents of the symbols can be found in Tables 1-2 of [34].

| Symbol | Description |
|---|---|
| p | voiceless bilabial stop and fricative |
| b | voiced labial stop and fricatve |
| m | bilabial nasal |
| f | voiceless labiodental fricative |
| v | voiced labiodental fricative |
| 8 | voiceless and voiced dental fricative |
| 4 | dental nasal |
| t | voiceless alveolar stop |
| d | voiced alveolar stop |
| s | voiceless alveolar fricative |
| z | voiced alveolar fricative |
| c | voiceless and voiced alveolar fricative |
| n | voiceless and voiced alveolar nasal |
| S | voiceless postalveolar fricative |
| Z | voiced postalveolar fricative |
| C | voiceless palato-alveolar affricative |
| j | voiced palato-alveolar affricate |
| T | voiceless and voiced palatal stop |
| 5 | palatal nasal |
| k | voiceless velar stop |
| g | voiced velar stop |
| x | voiceless and voiced velar fricative |
| N | velar nasal |
| q | voiceless and voiced uvular stop |
| X | uvular fricatives and pharyngeal fricatives |
| 7 | voiceless glottal stop |
| h | voiceless and voiced glottal fricative |
| l | voiced alveolar lateral approximate |
| L | all other laterals |
| w | voiced bilabial-velar approximant |
| y | palatal approximant |
| r | all varieties of “r-sounds” |
| i | high front vowel, rounded and unrounded |
| e | mid front vowel, rounded and unrounded |
| E | low front vowel, rounded and unrounded |
| 3 | high and mid central vowel, rounded and unrounded |
| a | low central vowel, unrounded |
| u | high back vowel, rounded and unrounded |
| o | mid and low back vowel, rounded and unrounded |

*long*, *man*, *many*, *moon*, *mouth*, *neck*, *not*, *rain*, *red*, *root*, *round*, *sand*, *say*, *seed*, *sit*, *sleep*, *small*, *smoke*, *stand*, *swim*, *tail*, *that*, *this*, *walk*, *what*, *white*, *who*, *woman*, *yellow*.

Regarding the classification of languages, the Glottolog genealogical classification is preferable over other available alternatives because it is the only one to classify every living or extinct language while providing brief pointers to justifications for all choices taken—however, a less conservative independent classification was used additionally in the main test (see below). We stratify languages geographically by dividing the world's landmass into six largely independent linguistic macro-areas: North America, South America, Eurasia, Africa, Greater New Guinea and Australia—these regions have a history of attested contact within them but little contact between them in prehistorical times [101].

## 3.3 DETECTING SOUND-MEANING ASSOCIATIONS

We aim to capture robust and widespread tendencies in sound-meaning associations, where "tendency" should be understood as a systematic bias in the frequency with which certain words tend to carry specific symbols in contrast to their baseline occurrence in other words. Crucially, a strong tendency does not imply that a signal has an extremely high frequency of occurrence, and conversely a very frequent sound-meaning co-occurrence is not sufficient evidence to discount chance. Importantly, whatever advantage a sound-meaning pairing might confer in terms of learning or processing, it has to be considered in the context of a myriad of competing factors that shape the phonetic and phonological fabric of words, from articulatory production costs [172] to systemic constraints due to the similarity with other lexical elements [236].

Our statistical approach consists in a series of tests where the presence of a symbol in a word is contrasted against a suitable subset of other words, and then the bias is evaluated across lineages. To begin, we calculate, for each concept and symbol, a genealogically balanced average ratio of the times they co-occur in a word of a language for which both symbol and concept are attested, and compare that quantity with a proper random counterpart. The fundamental statistic in our analysis is $p_{ij}$, the maximum likelihood estimator (i.e. the sample frequency) for the probability of finding that concept $i$ has at least one instance of symbol $j$, after randomly choosing a lineage, a language within the lineage and a dialect within the language (if any) in that sequential order. Naturally, this calculation is restricted to the set of dialects of languages for which the concept and the phone are at-

tested (which we will refer as $S_{ij}$); for each of those sets this quantity is formally:

$$p_{ij} = \frac{1}{|L|} \sum_{k=1}^{|L|} \left( \frac{1}{|L_k|} \sum_{l=1}^{|L_k|} \frac{1}{|L_{kl}|} \sum_{d=1}^{|L_{kl}|} \pi_{ij}^{kld} \right)$$

The sets $L$, $L_k$ and $L_{kl}$ are the sets of all lineages, languages within lineage $k$ and dialects of language $l$ within lineage $k$. $\pi_{ij}^{kld}$ is a binary variable that takes value 1 if there is at least one instance of symbol $j$ in the word for concept $i$ for dialect $d$ of language $l$ from lineage $k$ (always within the set $S_{ij}$) and 0 otherwise.

This computation is conservative in that all languages known to belong to the same genealogical group influence the aggregated statistics in the same way regardless of their size, but on the other hand it guarantees the minimum possible bias in the dependence of the languages' words. In order to avoid testing cases whose coverage is insufficiently wide before testing, we evaluated only those associations for which $S_{ij}$ comprises ten lineages in each of three different macro-areas at least.

Conversely, for each dialect of each language we calculated the proportion of words other than that associated with $i$ that have symbol $j$, and we note this as $\pi_{-ij}^{kld}$, and similarly the genealogical balanced average as $p_{-ij}$. These probabilities are used to produce $n_{sim} = 1000$ Monte Carlo simulations of symbol $j$ presence/absence for all the languages in $S_{ij}$ - the set of $p_{-ij}$ values resulting from these simulations will be called $\zeta_{ij}$. The purpose is to compare $\zeta_{ij}$ with $\pi_{ij}$ in order to answer the question: does symbol $j$ appear much more (or much less) often when a subset of words referring to concept $i$ is selected than in a randomly picked set of words from the same languages? The two-tailed P-value for a particular concept $i$ and symbol $j$ is then [185]

$$P = \frac{1}{n_{\text{sim}} + 1} \left( 2 \min\{ |x \in \zeta_{ij} : x \geq p_{ij}|, |x \in \zeta_{ij} : x \leq p_{ij}| \} + 1 \right)$$

where $|\cdot|$ is the cardinality of the set.

There are four potential sources of false positives in this scheme, for which we need to control.

First, the large number of tests performed require a control for type I errors. We perform a False Discovery Rate (FDR) analysis fixing the FDR rejection threshold to .05, which means that we will allow no more than 5% of false positives on average. For this purpose we use the method described in [217]. The basic idea is that the distribution of P-values comes from a mixture of a uniform distribution (that corresponds to the baseline of tests where no associations beyond chance are present) and a distribution concentrated near $P = 0$ of true positives. The method used here learns the mixture proportion of the uniform distribution from values $P$ from 1 down to a threshold that is adjusted in order to reduce the false non-discovery rate (FNDR).

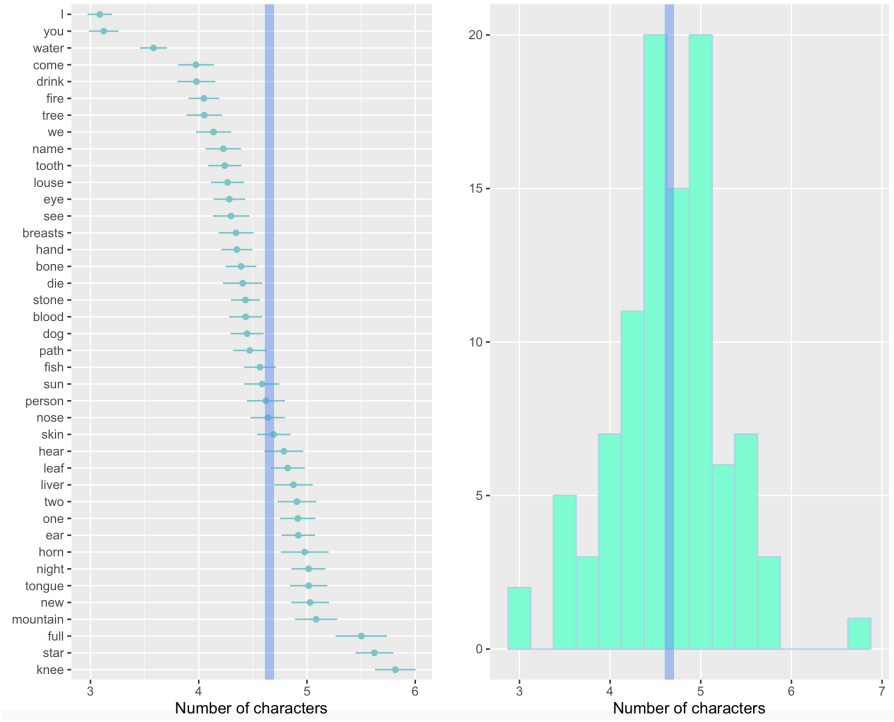


Figure 13: **Word length variation in ASJP**

On the left, genealogically balanced average of the number of characters for each of the 40 concepts with most coverage in ASJP. The horizontal bars represent approximate 95% CI for the average. On the right, distribution of the genealogically balanced average for all of the concepts in ASJP. In both graphs, the vertical blue bar represents the mean value across all concepts in ASJP.

This entire procedure was repeated with a different, less conservative, genealogical classification—the one provided by the World Atlas of Language Structures (WALS) [103]. For our analysis, we only considered associations that were below the defined FDR level according to both classifications. The fraction of the component of true negatives learned from both classifications was around 0.65.

Second, word length is trivially correlated with the chance of finding any particular symbol. There is considerable variance in the (genealogically balanced) length of the words in our dataset, with some pronouns, negation and basic verbs (like *say* and *give*) consisting only of about three symbols on average, whereas the length of some color words and body part terms contain is over five (see Fig. 13).

To control for this confound, for each language (and dialect) in $S_{ij}$, $n = 1000$ of independent simulations we sampled without replacement as many random symbols from words other than $i$ up to the length of word $i$. This effectively produces, for each word $i$, a random counterpart equivalent to shuffling all the symbols corresponding to all the the words of a language while keeping word lengths

constant. Over each of those sets, the same association test based on the Glottolog classification was performed.

Third, besides the mere number of symbols, word length might be a confound due to the fact that different phonotactic restrictions might apply accordingly. For instance, in a language that only allows CV structures and also prohibits the presence of word-initial liquids, no monosyllabic words will carry liquids. To remedy this, we repeated the same global test using the Glottolog classification this time comparing $p_{ij}$ with simulations obtained from words of exactly the same number of symbols in each language (and dialect).

In both of these two last procedures, we imposed a stricter cutoff: if any of the simulations yield a value of $p_{ij}$ equally or more extreme, we would reject the association as of potential interest.

Fourth, an important indicator for a consistent bias of a sound-meaning association is its ubiquitous nature. Finding that a sound-meaning association arises independently in areas with not strong contact in historical times is a strict yet important litmus test. Besides, some associations might result due to a large-scale areal contact or unresolved genealogy. With this idea in mind, for each macro-area with at least 10 independent lineages in $S_{ij}$, we analyzed the presence of a significant direction of association as in the main associations test—computing both empirical and random probabilities using only the languages of that area—with the difference that we flagged each macro-area specific association with $P \leq .1$. It should be noticed that this does not imply a softer rejection threshold than in the worldwide case: we only keep associations that display a bias consistent with the world-wide trend in at least half of the macro-areas, with the extra condition that no macro-area should exhibit a bias in the opposite direction.

To summarize: only associations that successfully satisfied all the requirements of the overall association test (with Glottolog and WALS classifications independently), the word length and the matched-length tests, and for which a consistent preference in at least half of the macro-areas could be found were considered "signals".

It should be noted that the overall testing scheme is conservative and that it is likely to have a large false negative rate. Also working against our analyses is the fact that the core set of concepts we use was originally gathered due to their exceptional phylogenetic persistence and resistance to borrowing, thus rendering them less likely to be adapted to potential functional biases that might underlie specific sound-meaning associations. Moreover, it is not clear a priori whether the granularity of our phonetic descriptions is sufficiently fine to capture widespread sound-meaning relations—for instance, the opposition between voiced and unvoiced consonants and between rounded and unrounded in vowels have been suggested to bear importance for sound-symbolism [124, 142], but each feature pair are usually con-

flated under a single symbol in the database. For these reasons, the associations found in our analyses should be regarded as providing a lower-bound estimate of the presence of non-arbitrariness in sound-meaning pairings.

## 3.4 STRONG WORLDWIDE ASSOCIATIONS

Our analysis detected 74 (positive and negative) signals, involving 30 concepts and 23 symbols. Signals will be described in terms of the most relevant information about them: the frequency of the symbol in the words corresponding to the concept ($p$), the ratio between that frequency and the frequency in other words ($RR$), the number of lineages that were analyzed for the global association ($n_l$) and the ratio between the number of areas where the association was independently found and the total number of tested areas ($a_s/a_t$). Table 5 and6 display the positive and negative signals, respectively.

Some concepts are associated with more than one signal. These are expected to be correlated; across languages it is often observed that there are preferences or restrictions with regard to the co-occurrence of symbols within one and the same word for either diachronic or synchronic phonotactic reasons. As an example, it is known that high front vowels trigger palatalization [15], so it is therefore not surprising that the voiceless palato-alveolar affricate C appears with i in the signals of *small*.

We analyze this statistically by taking sets of languages for which both the concept and the symbol associated with a pair of signals was present in at least ten lineages in each of (at least) three macro-areas. The association between signals—which we will refer to $A$ and $B$ here—was tested by means of a simple mixed effects logistic model,

$$\text{logit}(\text{signal A presence}) = \alpha_{\text{signal B presence}} + \alpha^{\text{lineage}}$$

where $\alpha_{\text{signal A presence}}$ is the coefficient related to the presence of signal A, and $\alpha^{\text{lineage}}$ is a random coefficient structured according to lineage. To the results obtained by comparing all the pairwise associations between signals belonging to the core 40 words, we applied a threshold on the FDR of 5%. About 12% of the 2062 cases satisfied this condition. Signals sharing a concept tend to be significantly associated in about 41% of the time, against only 8% of signals involving different concepts. The results of associations regarding same-concept signals and the genealogically balanced average effect on the presence of signal B on A can be found in Table 7.

The signals found in our analysis show a mixture of well-known and new associations. In line with the considerable literature on magnitude sound symbolism, the concept *small* was found to be associated with the high front vowel i ($RR$=1.58, $p$=.61, $n_l$=78, $a_s/a_t$=3/5), consistent with findings linking vowel height quality and size [109,

Table 5: Complete list of positive signals found in the ASJP database. The column 'Areal ratio' indicates the ratio between the number of areas where the signals are independently found with respect the total number of areas with minimum coverage. RR stands for "risk ratio". Family counts come from Glottolog [184].

| Concept | Symb. | $p_{ij}$ | $p_{-ij}$ | $\sigma(p_{-ij})$ | $\Delta$ | RR | Lineages | Areal ratio |
|---|---|---|---|---|---|---|---|---|
| ash | u | 0.516 | 0.270 | 0.043 | 0.25 | 1.91 | 68 | 3/5 |
| bite | k | 0.438 | 0.259 | 0.042 | 0.18 | 1.69 | 73 | 3/5 |
| bone | k | 0.311 | 0.223 | 0.016 | 0.09 | 1.39 | 333 | 3/6 |
| breasts | u | 0.376 | 0.257 | 0.018 | 0.12 | 1.46 | 317 | 4/6 |
| breasts | m | 0.326 | 0.200 | 0.016 | 0.13 | 1.63 | 320 | 4/6 |
| dog | s | 0.225 | 0.128 | 0.015 | 0.10 | 1.76 | 285 | 3/5 |
| ear | k | 0.319 | 0.224 | 0.017 | 0.09 | 1.42 | 338 | 4/6 |
| fish | a | 0.613 | 0.524 | 0.019 | 0.09 | 1.17 | 327 | 3/6 |
| full | p | 0.255 | 0.121 | 0.016 | 0.13 | 2.11 | 231 | 5/6 |
| full | b | 0.229 | 0.120 | 0.016 | 0.11 | 1.91 | 213 | 4/6 |
| hear | N | 0.199 | 0.127 | 0.018 | 0.07 | 1.57 | 182 | 3/6 |
| horn | k | 0.339 | 0.222 | 0.019 | 0.12 | 1.53 | 221 | 4/6 |
| horn | r | 0.271 | 0.155 | 0.019 | 0.12 | 1.75 | 191 | 3/6 |
| I | 5 | 0.129 | 0.063 | 0.015 | 0.07 | 2.06 | 136 | 4/6 |
| knee | u | 0.472 | 0.256 | 0.018 | 0.22 | 1.84 | 303 | 4/6 |
| knee | o | 0.406 | 0.239 | 0.017 | 0.17 | 1.70 | 291 | 4/6 |
| knee | p | 0.218 | 0.121 | 0.014 | 0.10 | 1.81 | 278 | 5/6 |
| knee | k | 0.374 | 0.226 | 0.018 | 0.15 | 1.66 | 305 | 5/6 |
| knee | q | 0.313 | 0.136 | 0.027 | 0.18 | 2.30 | 73 | 3/5 |
| leaf | p | 0.232 | 0.119 | 0.014 | 0.11 | 1.94 | 290 | 3/6 |
| leaf | b | 0.185 | 0.124 | 0.014 | 0.06 | 1.48 | 274 | 3/6 |
| leaf | l | 0.268 | 0.154 | 0.016 | 0.11 | 1.75 | 270 | 4/6 |
| name | i | 0.474 | 0.378 | 0.020 | 0.10 | 1.25 | 320 | 3/6 |
| nose | u | 0.351 | 0.255 | 0.018 | 0.10 | 1.38 | 325 | 4/6 |
| nose | n | 0.356 | 0.242 | 0.016 | 0.11 | 1.47 | 334 | 4/6 |
| one | t | 0.266 | 0.178 | 0.015 | 0.09 | 1.49 | 343 | 3/6 |
| one | n | 0.320 | 0.248 | 0.017 | 0.07 | 1.29 | 348 | 3/6 |
| red | r | 0.350 | 0.156 | 0.037 | 0.19 | 2.24 | 61 | 3/5 |
| round | r | 0.371 | 0.149 | 0.038 | 0.22 | 2.48 | 56 | 4/5 |
| sand | s | 0.325 | 0.126 | 0.034 | 0.20 | 2.58 | 65 | 3/5 |
| small | i | 0.613 | 0.389 | 0.043 | 0.22 | 1.58 | 78 | 3/5 |
| small | C | 0.416 | 0.081 | 0.029 | 0.33 | 5.12 | 61 | 3/4 |
| star | z | 0.158 | 0.063 | 0.018 | 0.10 | 2.52 | 96 | 3/5 |
| stone | t | 0.239 | 0.181 | 0.015 | 0.06 | 1.32 | 340 | 3/6 |
| tongue | e | 0.339 | 0.220 | 0.017 | 0.12 | 1.54 | 322 | 5/6 |
| tongue | E | 0.278 | 0.161 | 0.020 | 0.12 | 1.73 | 164 | 4/6 |
| tongue | l | 0.419 | 0.151 | 0.017 | 0.27 | 2.77 | 280 | 6/6 |
| we | n | 0.380 | 0.246 | 0.017 | 0.13 | 1.54 | 325 | 3/6 |

Table 6: Complete list of negative signals found in the ASJP database. The column 'Areal ratio' indicates the ratio between the number of areas where the signals are independently found with respect the total number of areas with minimum coverage. RR stands for "risk ratio". Family counts come from Glottolog [184]

| Concept | Symb. | $p_{ij}$ | $p_{-ij}$ | $\sigma(p_{-ij})$ | $\Delta$ | RR | Lineages | Areal ratio |
|---|---|---|---|---|---|---|---|---|
| bone | y | 0.065 | 0.122 | 0.013 | -0.06 | 0.54 | 312 | 3/6 |
| breasts | a | 0.422 | 0.524 | 0.020 | -0.10 | 0.81 | 329 | 3/6 |
| breasts | h | 0.093 | 0.149 | 0.016 | -0.06 | 0.62 | 254 | 3/6 |
| breasts | r | 0.083 | 0.175 | 0.015 | -0.09 | 0.47 | 290 | 3/6 |
| dog | t | 0.106 | 0.182 | 0.015 | -0.08 | 0.58 | 337 | 4/6 |
| drink | a | 0.421 | 0.533 | 0.020 | -0.11 | 0.79 | 310 | 4/6 |
| eye | a | 0.423 | 0.527 | 0.018 | -0.10 | 0.80 | 357 | 4/6 |
| I | u | 0.116 | 0.262 | 0.018 | -0.15 | 0.44 | 328 | 5/6 |
| I | p | 0.021 | 0.122 | 0.014 | -0.10 | 0.18 | 297 | 5/6 |
| I | b | 0.030 | 0.124 | 0.014 | -0.09 | 0.24 | 276 | 4/6 |
| I | t | 0.079 | 0.181 | 0.016 | -0.10 | 0.44 | 332 | 4/6 |
| I | s | 0.036 | 0.131 | 0.015 | -0.10 | 0.27 | 279 | 4/5 |
| I | l | 0.030 | 0.161 | 0.016 | -0.13 | 0.19 | 277 | 6/6 |
| I | r | 0.061 | 0.177 | 0.015 | -0.12 | 0.35 | 294 | 6/6 |
| name | o | 0.169 | 0.254 | 0.018 | -0.09 | 0.67 | 297 | 4/6 |
| name | p | 0.049 | 0.122 | 0.015 | -0.07 | 0.40 | 283 | 3/6 |
| nose | a | 0.391 | 0.524 | 0.019 | -0.13 | 0.75 | 339 | 4/6 |
| skin | m | 0.109 | 0.207 | 0.016 | -0.10 | 0.53 | 323 | 4/6 |
| skin | n | 0.170 | 0.256 | 0.016 | -0.09 | 0.66 | 329 | 4/6 |
| tongue | u | 0.164 | 0.264 | 0.017 | -0.10 | 0.62 | 327 | 3/6 |
| tongue | k | 0.167 | 0.232 | 0.017 | -0.07 | 0.72 | 334 | 4/6 |
| tooth | b | 0.054 | 0.126 | 0.014 | -0.07 | 0.43 | 282 | 4/6 |
| tooth | m | 0.130 | 0.205 | 0.016 | -0.08 | 0.63 | 335 | 4/6 |
| water | t | 0.066 | 0.184 | 0.015 | -0.12 | 0.36 | 345 | 6/6 |
| we | p | 0.052 | 0.121 | 0.015 | -0.07 | 0.43 | 288 | 5/6 |
| we | l | 0.064 | 0.160 | 0.016 | -0.10 | 0.40 | 268 | 5/6 |
| we | s | 0.077 | 0.129 | 0.015 | -0.05 | 0.60 | 273 | 3/5 |
| you | u | 0.149 | 0.259 | 0.017 | -0.11 | 0.58 | 316 | 3/6 |
| you | o | 0.165 | 0.246 | 0.017 | -0.08 | 0.67 | 306 | 3/6 |
| you | p | 0.046 | 0.124 | 0.014 | -0.08 | 0.37 | 289 | 3/6 |
| you | t | 0.072 | 0.182 | 0.015 | -0.11 | 0.40 | 322 | 5/6 |
| you | d | 0.045 | 0.129 | 0.015 | -0.08 | 0.35 | 264 | 4/6 |
| you | q | 0.043 | 0.146 | 0.029 | -0.10 | 0.29 | 75 | 3/5 |
| you | s | 0.049 | 0.131 | 0.015 | -0.08 | 0.37 | 271 | 4/5 |
| you | r | 0.053 | 0.180 | 0.016 | -0.13 | 0.29 | 284 | 6/6 |
| you | l | 0.030 | 0.159 | 0.016 | -0.13 | 0.19 | 266 | 6/6 |

Table 7: Dependencies between signals involving the same concept. The effect is the genealogically balanced mean change in probability of finding the first symbol given that the second is present (as estimated by the mixed model). Only entries with q-values smaller than 0.05 shown. See Materials & Methods for further details.

| Concept | Symb.1 | Symb.2 | Effect | Fam. tested |
|---|---|---|---|---|
| bone | y | k | -0.04 | 298 |
| bone | k | y | -0.14 | 298 |
| breasts | h | m | -0.04 | 237 |
| breasts | u | a | -0.16 | 314 |
| breasts | u | m | -0.10 | 309 |
| breasts | a | u | -0.16 | 314 |
| breasts | a | m | 0.18 | 317 |
| breasts | a | r | 0.11 | 285 |
| breasts | m | h | -0.10 | 237 |
| breasts | m | u | -0.08 | 309 |
| breasts | m | a | 0.12 | 317 |
| breasts | r | a | 0.03 | 285 |
| dog | s | t | -0.08 | 281 |
| dog | t | s | -0.04 | 281 |
| full | b | p | -0.17 | 175 |
| full | p | b | -0.21 | 175 |
| I | b | t | 0.02 | 264 |
| I | s | u | -0.02 | 265 |
| I | t | b | 0.04 | 264 |
| I | u | s | -0.07 | 265 |
| knee | k | q | -0.19 | 71 |
| knee | o | u | -0.28 | 273 |
| knee | q | k | -0.22 | 71 |
| knee | u | o | -0.29 | 273 |
| leaf | l | b | 0.10 | 217 |
| leaf | b | l | 0.09 | 217 |
| leaf | b | p | -0.18 | 226 |
| leaf | p | b | -0.21 | 226 |
| name | o | i | -0.06 | 290 |
| name | i | o | -0.12 | 290 |
| nose | a | n | 0.05 | 329 |
| nose | a | u | -0.09 | 321 |
| nose | n | a | 0.05 | 329 |
| nose | n | u | -0.05 | 319 |
| nose | u | a | -0.09 | 321 |
| nose | u | n | -0.06 | 319 |
| one | n | t | -0.07 | 338 |
| one | t | n | -0.06 | 338 |
| tongue | E | e | -0.17 | 142 |
| tongue | e | E | -0.16 | 142 |
| tooth | b | m | 0.03 | 272 |
| tooth | m | b | 0.02 | 272 |
| we | l | n | -0.04 | 257 |
| we | n | l | -0.19 | 257 |
| we | n | p | -0.06 | 279 |
| you | d | t | -0.04 | 253 |
| you | r | o | 0.02 | 254 |
| you | u | o | -0.10 | 285 |
| you | o | r | 0.15 | 254 |
| you | o | s | -0.04 | 252 |
| you | o | u | -0.11 | 285 |
| you | t | d | -0.04 | 253 |

186], and with the palatal consonant C ($RR$=5.12, p=.41, $n_l$=61, $a_s/a_t$=3/4), also in agreement with previous work [108, 109].

We also observed a strong association between *round* and r-sounds ($RR$=2.48, p=.37, $n_l$=56, $a_s/a_t$=4/5). While most recent research has emphasized the role of consonants in shape-sound meaning associations like this [78, 182], the usual hypothesis in this direction concerned the correlation between vowel roundedness and round objects [154] – association that appears as a tendency in our analyses without reaching the minimum statistical threshold established before. Both *small* and *round* have been linked to the phenomenon of cross-modal mapping [14, 47, 197]. Another property word, *full*, is endowed with a pair of signals involving voiced ($RR$=1.91, p=.22, $n_l$=213, $a_s/a_t$=4/6) and unvoiced bilabial stops ($RR$=2.11, p=.13, $n_l$=231, $a_s/a_t$=5/6).

Some of the strongest signals found correspond to body parts. *Tongue* was very strongly associated with the lateral 'l' ($RR$=2.77, p=.41, $n_l$=280, $a_s/a_t$=6/6) and the mid and low front vowels e ($RR$=1.54, p=.11, $n_l$=322, $a_s/a_t$=5/6) and E ($RR$ =1.73,p=.11 ,$n_l$=164, $a_s/a_t$=4/6). *Nose* was found to be associated most strongly with the alveolar nasal n ($RR$=1.47, p=.35, $n_l$=334, $a_s/a_t$=4/6), the high back vowel u ($RR$=1.38, p=.35, $n_l$=325, $a_s/a_t$=4/6). The link between *nose* and nasality has been noted previously [95], in particular in reference to the conjecture that body part terms used in phonation makes use of the distinctive qualities provided by the relevant organ [232].

*Breasts* was associated with the bilabial nasal consonant m ($RR$=1.63, p=.32, $n_l$=320, $a_s/a_t$=4/6) and the high back vowel u ($RR$=1.46, p=.37, $n_l$=317, $a_s/a_t$=4/6). Similar associations were found in the nursery terms for *mother*, a concept with which it often colexifies. It has been suggested that this might be due to the mouth configuration of suckling babies or to the sounds feeding babies produce [120, 231].

While this study lends support to a number of associations that were either elicited in experiments or conjectured based on a much smaller number of languages, it also provides telling negative evidence on others. Together with the association between high front vowels and the concept of small, there has been reports on a connection between back low vowels and the notion of big [124]. However, *big* ($n_l$=73) and *large* ($n_l$=74) and o did not show any relevant signature of association in our sample at the global level. Similarly, an analogous front/back vowel opposition has been proposed to hold between proximal and distal pronouns—the purported explanation being that proximal referents tend to be small whereas distal referents are usually large [124]. The concepts *this* ($n_l$=71) and *that* ($n_l$=74), however, do not show any associations with i and o (respectively).

## 3.5 ORIGINS AND NATURE OF THE ASSOCIATIONS

As discussed in the previous sections, there are multiple theories which attempt to elucidate why humans find that some sounds are more convenient or salient in association with certain meanings. How these hypothesized mechanisms lead to the widespread biases in vocabularies we find here is a complex question that is unlikely to be fully answered by the inspection of wordlists. Nonetheless, we can attempt to evaluate some of the potential consequences of those theories given the coarse level of detail of our data.

Functional advantages might increase the likelihood of signals being borrowed across languages in contact with one another, thus producing spatial diffusion patterns [231] (see Figure 2). The existence of opposing factors obscure definitive inferences in this direction, though: basic vocabulary items are particularly resistant to borrowing but unresolved genealogy involving nearby languages would be confounded with borrowing. In the same direction, large populations have been claimed to be more efficient at gaining and retaining non-arbitrary sound-meaning associations given a potential functional value [231], which is coherent with recent evidence from some Austronesian languages showing that larger populations gain new words at a faster rate [33].

We determined whether present-day log population size and log distance to the nearest genealogically unrelated language bearing the (positive) signal are effective predictors for signal presence. For each positive signal we calculated the great circle distances—i.e., the distance in kilometers of the shortest geodesic connecting two points in the surface of the Earth—involving all languages having both the relevant symbol and concept (but not necessarily the signal) and their nearest language from a different lineage that has the (positive) signal (dnn). More precisely, the hypothesis is that small distance from a language that has a signal will influence the likelihood of signal presence in a given language. Only signals belonging to the group of 28-40 better attested concepts were used for the analysis, and only one dialect per language was chosen. Extinct languages were excluded from the analyses.

For the testing we used a generalized logistic model with random effects:

$$\begin{aligned}\operatorname{logit}(\mathbb{E}[\text{signal presence}]) = \alpha + (\beta_{\text{dnn}} + \beta_{\text{dnn}}^{\text{lineage}}) \log(1 + \text{dnn}) \\ + \beta_{\text{pop}} \log(\text{population}) + \alpha^{\text{lineage}}\end{aligned}$$

where the superscripted coefficients ($\beta_{\text{dnn}}^{\text{lineage}}$ and $\alpha^{\text{lineage}}$) are random effects structured according to the lineage. Lineage as a random intercept is introduced as a means of accounting for the varying baseline presence of the signals within lineages, and their presence as random slopes aims to capture the fact that lineages have spread with

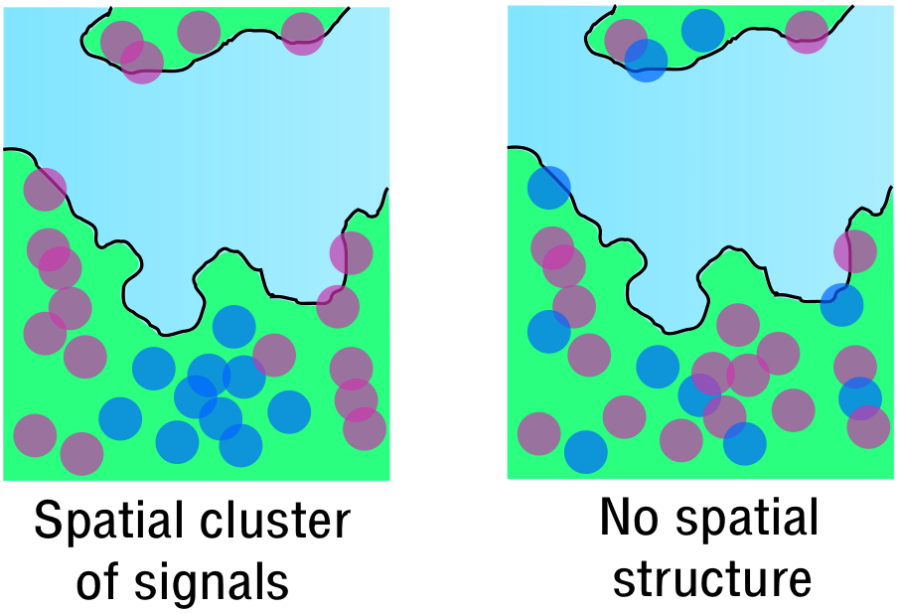


Figure 14: **Spatial scenarios for signal distributions**

Competing configurations of the spatial distribution of the tested languages. Blue and fuchsia dots represent languages with and without a specific signal, respectively. In the panel to the left, the likelihood of a language having the signal is correlated with its geographical distance to its nearest neighbor, and on the right there is no spatial structure.

Table 8: Estimated parameters ($\beta$), genealogical balanced mean probability difference (difference in 1-10000 km. reference) and P-values for the distance to nearest neighbor (dnn) and population model, displayed only for the signals and variables that reached significance at $\alpha = 0.05$. See main text for details.

| | | Dist. to nearest neighbor | | | Population | | |
|---|---|---|---|---|---|---|---|
| Concept | Symb. | $\beta$ | diff. | P-value | $\beta$ | diff. | P-value |
| stone | t | -0.59 | -0.29 | 0.01 | - | - | - |
| full | p | -0.54 | -0.44 | 0.01 | - | - | - |
| dog | s | -0.44 | -0.18 | 0.05 | 0.79 | 0.06 | $< 10^{-3}$ |
| tongue | E | -0.36 | -0.34 | 0.03 | - | - | - |
| knee | o | -0.26 | -0.27 | 0.03 | - | - | - |
| knee | u | -0.26 | -0.25 | 0.02 | - | - | - |
| nose | n | -0.24 | -0.20 | 0.04 | - | - | - |
| fish | a | - | - | - | 1.01 | 0.18 | $< 10^{-3}$ |
| knee | p | - | - | - | -1.09 | -0.12 | $< 10^{-3}$ |
| leaf | b | - | - | - | 0.57 | 0.06 | 0.01 |
| leaf | p | - | - | - | -0.51 | -0.05 | 0.04 |
| name | i | - | - | - | -0.42 | -0.08 | 0.01 |
| one | t | - | - | - | -0.58 | -0.06 | 0.002 |
| star | z | - | - | - | 0.86 | 0.05 | 0.05 |
| tongue | e | - | - | - | -0.36 | -0.06 | 0.03 |
| tongue | l | - | - | - | 1.17 | 0.18 | $< 10^{-3}$ |

different rates across the globe. The logarithmic transforms aims to reduce the effect of population and distance outliers. P-values were estimated through an asymptotic likelihood ratio test. Apart from the estimated coefficients, we calculated the genealogical balanced mean difference in probability of having a signal for two reference points, one variable at a time. For population, the difference was calculated between fixing all languages' populations to 10000 individuals and a single individual, and for dnn between 1000 km—which is roughly the maximum radius of linguistic areas as defined in AUTOTYP—and 0 km (which correspond to the situation where both languages as spoken at the same place). The results can be observed in Table reftab:spatial.

At $\alpha = 0.05$, log population turned out to be significant in about one third of the cases, but the effect was small and as many times positive as it was negative, which rules out a consistent role for population. Only one fifth of the signals showed sensitivity to the distance of nearest neighbors with signal, with all of the cases having an effect in the predicted direction by our model. On average, and in contrast to the case in which a language and its signal-bearing nearest ge-

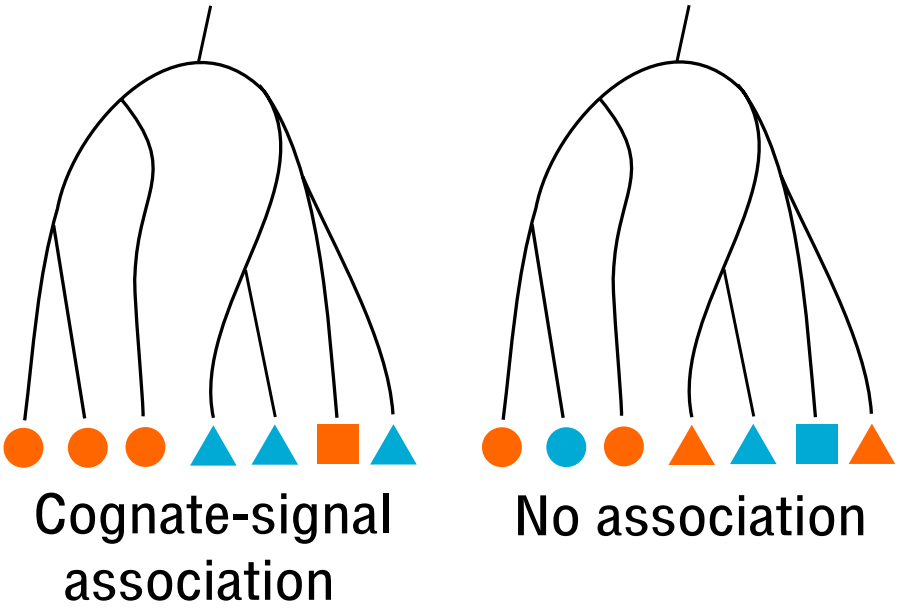


Figure 15: **Genealogical scenarios for signal distribution** Genealogical trees of languages where leaves are words for specific referents. In the figure to the left, cognate classes (depicted as different shapes) are associated with signal presence (blue shapes), whereas to the right there is no such correspondence.

nealogically unrelated neighbor are spoken in exactly the same place, the probability of finding the signal also in the language drops by 28%.

From a historical perspective, it has been suggested that sound-meaning associations might be evolutionarily preserved features of spoken language [187], potentially hindering regular sound change [186]. Furthermore, it has been claimed that widespread sound-meaning associations might be vestiges of one or more large-scale prehistoric proto-languages [114]. Tellingly, some of the signals found here feature prominently in reconstructed "global etymologies" [202, 212] that have been used for deep phylogeny inference [188]. If signals are inherited from an ancestral language spoken in remote prehistory, we might expect them to be distributed similarly to inherited, cognate words; that is, their distribution should to a large extent be congruent with the nodes defining their linguistic phylogeny (see Figure 3 for illustration).

A direct evaluation of this hypothesis is infeasible due to the absence of etymological dictionaries for all but a few families. More precisely, a proper phylogenetic test in the context of language history would comprise some kind of data carrying a phylogenetic signal (like cognate sets or collections of regular sound changes) and a sound evolutionary model that would lead to a tree or a distribution

of trees. Unfortunately, such trees exist for only a handful of language families.

However, it can be tested indirectly given that cognate words are expected to be more similar to one another than non-cognates [213]. If it is a correct hypothesis that signals render words less prone to change and that they are prehistoric vestiges, then, after controlling for concept, symbol, and lineage, we would expect to find that the similarity among words is predicted by signals.

The distance between words used here is the Levenshtein distance, which has found several uses in linguistics and often correlates with perceptual, processing and other meaningful lexical distances differences [4,5]. The Levenshtein distance between strings x and y LD(x,y) is defined as the minimum number of edits, additions or deletions of characters necessary to make two strings identical. For instance, 'Zultus' and 'sulus'—*star* in Uyghur and Sakha (two Turkic languages) respectively, have a Levenshtein distance of 2: a change of 'Z' to 's' and the deletion of 't' in the Sakha word. The normalized Levenshtein distance is simply $l = LD(x,y)/\max(|x|,|y|)$

For every family with at least six languages and every combination of concept and symbol, we calculated the Levenshtein distance between all members of two groups: word pairs for a concept belonging to a combination, and word pairs for a concept sharing at least one symbol but not the symbol relevant for the combination. For instance, given a family with three languages having the forms *ana*,*ena* and *ete* for the concept "rock", and considering the combination rock-n, we will have the two following groups: (a**n**a,e**n**a) and (**e**na,**et**e). Families with less than three distances in any of the groups were excluded from the analysis.

In order to summarize the previous information, we calculated, for each family, the probability of choosing a distance in the signal-sharing group and another in the non-signal-sharing group and finding that the first is smaller than the second ($\Pr(l_s < l_{-s})$). The larger this quantity, the more reliable an estimator of wordform similarity the association is.

Then we implemented the following beta regression mixed model with logistic link function and constant precision parameter:

$$\begin{aligned}\operatorname{logit}(\mathbb{E}[\Pr(l_s < l_{-s})]) = &\sum_{\text{concepts}} \beta_i I_i + \sum_{\text{symbols}} \beta_j I_j \\ &+ \alpha_{\text{signalhood}} + \alpha^{\text{lineage}}\end{aligned}$$

where the $i$ and $j$ indexes run over the set of concepts and symbols, respectively, the coefficient "signalhood" indicates whether the combination of concept and symbol is to be found in Table S2. 'signalhood' was coded as a single level common to all individual positive signals. $\alpha^{\text{lineage}}$ stands for a random intercept according to lineage. In order to cope with a few values of $\Pr(l_s < l_{-s})$ identical to 1 (that

account for less than 0.5% of the data) we applied the transformation $t(x) = (x(N-1)+0.5)/N$ to the values. As a way of accounting for the more robust evidence provided by lineages with a large number of distance pairs to be compared, we included a weight for each observation equal to the logarithm of the number of such pairs involved—however, the results did not differ considerably from the unweighted case.

Overall, the model quality is heavily dominated by lineage: 86% vs. 3% of explained deviance with and without the lineage random effect, respectively. Signal presence (while significant) has a negligible effect in the opposite direction than predicted: the genealogically balanced average effect is less than a 0.5% decrease in similarity for those words sharing a signal-related symbol compared to those sharing some other symbol.

Consistency in word position is important for establishing cognacy [118, 213]. Further support for the idea that signals are not residuals of deep history comes from the analysis of the position within the word in which they occur, in particular whether they have a clear word-initial bias.

We simulate, for each language and signal, random positions of the relevant signal-associated symbol based on all the available positions in the word according to the consonant/vowel distinction. Concretely, we calculate the number of times the phone is initial when its simulated counterpart is not, averaging genealogically and respecting the vowel and consonant template of each word. Then we compare this quantity in the original word list against $n = 1000$ simulations and consider those cases in which the original bias is larger than 95% of the simulated cases. These results can be observed in Table 9.

All in all, we find that signals do not have a consistent cross-linguistic preference or dispreference in this respect beyond well-established cross-linguistic phonotactic patterns, such as the avoidance of liquids or the prevalence of dorsal and labial stops in word-initial position [145, 196] (see Supplemental Methods and Table S5).

In perspective, these results suggest that although it is possible that the presence of signals in some families are symptomatic of a particularly pervasive cognate set, this is not the usual case. Hence, the explanation for the observed prevalence of sound-meaning associations across the world has to be found elsewhere [36].

## 3.6 CONCLUSION

We have demonstrated that a substantial proportion of words in the basic vocabulary are biased to carry or to avoid specific sound segments, both across continents and linguistic lineages. Given that our analyses suggest that phylogenetic persistence or areal dispersal are unlikely to explain the widespread presence of these signals, we are

Table 9: Analysis of word-initial position bias. Bias measure how more or less frequently the symbol appears in word initial position for that concept. Lineages counts how many lineages had at least one language for which the analysis could be performed. See Materials & Methods for more details.

| Concept | Symb. | Bias | Lineages |
|---|---|---|---|
| bite | k | 0.20 | 42 |
| bone | k | 0.09 | 162 |
| breasts | u | -0.06 | 185 |
| breasts | m | 0.05 | 152 |
| ear | k | 0.07 | 159 |
| fish | a | 0.05 | 249 |
| full | b | 0.11 | 81 |
| full | p | 0.12 | 100 |
| horn | r | -0.23 | 82 |
| horn | k | 0.15 | 115 |
| knee | o | 0.10 | 177 |
| knee | p | 0.09 | 104 |
| knee | k | 0.07 | 177 |
| knee | q | 0.19 | 35 |
| leaf | l | -0.14 | 120 |
| one | n | -0.07 | 175 |
| red | r | -0.24 | 28 |
| tongue | l | -0.09 | 160 |

left with the alternative that the signals are due to factors common to our species, such as sound symbolism, iconicity, communicative pressures or synaesthesia. We expect future research to further elucidate the role and interaction of these factors in driving the observed sound-meaning association biases, and to extend the scope of our findings to a broader portion of the vocabulary.

The outcome of our analyses have consequences for historical and comparative linguistics, where it has been suggested that there is a small set of ultra-conserved words that are particularly useful for establishing ancient genealogical relations beyond the limits of the comparative method [188]. However, some of these words are involved in the signals discovered here: *we* is associated with the alveolar nasal, *hear* with the velar nasal, and *ash* with the vowel u. Thus, proposals of far-reaching etymologies based on words of similar form and meaning should be accompanied by an evaluation of whether the observed lexical similarities might have resulted from the kinds of signal discussed in this paper rather than common inheritance. More generally, even though it is unclear whether the locus of the emergence of signals is in the invention or historical development of lexical roots, our findings have implications for the study of the dynamics of lexical phonology.

In summary, our results provide new insights into the constraints that affect how we communicate, suggesting that despite the immense flexibility of the worlds languages, some sound-meaning associations are preferred by culturally, historically and geographically diverse human groups.

# Part III

# EXPLAINING DIVERSITY IN THE WORLD'S LANGUAGES

# 4

# CREOLES AS A TYPOLOGICAL GROUP

## 4.1 EXTREME CONTACT LANGUAGES

The label "creole" is given to languages that are considered to be the result of extreme contact situations, prototypically (but not exclusively) as the result of the exploitation of slave plantations following the European colonial expansion since the end of the Middle Ages. As with all other languages creoles have a considerable share of words that can be traced back to another language or languages[1]. In the case of creoles, those languages -referred as *lexifiers*- were most of the time dialects of European languages. English speakers will be able to get a sense of the following extract (meant to be read following English writing conventions) from the New Testament:

> So King Herod sen fa de man dem dat done come fom de east fa meet wid um, bot e ain tell nobody bout de meetin. Den Herod aks dem man fa tell um de zact time wen dey fus see dat staa. E tell um say, "Oona mus go ta Betlem an look roun good fa de chile. Wen oona find um, mus come back an leh me know, so dat A kin go mesef fa woshup um op too."

The piece is in Gullah, a creole that is still spoken in the states of Georgia and South Carolina in the United States. Gullah, a prototypical creole, originated from the interaction between English — its lexifier — and a number of African languages from the Guinea coast, which are collectively labelled as *substrate* languages. Beyond proper nouns, many words of English origin can be easily spotted, including verbs (*tell*, *go*, *say*), nouns (*kin* "king", *man*, *chile* "children"), adjectives and adverbs (*zact* "exact", *good*). For most languages, the lexical material is transmitted along with the grammar. This enables the possibility of historical linguistics, in fact: by detecting vocabulary of common origin (and regular sound changes) in a set of languages, linguists are able to make predictions about the common ancestry of the underlying grammars as well. However, for creoles, there is no

1 As an exception one could refer here to some *secret languages*, which explicitly alter the words of a target language (or directly create new artificial words) in order to make it unintelligible to outsiders.

consensus about whether this is the also the case. The theories that have been put forward in the last half-century about the origin of creole grammars range from the spectacular to the dull. While some regard creoles as average languages (putting aside their particular social underpinnings), others think they might be a linguistic Holy Grail that could ultimately shed light on fundamental issues such as the evolution of human language, the importance of communication for shaping language structures and the strength of innate linguistic biases.

## 4.2 ORIGINS OF CREOLES

Recurrently, researchers have pointed out that there seems to be a number of features shared by all or most of creole languages [208, 234]. Considering some of the recurrent (but not omnipresent) features there is the SVO word order, the realization of tense, aspect and mood markers as free morphology, and the lack of noun classes, in particular gender — for an excellent review see [51]. Perhaps following the intuition from classic historical linguistics, the first ideas about these commonalities attempted to find a single root to all creoles. The candidate for siring a large number of creole languages was thought to be a *pidgin* — which is, as a first approximation, a restricted communication code with an extremely limited lexicon that does not have any native speakers — that served for the purpose of commerce and exchange in the Atlantic between the XV and XVI centuries and that was based on Portuguese [223]. In many cases the slave trade that fed the creole-speaking populations can be traced back to Portuguese settlements in the African west coast, where slaves and their exploiters used the pidgin. That common origin would explain some otherwise curious parallels in lexica and grammar. To take an example, creoles like Sranan, Negerhollands and Haitian (none of which developed in a Portuguese-speaking setting) all have a similar locative preposition *na* with a number of functions that resemble the Portuguese *na* (LOC.SG.FEM).

In contrast, Bickerton has argued that creoles exhibit shared properties because they represent an initial state of the possible grammars that are available to all humans, before language change and contact make them drift away from that configuration [27]. In order to introduce Bickerton's argument we should refer to another idea that suggests creole languages develop from pidgin languages via *nativization*: pidgins turn into a full-fledged (creole) language once they are learned as first language. Bickerton's battle horses are Hawai'ian Pidgin and Hawai'ian Creole. In the XIX century, a blooming sugar cane industry in the island witnessed a diverse migration of labor workers coming from China, the Philippines, Japan, other Pacific islands and European countries, besides the native population and the En-

glish. By the end of the century a more or less stable pidgin emerged as an ancillary communication means for the linguistically diverse community of workers. With striking speed, in perhaps a single generation, a creole spoken by the descendants of those workers could be identified. The commonalities between the original pidgin and the creole are considerable, which led Bickerton to argue that the pidgin learned as L1 (first language) was enriched with the full expressive machinery of any other natural language by means of a genetic blueprint characteristic of our species. Furthermore, Hawai'ian Pidgin seems to exhibit considerable variation according to the ethnolinguistic affiliation of the speakers — as an example, pidgin speakers of Japanese origins would use SOV order while Filipinos would produce VS sometimes, which turns out to be a transparent substrate effect since Japanese is SOV while many Philippines' languages are VS. In Hawai'ian Creole, basic word order seems to be a rigid SVO regardless of ethnolinguistic background, and many grammatical features that were largely optional before (like the marking of tense and aspect) now are obligatorily expressed. The fact that a handful of features (that will be discussed in the next section) seem to emerge in other creoles as well, following a hypothetically similar trajectory — that is, having a preceding pidgin stage— led Bickerton to conjecture that those might correspond to some sort of "default" configuration for languages, which in his early work was used to argue for creoles as a window to the process of language evolution.

There is an important issue with the sample of creoles on which Bickerton based his observations, though. Although the genealogy of creole languages cannot be accounted for in a transparent way by means of trees as with other (but certainly not all, and maybe not even the majority of) languages, no researcher would deny the fact that at least some of the typological features of creole languages are in close connection to those of its parents. This serves as a prelude to the issue being treated in this chapter: the imbalance in most of the compiled data on creoles is dramatic; the large majority of existing (and described) creole languages have as a lexifier at least one of a few West European languages — notably English, Dutch, Spanish, French and Portuguese — or languages from the Macro-Sudan belt as substrates [98]. Actually, when creoles outside of those ancestry groups are taken into account, violations to some presumed universal features of creoles can be found. Baker [11] mentions the case of Yilan Creole, a recently discovered language spoken in a few villages of Northeast Taiwan that emerged from the contact between Atayal (an Austronesian language) and Japanese [243]. Yilan Creole has a few properties (like SOV word order) that go against the proposals Bickerton (and others) have put forward in relation to the universal properties of creoles.

Those who eschew the idea of creoles having universal properties fail to agree on what is the relative relevance of a creole's ancestry in the genesis of its grammatical structure. The *relexification hypothesis* suggests that most creole languages are substrate grammars coated with a vocabulary from the lexifier [132]. Speakers of substrate languages had a very limited access to the lexifier, and in consequence they adopt labels that match in some way the references present in their own lexicon. Naturally this leaves out all the material for which an adequate pairing could not be established — notably functional words or morphology standing for grammatical features absent in the substrates, which would explain why creole languages seem to be isolating languages in general. Lefevebre proposes as a paradigmatic example the analysis of the semantics of the word *ansasinen* in Haitian Creole, which is a clear borrowing from French (*assassiner*). While in both languages the word has the meaning "to murder", in Haitian Creole it also means "to mutilate". Strikingly, Fongbe (one of the West African languages that served as a substrate) has a word (*hù*) with exactly that polysemy pattern, which suggests that it is still in place in Haitian Creole, although with a French label [132].

On the other hand, some have argued that creoles are actually regular offspring of their lexifiers with some influence from their substrates, similarly to Romance languages in relation to Latin [39]. One of the theories consistent with this idea places considerable weight on the original population of speakers of the lexifier, which, in analogy with genetics, is referred as the "founder principle" [168]. As with the relexification hypothesis, the inspiration for this idea comes from French-based creoles. A simplified account of the theory follows: the language of the first settlers (who many times spoke *koiné* varieties of the languages) would be learned with accuracy by the first groups of slaves. The success of the plantations increased the need for manpower, and soon the Europeans (and the first slaves) were outnumbered. This complicated the access to the language for the newcomers, and thus successive generations shaped the language according to these restrictions (and the linguistic features they brought in from their own languages). While it is true that the special circumstances of creoles might have accelerated the pace of language change, the kind of general processes they undergo are common to a large number of (non-creole) languages, and they receive the name of "contact-induced shift". In this respect, there is nothing special about creoles as a distinctive *linguistic* group.

In the last years there has been a revival of the idea of creole being a coherent group with universal (or quasi-universal) properties beyond the influence of their ancestry and far from ideas of hard-wired linguistic biases. The argument goes that creoles emerge in a situation of extreme pressure for communication, and as such some of the resources they display might be considered as natural or economic

against alternatives that attempt to handle other possible constraints as well (or that have accumulated over the years a number of not obviously functional features). In words or Bakker [12]:

> If we think of creoles as the result of retention (from superstrate most significantly a partial lexicon and grammar; and substrate most significantly in phonology and syntax), loss (from superstrate part of the grammar) and reconstitution (from neither substrates or superstrates), then creoles shed light on what properties are necessary in languages. Why do creoles almost universally develop tense and aspect systems, plurality, negation, an article system, etc. through grammaticalization, after they have been lost, but never e.g. gender? The solution may be found in pragmatic and cognitive saliency of certain semantic distinctions. In biological life, the existence of genders and sex are needed for the reproduction and survival of our species, but in reconstituted languages grammatical gender would be counterproductive, as it is altogether superfluous. [...]
>
> If one considers creoles unexceptional continuations of the lexifier, following the claim that the change from Latin to French was of the same magnitude as the change from French to Haitian, there is a big problem here. Why would there be no evidence of any creole with grammatical gender, not even within 40 to 400 years of development, whereas almost no Indo-European lost gender during 6,000 to 10,000 years? The explanation is obvious, and that is that creoles and pidgins never had gender. Pidgins started from zero, when the creators started adopting words and features they found useful from the lexifier and their mother tongues. The creoles built on that. No one found gender marking useful so they didn't adopt that. Creolization was preceded by a process of loss call it pidginization or simplification. And in contrast to other semantic distinctions, the creators of the creole wisely avoided the reconstitution of gender distinctions. Smart.

These ideas are often linked with the slippery concept of "complexity", that became installed in the discussions about creoles after the publication of a paper in 2001 with the stark title: "The world's simplest grammars are creole grammars" [156]. According to the researchers sympathetic to this idea, a part of the grammar of a language can be considered as more complex than the equivalent of another if the first involves more distinctions or elements (other things being equal). Thus, the ten tones of Trique (Oto-Manguean) are more complex than the five of Thai, and the fifteen cases of Finnish are

more complex than the two of Aleut (Eskimo-Aleut). Whether these distinctions bear any impact for the processing, production or acquisition of language is still to be demonstrated. Good makes the point that, in the light of an imperfect access to the ancestral languages, it makes sense to expect a *paradigmatic* rather than *syntagmatic* reduction in complexity [91]. Roughly, while the successful transmission of a paradigm — like the different pronouns of a language — involves the individual transmission of each of its members or generating rules, lexical items or constructions (like the passive) can in principle be passed unitarily. In his own words,

> Consider, for instance, what is required for the string of phonemes associated with cat to be transferred into a jargon from English. All that is needed is for /ket/ to be used in by a single member of the jargon community and for the other members to understand what is being referred to, at which point they may re-use this form to refer to similar entities. By contrast, for the English singular/plural distinction to be transferred, logically speaking, two linguemes must be transferred, e.g., cat and cats, since it is only possible for plural marking to enter the new contact variety if the coding of plurality via an affix is discoverable on the basis of the forms being used by its speakers. For this to happen, at least one singular and one plural noun must be transferred.

Last but not least, at this point it should be clear that there are two implicit formal models of creole profile being discussed in the literature. One consists in a more often than not fixed template of features to which creoles develop into or are born with. The bioprogram hypothesis and the pan-creole features of Bakker and Daval-Markussen [51] fall into this category. The other is instead of a more probabilistic nature and it involves perhaps a larger range of linguistic phenomena: there are certain features that we will expect to see more frequently in creoles, but it is the large number of the occurrences of those which characterizes creoles in toto instead of a few fixed traits. Researchers that discuss creoles from the vantage point of complexity endorse this characterisation.

## 4.3 A STATISTICAL TURN

In perspective, the gains for settling the issue of whether there is a creole profile cannot be overestimated. If there is such a thing, then creoles would be the most outstanding natural experiment to tell apart what is necessary from what is contingent in the structure of languages. If not, it would argue in favour of the idea that cultural evolution of language is strong and robust even in the extreme

situations of creole origin. Beyond the (already substantial) theoretical profit, there are concrete practical consequences for having the answer. Daval-Markussen, after determining a set of diagnostic features for the creole profile, find that Hmong Njua (a Hmong-Mien language from Laos) complies with it [50]. Reviewing genetic and linguistic evidence he finds that the case for that language being a creole is considerable. In other words, a positive answer to the creole profile question would provide us with a relatively inexpensive way of making inferences about the sociocultural past of languages for which direct knowledge might be limited.

Given the importance of the topic, and from an outsider perspective, it might seem puzzling that such a large number of (in principle) mutually excluding theories on the emergece of creoles could exist. But testing the nature of the origin of creoles and, in particular, whether there is such a thing as a creole profile are conspicuously difficult tasks. For starters, most creoles developed in a very short amount of time — a few generations versus centuries in the case of West European languages — and they usually lack appropriate documentation. Most of what is known about the early stages of creoles comes from religious orders (that attempted to acquire the language for spreading the New Testament in the populations) and secular Europeans associated with the colonies.

With respect to the substrate languages the situation is close to helpless but due to a different reason: most of the time there is only very vague information about which languages those were in the first place. Labor workers and slaves involved in creoles came from virtually every corner of the planet — just in the West Indies one could find speakers of East Asian, Pacific, West European and Native American languages. True, a large number of slaves came from the Atlantic coast of Africa and in particular from the Guinea Region, which, unfortunately for the scientific question at hand, happens to be the second most linguistically diverse area in the world after Papua New Guinea. In this context, proponents of the relexification hypothesis have been accused of triggering the *Cafeteria Problem*: by looking hard enough among the many African languages of that region, one would be able to find at least one that matches some aspect of the creole being studied [167].

While lexifiers are in a better position as comparison standards for creoles (since West European languages are the most studied languages in linguistics) it has been argued that the varieties that contributed to the genesis of creole languages were not the standard varieties but were instead particular sociolects that sometimes could diverge from their source. An interesting case is the variety of English spoken by pirates in the Caribbean just before the explosion of sugar cane plantations [53]. It is thought that groups of aboriginal Americans and African slaves might have picked up their first European

languages through contact with pirates. Pirate's language was heavily influenced by other low prestige varieties of English as well as the languages of the other peoples involved in maritime trade. Pirates were skilled in multiple languages apparently; the infamous Henry Morgan spoke English, Welsh and French. Strikingly, in the pirate ships' logs it is possible to find features that are absent in standard English but present in many of its creoles, such as for instance the inflection patterns of copula verbs.

Finally, it is in general not clear what would be the value of predictions built on generalizations (and theories) of language change of non-creole languages. Whether the study of long-scale processes — like the great vowel shift, that remodelled the vowel space of English over a time lapse of 300 years — can tell us anything about creole development, that typically occurs in a few generations, is still to be answered. On top of that (and perhaps barring some changes in phonological systems) even for non-creole languages the number of diachronic laws is modest, and they are almost exclusively better casted as tendencies instead of deterministic forces.

This state of affairs redirected the attention from purely linguistic to more statistical grounds. The implicit (and sometimes perhaps concealed) philosophy behind the quantitative and statistical work I will review in the next section is that we can attempt to solve the creole profile question based on our (statistical) analysis of data. If a reasonable analysis shows that some features are attached to creoles and less so to non-creoles, then we could call that the creole profile, and the results thus obtained should fuel the development of a theory about creoles.

## 4.4 CAVEATS FOR THE TESTING OF THE CREOLE PROFILE

The first natural step in this analysis is to evaluate the characteristics of the data on creoles at hand. To the extent of my knowledge, all databases where creole typological information features extensively have been built with those particular languages in sight, in contrast to other typological databases that sacrifice the description of potentially interesting traits for a region or a phylogeny for the sake of increased cross-linguistic comparability. This has likely an impact on the chosen features. For instance, although about one third of the languages in the world are described as displaying an ergative alignment system, no creole database systematically codes for this, the reason being very probably that there does not seem to be any creoles falling within that category [156]. On the other hand, while the presence of prenasalized stops is not coded in many cross-linguistic databases, it is usually brought into question regarding creoles - which does not come as a surprise since many West African languages (which are the usual substrates) have them. All this means that, from the onset, the subset

of variables present in those databases share a bias towards aspects that are deemed to be of interest for creolists. It should not come as a surprise then if a handful of those features are enough to draw a typological line between creoles and non-creoles, as it was found [51].

Even more, different aspects of grammars tend to be not represented with the same number of detail or through the same number of entries in these databases. On top of that there are functional and diachronic affinities between features (see Chapter 2). Ultimately this translates into highly correlated and redundant features. This could be simply solved by studying those correlations from independent samples. However, creoles pose a formidable challenge: determining the correlations of features from creole data assumes the independence of those languages, but in principle the complementary reading is plausible as well: it might be that the features are correlated as a result of those languages being creoles. Inferring the dependencies from non-creole languages also runs into the same conundrum.

To worsen the situation, the range of the number of variables analyzed by creolists is considerably wide, and this has never been explicitly taken into account. Trivially, a larger number of available dimensions is more likely to yield a separating criterion for any proposed classification.

Given this cumbersome landscape, the best strategy to adopt is not to screen the variables at all. The idea would be that if there is any bias in the distribution, number or correlational structure of a set of features, then this should translate into a better classificatory power of *any* arbitrary group of independent languages. Comparing how well we can distinguish creoles from a structurless sample of languages of the world in contrast to random assignments of languages into two groups appears as a sensible solution for this issue.

As a second stage, claims about any profile-defining features should address the question of whether the affiliation of creoles to their ancestral languages explains the regularities. From a purely data-based perspective, creole-defining features should be ostensible independent from their ancestry unless there is a good alternative explanation that could account the accidental coincidence. Repeating the previous testing without those features for which there is substantial statistical evidence of their association with an ancestral lineage will be a litmus test for a robust creole profile.

## 4.5 EXPLORATORY ANALYSES AND ANCESTRY-RELATED FEATURES

### 4.5.1 *Data*

For all further tests I use the largest published dataset on the structure of creole languages, the Atlas of Pidgin and Creole Languages (APiCs) [160]. This survey shares a number of typological features

that can be readily compared with another large database of languages, the World Atlas of Language Structures (WALS) [103], which was introduced already in Chapter 2. This set of common features comprises properties of nominal and verbal phrases, word order patterns, clausal syntax, a bit of phonology and locus of coding. While this implies a considerable loss of information in comparison with the full APiCs (which has 131 features in total), a quick assessment of the abandoned variables shows that those were the most idiosyncratic to creoles and pidgins, so this strategy helps alleviating the bias discussed before.

As a comparison group I choose a genealogically balanced sample with particularly good coverage in the WALS for the set of shared languages. None of the languages of the sample have been shown to be a creole[2], and they all belong to different genera.

A few languages from the APiCs database were removed due to them being represented by means of a very close dialect or variety in order to avoid over-counting essentially the same language. After removing both languages and features with low coverage, I ended up with a set with 53 creole languages and a balanced sample of 106 non-creole languages. 41 out of the 48 shared features between APiCs and WALS have more than a 75% in our sample and thus were retained for analysis. About 8% of the relevant feature values were missing in the aggregated data. Most of those come from the non-creole set (which exhibits 12% of missing entries); creoles have an almost perfect coverage, with less than 2% of missing data (see Table 10).

### 4.5.2 *Exploratory analysis*

An exploratory analysis of the data was conducted to detect eventual oddities and to have a better grasp of the typological dispersion of the languages at hand.

The first question that could be readily answered is whether there is any evidence for the idea that the distribution of features within the creoles' set is not readily comparable to the typical world-wide distribution in non-creole languages. As a rough model, I approximated the later with the Laplaced-smoothed empirical frequencies observed in the whole WALS — not only on the reference subset described before — and then evaluated the P-value corresponding to a multinomial test applied to the creole languages. The results can be observed in Table 11.

The majority of features are distributed in a way that is decidedly different from the worldwide, non-creole sample. Between a fifth and

2 However, Chamorro has been regarded in the past a Spanish creole due to the large amount of vocabulary it took from that language, and Hmong Njua has been flagged as a potential creole based on the work of Daval-Markussen (as discussed in the main text)

| Feature description | Levels | Creole cov. | WALS cov. |
|---|---|---|---|
| Order of subject, object and verb | 6 | 1.00 | 0.95 |
| Order of genitive and noun | 2 | 1.00 | 1.00 |
| Order of adjective and noun | 4 | 1.00 | 0.98 |
| Order of adposition and NP | 5 | 1.00 | 0.98 |
| Order of demonstrative and noun | 4 | 1.00 | 0.98 |
| Order of numeral and noun | 4 | 1.00 | 0.95 |
| Order of RC and noun | 6 | 1.00 | 0.90 |
| Pos. of interrogative phrases | 3 | 1.00 | 0.92 |
| Gender in indep. pers. pronouns | 5 | 1.00 | 0.99 |
| Inclusive/exclusive dist. in indep. pers. pron. | 4 | 1.00 | 1.00 |
| Politeness in pronouns | 4 | 0.85 | 0.94 |
| Indefinite pronouns | 5 | 1.00 | 0.74 |
| Occurrence of nominal plurality | 6 | 1.00 | 0.72 |
| Coding of nominal plurality | 8 | 1.00 | 0.97 |
| Definite articles | 5 | 1.00 | 0.87 |
| Indefinite articles | 5 | 1.00 | 0.82 |
| Pronominal and adnominal demonstratives | 3 | 1.00 | 0.75 |
| Distance contrasts in demonstratives | 5 | 1.00 | 0.86 |
| Ordinal numerals | 8 | 0.89 | 0.84 |
| Numeral classifiers | 3 | 1.00 | 0.74 |
| Locus of marking in pos. NP | 5 | 1.00 | 0.89 |
| Suppletion acc. to tense and aspect | 4 | 1.00 | 0.96 |
| Prohibitive | 4 | 0.98 | 0.92 |
| Alignment of case marking and full NP | 5 | 1.00 | 1.00 |
| Alignment of case marking of pron. | 7 | 1.00 | 0.93 |
| Ditrans. constructions, give | 4 | 1.00 | 0.90 |
| Expression of pronominal subjects | 6 | 1.00 | 0.89 |
| Comitatives and instrumentals | 3 | 0.98 | 0.72 |
| Nominal and verbal conjunction | 3 | 1.00 | 0.80 |
| Zero copula for predicative nominals | 2 | 1.00 | 0.83 |
| Nominal and locational predication | 2 | 1.00 | 0.83 |
| Intensifiers and relfexive pronouns | 2 | 0.89 | 0.75 |
| Reciprocal constructions | 4 | 0.96 | 0.73 |
| Applicative constructions | 7 | 1.00 | 0.93 |
| Relativization on subjects | 4 | 0.89 | 0.79 |
| Want complement subjects | 5 | 1.00 | 0.73 |
| Negative morphemes | 6 | 0.85 | 0.97 |
| Neg. indef. pron. and predicate neg. | 4 | 1.00 | 0.67 |
| Polar questions | 6 | 1.00 | 0.95 |
| Tone | 3 | 0.98 | 0.96 |
| Vowel nasalization | 2 | 1.00 | 0.93 |

Table 10: **Number of levels and coverage of the grammatical features analyzed in APiCS and WALS**

| Feature | P-value |
|---|---|
| Order of subject, object and verb | $< 10^{-3}$ |
| Order of adjective and noun | $< 10^{-3}$ |
| Order of adposition and NP | $< 10^{-3}$ |
| Order of numeral and noun | $< 10^{-3}$ |
| Order of RC and noun | $< 10^{-3}$ |
| Pos. of interrogative phrases | $< 10^{-3}$ |
| Inclusive/exclusive dist. in indep. pers. pron. | $< 10^{-3}$ |
| Indefinite pronouns | $< 10^{-3}$ |
| Occurrence of nominal plurality | $< 10^{-3}$ |
| Coding of nominal plurality | $< 10^{-3}$ |
| Definite articles | $< 10^{-3}$ |
| Indefinite articles | $< 10^{-3}$ |
| Pronominal and adnominal demonstratives | $< 10^{-3}$ |
| Distance contrasts in demonstratives | $< 10^{-3}$ |
| Ordinal numerals | $< 10^{-3}$ |
| Numeral classifiers | $< 10^{-3}$ |
| Locus of marking in pos. NP | $< 10^{-3}$ |
| Prohibitive | $< 10^{-3}$ |
| Alignment of case marking and full NP | $< 10^{-3}$ |
| Alignment of case marking of pron. | $< 10^{-3}$ |
| Ditrans. constructions, give | $< 10^{-3}$ |
| Expression of pronominal subjects | $< 10^{-3}$ |
| Comitatives and instrumentals | $< 10^{-3}$ |
| Applicative constructions | $< 10^{-3}$ |
| Relativization on subjects | $< 10^{-3}$ |
| Want complement subjects | $< 10^{-3}$ |
| Negative morphemes | $< 10^{-3}$ |
| Polar questions | $< 10^{-3}$ |
| Order of genitive and noun | 0.001 |
| Tone | 0.002 |
| Nominal and locational predication | 0.006 |
| Reciprocal constructions | 0.019 |
| Suppletion acc. to tense and aspect | 0.029 |
| Neg. indef. pron. and predicate neg. | 0.041 |
| Gender in indep. pers. pronouns | 0.053 |
| Nominal and verbal conjunction | 0.069 |
| Vowel nasalization | 0.075 |
| Order of demonstrative and noun | 0.170 |
| Zero copula for predicative nominals | 0.434 |
| Intensifiers and reflexive pronouns | 0.485 |
| Politeness in pronouns | 0.582 |

Table 11: **P-values of creole features in relation to their worldwide typicality.**

a quarter of all features will be rejected under conventional statistical thresholds.

For the purpose of investigating the internal variation within the creole languages, I used a simple Gower similarity, which is defined as the fraction of shared features between two languages that have the same value in each of them,

$$G(i,j) = \frac{\sum_{\phi \in \Phi_{ij}} \delta(\phi_i = \phi_j)}{|\Phi_{ij}|} \tag{13}$$

where $i$ and $j$ parametrize languages and $\Phi_{ij}$ is the set of features that are coded for languages $i$ and $j$.

There are no two languages in our sample with similarity equal or larger than 0.85. By taking all the languages with a similarity of 0.8 at least we see a clear clustering of some creoles according to their lexifiers, as shown in Figure 16.

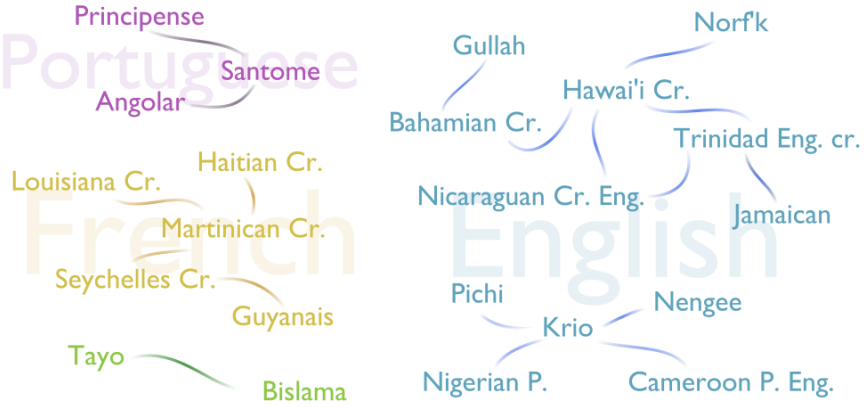


Figure 16: **Superficial creole clusters**

Similarity graph for creole languages with over 80% of shared features coded with the same value. Labels indicate the common lexifier of the languages.

Superficial clustering techniques like this cannot be taken as anything else than suggestive evidence that cannot stand alone as a proof. All features contribute equally likely to the aggregated similarity, whereas the expectation is that there might be a dramatic difference in the value certain properties have as diagnostic of creolness. More critically, there is no treatment to the obvious collinearity patterns present in the features, as discussed in the previous section.

As a second exploratory step, I studied how well individual variables predicted class membership of languages (creole/non-creole) and viceversa. A few univariate measures of association for categorical values between typological variables and the class label were used. $G^2$ tests of independence (based on the $\chi^2$ approximation) were per-

formed on the whole set, as well as regular significance testing of the specific measures described next.

Goodman-Kruskal $\lambda$ and $\tau$ statistics were used in order to evaluate the pairwise predictive power. Goodman-Kruskals $\lambda$ is the relative reduction of classification error due to the independent variable $X$ given the baseline of the modal frequency of the dependent variable $Y$,

$$\lambda(x;y) = \frac{1 - \max_y \Pr(Y = y) - \sum_i \max_y \Pr(Y = y|X = x_i)}{1 - \max_y \Pr(Y = y)}$$

Goodman-Kruskals $\tau$, instead, uses as baseline the strategy of choosing each value in the support of $Y$ according to its probability. If we call $\hat{f} = \sum_{y \in \mathcal{Y}} \Pr(Y = y)^2$, then

$$\tau(x;y) = \frac{1 - \hat{f} - \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \Pr(Y = y|X = x)^2}{1 - \hat{f}} \tag{14}$$

$\tau$ and $\lambda$ are not symmetric. When the direction of prediction is from the feature values to the membership I use the subscript 'F', and 'M' in the other case. Results can be observed in Table 12.

For about 85% of the features there is a detectable association between their values and the language membership. However, for only 25% of the features the gain in predictive power (for at least one of the measures and directions) is meaningful.

Critically, it is important to screen out those features for which some evidence of ancestry origin could be established. I mentioned before that pinning down individual languages from WALS acting as lexifiers and substrates is unfeasible. However, we do have access to more general information about the genealogical (or typological) groups to which they belonged to. The idea is then simply to establish whether individual features show sufficient dependency with *any* group of their ancestry and flag them as potentially inherited. There is a tradeoff between how specific these levels can be and the possibility of statistically demonstrating any dependency, so for this reason I latched to whatever grouping level was available where some variation could be appreciated. Naturally, statistical associations can be detected only when there is enough variation in the data. This does not rule out the possibility that there is a direct influence of the ancestry over creoles for those features, but just that the point exceeds what can be argued based on the primary analysis of the data solely.

At the lexifier level, a difference between Germanic and Romance languages can be readily established within the West European languages. Austronesian and Macro-Sudan languages also contribute significatively to the substrates. Finally, the European/Macro-Sudan "biclan" — implying European lexifier as well as substrate from Macro-Sudan— is by far the most commonly attested combined pair of ancestry lineages.

| Feature | $\lambda_F$ | $\lambda_M$ | $\tau_F$ | $\tau_M$ |
|---|---|---|---|---|
| Comitatives and instrumentals | 0.54 | 0.51 | 0.45 | 0.31 |
| Occurrence of nominal plurality | 0.53 | 0.35 | 0.37 | 0.18 |
| Expression of pronominal subjects | 0.47 | 0.37 | 0.41 | 0.19 |
| Indefinite pronouns | 0.43 | 0.38 | 0.37 | 0.24 |
| Indefinite articles | 0.38 | 0.27 | 0.36 | 0.13 |
| Prohibitive | 0.36 | 0.21 | 0.34 | 0.15 |
| Order of subject, object and verb | 0.34 | *0.19* | 0.35 | 0.20 |
| Coding of nominal plurality | 0.32 | 0.23 | 0.30 | 0.09 |
| Ditrans. constructions, give | 0.30 | *0.15* | 0.24 | 0.09 |
| Pronominal and adnominal demonst. | 0.26 | *0* | 0.14 | *0.08* |
| Order of adposition and NP | *0.24* | 0.33 | 0.33 | 0.25 |
| Locus of marking in pos. NP | *0.19* | *0* | 0.18 | 0.04 |
| Polar questions | *0.19* | *0.03* | 0.16 | 0.06 |
| Distance contrasts in demonstratives | 0.17 | *0* | 0.17 | 0.05 |
| Want complement subjects | 0.17 | *0* | 0.23 | 0.06 |
| Relativization on subjects | 0.13 | *0* | 0.14 | 0.03 |
| **Reciprocal constructions** | *0.06* | *0* | *0.04* | *0.01* |
| Definite articles | *0.06* | *0.06* | 0.16 | *0.09* |
| **Neg. indef. pron. and predicate neg.** | *0.06* | *0* | *0.05* | *0.01* |
| Ordinal numerals | *0.04* | *0* | 0.13 | 0.02 |
| Pos. of interrogative phrases | *0.04* | 0.28 | 0.14 | 0.10 |
| Order of genitive and noun | *0.02* | *0.13* | 0.09 | 0.13 |
| Order of RC and noun | *0.02* | *0* | 0.18 | 0.09 |
| Tone | *0.02* | *0* | 0.07 | 0.02 |
| Order of adjective and noun | *0* | 0.20 | 0.07 | 0.05 |
| **Order of demonstrative and noun** | *0* | *0* | *0.01* | *0* |
| Order of numeral and noun | *0* | *0* | 0.10 | 0.08 |
| Gender in indep. pers. pronouns | *0* | *0* | *0.04* | 0.01 |
| Incl./excl. dist. in indep. pers. pron. | *0* | *0* | 0.09 | 0.05 |
| **Politeness in pronouns** | *0* | *0* | *0.03* | *0* |
| Numeral classifiers | *0* | *0* | 0.08 | 0.05 |
| **Suppletion acc. to tense and aspect** | *0* | *0* | *0.03* | *0.01* |
| Alignment of case marking and full NP | *0* | *0* | 0.14 | 0.08 |
| Alignment of case marking of pron. | *0* | *0.07* | 0.10 | 0.07 |
| Nominal and verbal conjunction | *0* | *0* | *0.04* | *0.09* |
| **Zero copula for predicative nominals** | *0* | *0.02* | *0.01* | *0.01* |
| **Nominal and locational predication** | *0* | *0* | *0.02* | *0.02* |
| **Intensifiers and reflexive pronouns** | *0* | *0* | *0.01* | *0.01* |
| Applicative constructions | *0* | *0* | 0.15 | 0.09 |
| Negative morphemes | *0* | *0* | 0.19 | 0.10 |
| **Vowel nasalization** | *0* | *0* | 0.02 | *0.02* |

Table 12: Goodman-Kruskal's statistics on the prediction of membership from features (F-index columns) and the other way around (M-index columns).

Simple Fisher's exact tests between each feature and ancestry group were performed. For instance, consider the case of the relative order of the genitive and the noun and the lexifiers (Table 13)

| Order of genitive and noun | Germanic lex. | Romance lex. | Other |
|---|---|---|---|
| GenN | 13 | 3 | 1 |
| NGen | 3 | 19 | 5 |
| Both equally likely | 8 | 1 | 0 |

Table 13: Creole counts for the order of genitive and noun according to the lexifier

Without the need of invoking any theory of feature transmission, our data strongly suggests that lexifiers do play a role in the manifestation of this variable in their corresponding creoles. Strikingly, for the majority of features this is the case as well, as it can be seen in Table 15.

## 4.6 PROBABILISTIC PROFILE CLASSIFICATION

### 4.6.1 *Random forests*

As discussed before, one possible way in which creolization could have impacted the typology of a language is by means of biasing probabilistically different properties without the need of imposing any deterministic template. In this case, the effort is geared towards modelling the conditional probability distribution of the class with respect to the variables,

$$\Pr(\text{class}|F_1, F_2, \ldots, F_N) \tag{15}$$

As a way of tackling this problem, for instance, one could estimate unseen feature combinations by imputing a binominal distribution based on the sample distribution of k-nearest neighbors, where k would be selected by cross-validation or some other classification efficiency measure. Setting aside the issue that the data is quite sparse, I find there is a more troubling conceptual problem with such procedures. By taking nearest neighbors (or any other strategy consisting on local estimates) we become prey of the democracy of features, as I discussed before in relation to clustering techniques by focusing on the feature correlation issue. Some of the languages in our sample — like the two varieties of Chabacano, Ternate and Zamboanga — are different enough to be considered as independent points yet extremely similar in other respects, and they might perfectly be regarded as dialects. Presumably, the commonalities among them go beyond a hypothesized creole profile since they have been in contact with similar languages as well. Because of that, even a very efficient

| Feature | Lexifier | Substrate | Biclan |
|---|---|---|---|
| Order of subject, object and verb | 0.42 | 0.01 | 0.03 |
| Order of genitive and noun | $< 10^{-2}$ | 0.38 | 0.16 |
| Order of adjective and noun | $< 10^{-2}$ | 0.65 | 0.85 |
| **Order of adposition and NP** | 0.36 | 0.36 | 0.34 |
| Order of demonstrative and noun | 0.01 | 1.00 | 1.00 |
| Order of numeral and noun | $< 10^{-2}$ | 0.08 | 0.07 |
| **Order of RC and noun** | 0.74 | 0.22 | 0.08 |
| Pos. of interrogative phrases | $< 10^{-2}$ | $< 10^{-2}$ | $< 10^{-2}$ |
| Gender in indep. pers. pronouns | 0.01 | 0.20 | 0.18 |
| Incl./excl. dist. in indep. pers. pron. | 0.41 | $< 10^{-2}$ | 0.04 |
| **Politeness in pronouns** | 0.25 | 0.12 | 0.23 |
| Indefinite pronouns | 0.97 | 0.05 | 0.30 |
| Occurrence of nominal plurality | 0.02 | 0.60 | 0.62 |
| Coding of nominal plurality | $< 10^{-2}$ | 0.02 | 0.04 |
| Definite articles | $< 10^{-2}$ | 0.01 | $< 10^{-2}$ |
| Indefinite articles | $< 10^{-2}$ | 0.02 | 0.01 |
| Pronominal and adnominal demonst. | 0.41 | 0.06 | 0.02 |
| **Distance contrasts in demonstratives** | 0.26 | 0.11 | 0.36 |
| **Ordinal numerals** | 0.12 | 0.48 | 0.37 |
| **Numeral classifiers** | 0.47 | 0.34 | 0.58 |
| Locus of marking in pos. NP | $< 10^{-2}$ | $< 10^{-2}$ | $< 10^{-2}$ |
| **Suppletion acc. to tense and aspect** | 0.45 | 0.80 | 0.42 |
| Prohibitive | 0.06 | $< 10^{-2}$ | $< 10^{-2}$ |
| Alignment of case marking and full NP | 0.01 | $< 10^{-2}$ | $< 10^{-2}$ |
| Alignment of case marking of pron. | 0.02 | 0.42 | 0.24 |
| Ditrans. constructions, give | 0.09 | $< 10^{-2}$ | $< 10^{-2}$ |
| Expression of pronominal subjects | $< 10^{-2}$ | $< 10^{-2}$ | $< 10^{-2}$ |
| **Comitatives and instrumentals** | 0.08 | 0.14 | 0.12 |
| Nominal and verbal conjunction | 0.03 | 0.17 | 0.77 |
| Zero copula for predicative nominals | 0.21 | 0.05 | 0.03 |
| Nominal and locational predication | 0.03 | 0.49 | 0.78 |
| Intensifiers and relfexive pronouns | 0.01 | 0.03 | 0.21 |
| **Reciprocal constructions** | 0.18 | 0.62 | 1.00 |
| Applicative constructions | 0.01 | 0.31 | 0.18 |
| **Relativization on subjects** | 0.22 | 0.50 | 0.26 |
| **Want complement subjects** | 0.40 | 0.19 | 0.15 |
| **Negative morphemes** | 0.35 | 0.70 | 0.24 |
| Neg. indef. pron. and predicate neg. | 0.04 | 0.03 | 0.08 |
| **Polar questions** | 0.10 | 0.64 | 0.49 |
| Tone | 0.04 | 0.01 | $< 10^{-2}$ |
| Vowel nasalization | 0.01 | 0.19 | 0.38 |

Table 15: Association between features and ancestry group. The numbers are P-values from Fisher's exact test. Feature labels in bold reflect no association detected for any of the three ancestry groups.

estimate of the conditional probability distribution might be latching on the *wrong* features.

Our estimate method should make use of features for which *global* evidence of an association with the creole profile exists. Because of this, I turn my attention to a powerful classification technique with state-of-art performance: random forests.

Random forests consist of an ensemble of unbiased binary conditional independence trees (CIT) [113]. CIT are decision trees that are the outcome of the following recursive algorithm

1. Find the explanatory variable that is most strongly associated with the response variable. Here this was done by means of choosing the minimum P-value from all the dependency tests between response and explanatory variables.

2. Partition the space in two parts according to the explanatory variable found in the previous step, in such a way that the association between the values of this variable and the response is maximized.

3. For each of the two partitions induced in the previous step, go to 1) and repeat the process until no variable is significantly associated with the response.

In our case, such a procedure would induce an hypothetical tree as the one schematically shown in Figure 17

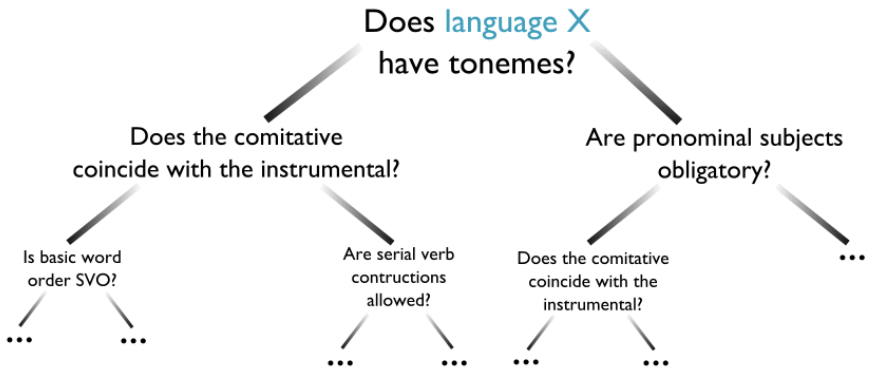


Figure 17: **Classification tree**

Schematic representation of a conditional inference tree (CIT).

By using only those features that show a correlation at the global level, we guarantee that the induced probability distribution is then theoretically relevant. It is crucial to notice that the correlation is calculated by assuming independent languages. This is a very reasonable assumption for the WALS sample, but some — those that defend some degree of affiliation between creoles and their ancestry

— might object that this is exactly not the case. However, the conflict should be resolved by this same process: if creolness trumps ancestry and it brings stronger and more reliable cues, then it should be the case that its associated features will emerge often in the construction of the trees.

CIT is a powerful technique, although it induces hard decision boundaries in the data. This can be solved by means of generating an ensemble of trees and then combining the individual predictions, a general machine learning strategy referred as *boosting* [205]. Such collection of trees is dubbed a forest. The tests used in our case follow [218] in order to guarantee that the number of levels a variable has does not influence its likelihood of being chosen; this also determined that about 60% of the data points (languages in our case) are sampled without replacement for every tree.

In definitive, I used random forests to determine how efficiently creoles can be distinguished from non-creoles in our data. I produced a comparison baseline by randomizing the classes labels 50 times for each of the 30 imputations and evaluating the classificatory properties of random forests on each of the sets.

### 4.6.2 *Results*

Applying these techniques to our data require complete datasets. I have introduced before the characteristics of a multiple imputation scheme based on a Gibbs' sampler and a conditional multinomial model — the same algorithm was used here [207]. Data were imputed independently for creoles and non-creole languages — while in practical terms this was observed to have little impact on the actual filled values, formally it guarantees that if there is any bias introduced by the procedure this would go in the direction of making both groups more homogenous internally. $n = 30$ independently imputed sets were produced. The imputed sets share on average 95% of their feature values and no relevant variance whatsoever.

When using all features, the full set of creoles can be distinguished from non-creoles efficiently, with good OOB precision and recall values (see Table 16 and Figure 18). In contrast, none of the 100 random groups per imputation matched any of the indices, and they exhibit poor classificatory properties (see Table 16).

Applying the same procedure now only to the features not flagged as ancestry-dependent, we get the results in Figure 19

Again, a distinction between creoles and non-creoles can be determined on a statistical basis, with the random case being considerably worst than the empirical (see Table 16).

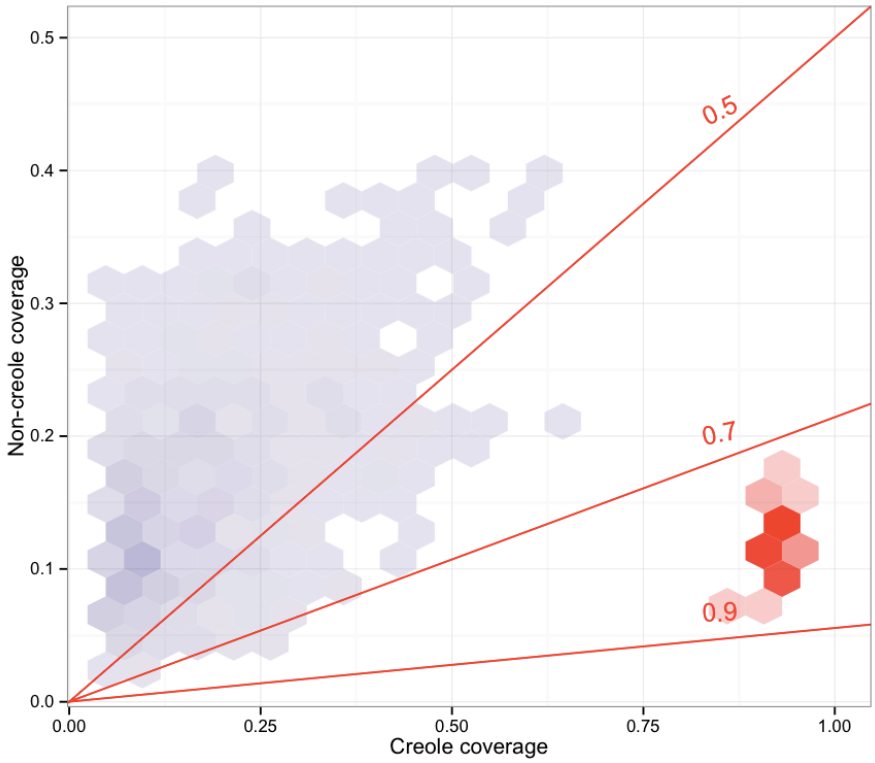


Figure 18: **Random forest classification with all features**

Distributions for the creole and non-creole coverages for the empirical (red) and the random (blue) classification based on random forests and all variables. Color intensity increases with density of all datasets analyzed for that group. The red lines correspond to classifiers of precision 0.9, 0.7 and 0.5 (from bottom to top).

### 4.6.3 *Variable importance*

Once a forest has been constructed, it is possible to measure the relative importance every variable has for the purpose of classification. In every tree built for the random forest, a number of observations are left out and are not being used for the construction of the classifier — these receive the name of Out Of Bag (OOB) observations. For each predictor, then, some metric of classification accuracy of the OOB sample is applied to each tree, which is then contrasted with the same measure applied after the permutation of the predictor. The idea is clear: if a predictor does not have any bearing on the the dependent variable, then the permutation should not impact the classification accuracy.

There are two further aspects of this procedure. As discussed at the beginning of this section, random forests were used to handle the potential correlations among predictors. A simple approximate

| | | Recall | | Precision | |
|---|---|---|---|---|---|
| Features | Type | mean | SD | mean | SD |
| All | Empirical | 0.92 | 0.01 | 0.79 | 0.03 |
| All | Random | 0.10 | 0.11 | 0.21 | 0.18 |
| Reduced | Empirical | 0.91 | 0.01 | 0.74 | 0.04 |
| Reduced | Random | 0.07 | 0.11 | 0.15 | 0.19 |

Table 16: Precision and recall values of the creole class with random forests, in both empirical and random cases for the whole and reduced set of features.

solution to this issue in the determination of variable importance is to perform the permutations of the predictors by blocks that correspond to the levels of its corresponding covariates. Heuristically, I used the rule $P \leq 0.05$ to define the set of relevant covariates. Notice that the output of the algorithm could be dramatically affected by the threshold: in a case of strongly correlated independent variables, the multiplicity of variables on which it has to be conditioned on for the permutation could lead to a case in which there is virtually *no* difference between the permuted and the original variable.

Finally, imbalanced classes might produce inflated variable importance measures. If a particular class is extremely frequent, permutation of a few predictors is not expected to change drastically the classification accuracy of this class (since random forests will anyhow try to assign most of the instances to this class). On the contrary, the minority class, which is perhaps detected thanks to a few variable combinations, is likely to be less robust to permutations. These considerations are not built in into the usual estimators of variable importance. A useful alternative is to calculate, before and after the permutations, the AUROC (Area Under Receiver Operator Characteristic), which is simply the fraction of OOB pairs of different classes to which the algorithm correctly predicts their membership [122].

In sum,

$$\mathrm{VI}_j = \frac{1}{|\mathrm{trees}|} \sum_i \mathrm{AUROC}_{ij} - \mathrm{AUROC}_{i\pi(j)} \tag{16}$$

### 4.6.4 *Results*

After pruning all those variables with marginal variable importance, a small number of them seems to carry the largest independent contribution to the classification, as it can be seen in Figure 20

As it can be witnessed, there is some variation in the variable importance scores across imputations.

The evaluation of variable importance for the set of independent features is completely consistent, as seen in Figure 21

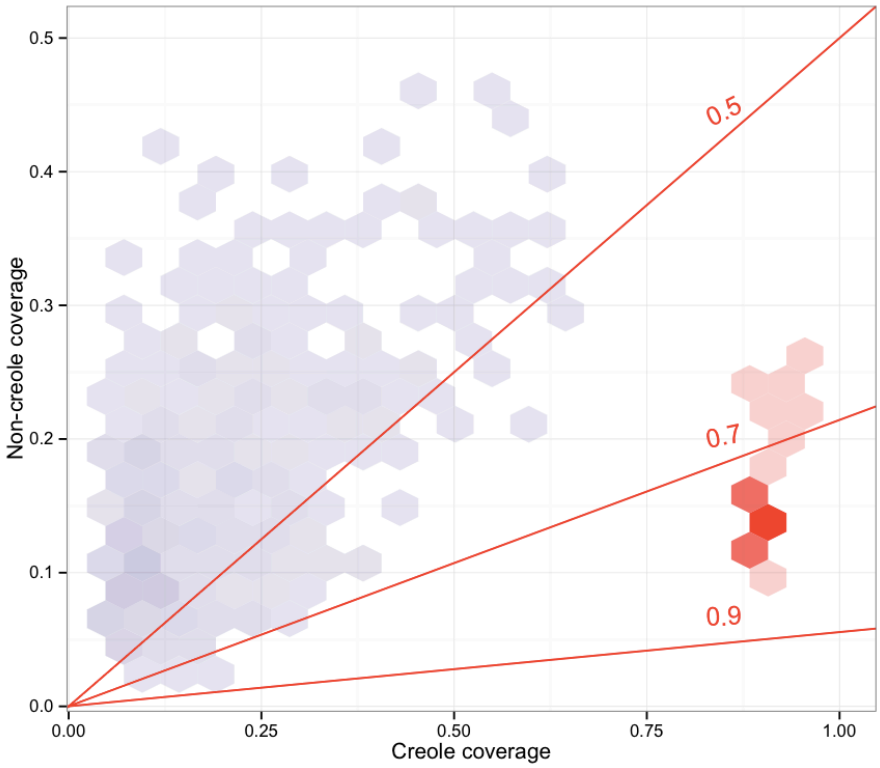


Figure 19: **Random forest classification with non-genealogical features**

Distributions for the creole and non-creole coverages for the empirical (red) and the random (blue) classification based on random forests and only those variables for which no genealogical association could be established. Color intensity increases with density of all datasets analyzed for that group. The red lines correspond to classifiers of precision 0.9, 0.7 and 0.5 (from bottom to top).

The relative order of values is the same as in the set with all features, once the ancestry-dependent ones are removed.

## 4.7 PROTOTYPE PROFILE CLASSIFICATION

### 4.7.1 *Associations rule mining*

In this scenario the goal is to find a *specific* configuration of features —the prototype— with the ability of predicting creole languages and viceversa. Instead of determining, over all features and values, what is the probability that a particular configuration corresponds to a creole language, here I will seek to find the best fixed combination of

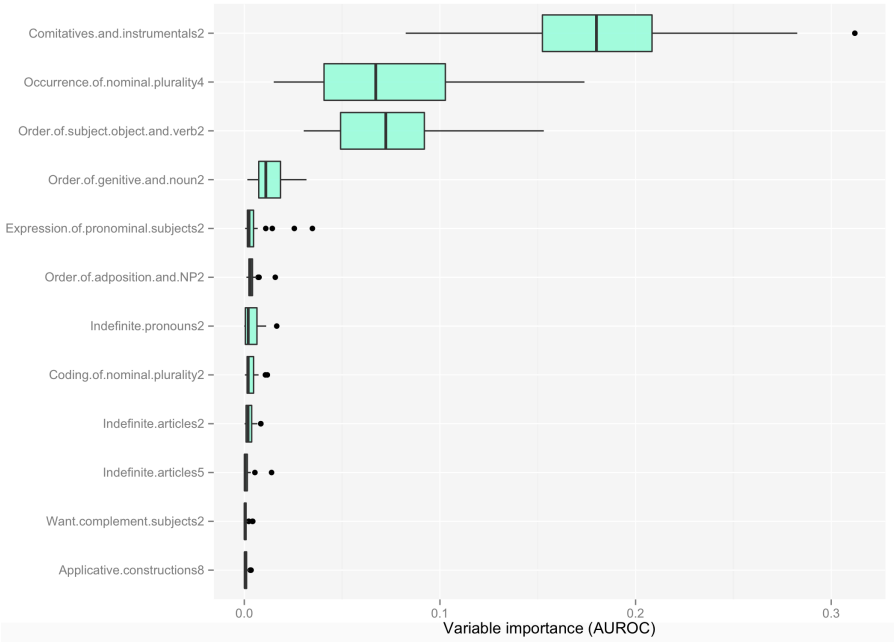


Figure 20: **Variable importance based on all features**

Boxplots of variable importance as inferred from a random forest based on all features.

feature values. Statistically speaking, a good prototype (involving feature values $F_1 = f_1, F_2, \ldots, F_n$) will be one with high values of

$$\Pr(\text{class} = \text{creole} | F_1 = f_1, F_2, \ldots, F_n)$$

and

$$\Pr(F_1 = f_1, F_2, \ldots, F_n | \text{class} = \text{creole})$$

In other words, the first conditional probability determines the recall of a particular creole prototype (i.e. what is the fraction of all creoles that respond to that precise pattern) whereas the second is directly related to precision — namely, what is the fraction of creoles over all languages that exhibit those properties. A high value of recall can be obtained if a particular feature value is shared across all the languages —which would mean that is trivially shared by all creoles, specifically— and, similarly, a high value of precision can be achieved if a complex rule that only fits one creole is proposed. In order to avoid such irrelevant cases, it is necessary to use a measure that combine both aspects of classification. A standard solution is the weighted harmonic mean of accuracy and recall,

$$F_\beta = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

which is called as *F-measure* [195]. The $\beta$ parameter determines the relative importance of recall over precision — when recall =

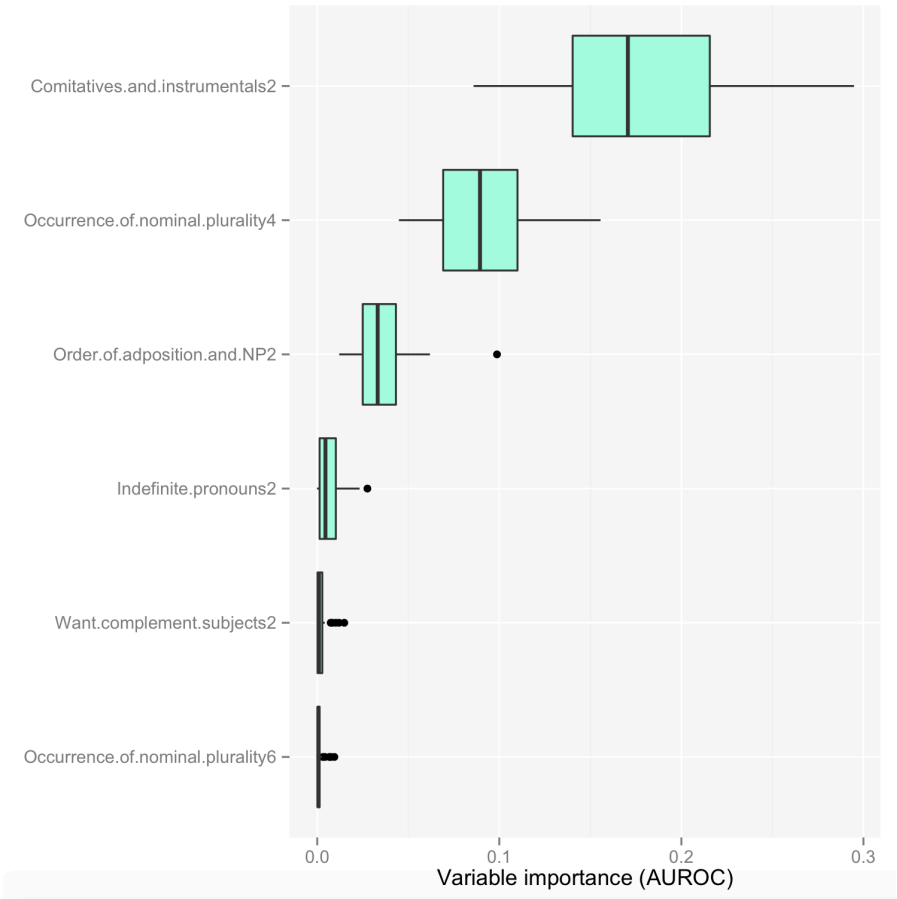


Figure 21: **Variable importance for random forests using features without genealogical association**

Boxplots of variable importance as inferred from a random forest based on features without genealogical association.

$\beta \cdot$ precision, $F_\beta$ = precision. Notice that the F-measure does not consider the true negatives — the proportion of languages that are not creoles, in this case. More and less non-creole languages in the sample will characteristically produce smaller and larger values of precision (and F-measure.) However, and beyond the fact that the two classes of languages are not extremely unbalanced in our case study, here we are interested in the quality of classification of creoles in relation to other comparable groupings — any positive or negative bias will become clear in the randomized sets of languages.

Hence, the goal is to mine the data in order to retrieve the best rules according to some measure of classification efficiency. From the point of view of computational complexity, a brute-force search is suboptimal but feasible anyhow. An efficient alternative is the classic breadth first algorithm *a priori* [4].

The structure of the rules inferred is simply given by

$$F_1 = f_1, F_2, \ldots, F_n \implies \text{class} = \text{creole} \tag{17}$$

where $n$, the number of features on the left-hand side of the rule, is referred as the *rule length*.

As with the probabilistic profile analysis, I compared the rules induced by the apriori algorithm in the empirical case against 1500 randomizations of the data's classes (50 for each of the 30 imputed datasets). For each of the real and randomized datasets and a given prototype length — the number of features that compose the prototype — I found the best candidate according to four information retrieval measures: recall, $F_1$, $F_2$ and $F_{0.5}$.

### 4.7.2 *Results*

In the empirical data, a number of strong candidates for creole prototype were found with the a priori algorithm. Results can be observed in Figure 22.

For all of the rule sizes and the classification measures, the distribution of all the best values for the empirical data achieves larger values than those of the randomized versions. It is interesting to note that, for length 1, the empirical values of recall are not much higher than that of the random groups — this is simply a reflection of the fact that some typological features are extremely frequent. All of the empirical values for all the measures and rule lengths are above 0.7. As expected, the values of recall and $F_2$ shrink as rule length increases, whereas the opposite occurs with $F_1$ and $F_{0.5}$.

Interestingly, similar results are found when the reduced set of features with no ancestry association is used, as it can be seen in Figure 23

## 4.8 MISCLASSIFICATION PATTERNS

So far we have shown that both implementations of the creole profile yield an efficient classification scheme that distinguishes creoles from non-creoles. However, both takes on the creole profile end up with a classification scheme that is not perfect. The analysis of the cases that are misclassified — namely, non-creole languages that are flagged as creoles — plays a crucial role in the discussion about what is precisely that our classification schemes are capturing.

We center our attention on those languages that were on average misclassified at least half of the time (for at least one rule length, in the case of the prototype profile, and overall for the probabilistic profile). The misclassified languages according to the prototype profile can be found in Table 4.8 (all features) and Table 4.8 (reduced set of features).

| Language | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Berbice Dutch | 0.00 | 0.11 | 0.56 | 0.84 |
| Diu Indo-Portuguese | 0.43 | 0.46 | 0.50 | 1.00 |
| Ghanaian Pidgin English | 0.12 | 0.39 | 0.60 | 0.87 |
| Kikongo-Kituba | 0.46 | 0.07 | 0.04 | 0.51 |
| Korlai | 0.75 | 1.00 | 1.00 | 1.00 |
| Kriol | 0.00 | 0.30 | 0.51 | 0.25 |
| Lingala | 0.56 | 0.14 | 0.40 | 0.68 |
| Papia Kristang | 0.10 | 0.08 | 0.36 | 0.89 |
| Sri Lanka Portuguese | 0.54 | 0.95 | 1.00 | 1.00 |
| Ternate Chabacano | 0.54 | 0.95 | 0.99 | 1.00 |
| Zamboanga Chabacano | 0.54 | 0.95 | 0.99 | 1.00 |
| Arabic (Egyptian) | 0.57 | 0.58 | 0.53 | 0.73 |
| Apurin | 0.57 | 0.46 | 0.00 | 0.00 |
| Bagirmi | 0.58 | 0.59 | 0.57 | 0.75 |
| English | 0.79 | 0.93 | 0.97 | 0.99 |
| Ewe | 0.72 | 0.62 | 0.02 | 0.00 |
| Finnish | 0.57 | 0.30 | 0.00 | 0.00 |
| Guaran | 0.57 | 0.12 | 0.00 | 0.00 |
| Hausa | 0.44 | 0.86 | 0.60 | 0.32 |
| Hmong Njua | 0.67 | 0.43 | 0.66 | 0.64 |
| Indonesian | 0.62 | 0.53 | 0.86 | 0.39 |
| Ju'hoan | 0.54 | 0.61 | 0.24 | 0.03 |
| Khmer | 0.65 | 0.54 | 0.51 | 0.39 |
| Khasi | 0.69 | 0.41 | 0.50 | 0.00 |
| Lango | 0.32 | 0.58 | 0.53 | 0.25 |
| Latvian | 0.79 | 0.70 | 0.57 | 0.01 |
| Mapudungun | 0.57 | 0.17 | 0.00 | 0.00 |
| Maung | 0.62 | 0.57 | 0.53 | 0.66 |
| Maybrat | 0.35 | 0.63 | 0.54 | 0.26 |
| Mandarin | 0.57 | 0.13 | 0.00 | 0.00 |
| Ngiyambaa | 0.53 | 0.07 | 0.00 | 0.00 |
| Nahuatl (Tetelcingo) | 0.50 | 0.57 | 0.51 | 0.47 |
| Russian | 0.67 | 0.55 | 0.50 | 0.00 |
| Spanish | 0.58 | 0.58 | 0.51 | 0.00 |
| Swahili | 0.32 | 0.45 | 0.51 | 0.25 |
| Taba | 0.42 | 0.54 | 0.66 | 0.25 |
| Thai | 0.61 | 0.40 | 0.50 | 0.64 |
| Tiwi | 0.33 | 0.59 | 0.50 | 0.25 |
| Vietnamese | 0.78 | 0.51 | 0.51 | 0.64 |
| Wich | 0.60 | 0.56 | 0.56 | 0.68 |
| Yoruba | 0.78 | 0.51 | 0.52 | 0.73 |

Table 17: Misclassification percentages for the average across all the best rules according recall, $F_1$, $F_{0.5}$ and $F_2$ for a given rule length using all features. Only languages misclassified more than 50% for at least one rule length are displayed.

| Language | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Ambon Malay | 0.00 | 0.27 | 0.89 | 0.95 |
| Angolar | 0.09 | 0.00 | 0.13 | 0.72 |
| Berbice Dutch | 0.35 | 1.00 | 1.00 | 1.00 |
| Fa d Ambo | 0.09 | 0.00 | 0.13 | 0.72 |
| Ghanaian Pidgin English | 0.75 | 1.00 | 1.00 | 1.00 |
| Gullah | 0.00 | 0.27 | 0.89 | 0.95 |
| Korlai | 0.91 | 1.00 | 1.00 | 1.00 |
| Kriol | 0.16 | 0.25 | 0.56 | 0.93 |
| Santome | 0.09 | 0.00 | 0.13 | 0.72 |
| Sri Lanka Portuguese | 0.91 | 1.00 | 1.00 | 1.00 |
| Arabic (Egyptian) | 0.60 | 0.52 | 0.58 | 0.62 |
| Arapesh (Mountain) | 0.54 | 0.35 | 0.24 | 0.23 |
| Bagirmi | 0.60 | 0.41 | 0.35 | 0.38 |
| Berber (Middle Atlas) | 0.60 | 0.44 | 0.40 | 0.43 |
| Barasano | 0.56 | 0.00 | 0.00 | 0.00 |
| Chamorro | 0.64 | 0.58 | 0.63 | 0.65 |
| English | 1.00 | 1.00 | 1.00 | 1.00 |
| Ewe | 0.56 | 0.00 | 0.00 | 0.00 |
| Greek (Modern) | 1.00 | 0.88 | 0.57 | 0.56 |
| Hausa | 1.00 | 1.00 | 1.00 | 0.99 |
| Hmong Njua | 0.57 | 0.47 | 0.34 | 0.03 |
| Indonesian | 0.51 | 0.25 | 0.00 | 0.00 |
| Irish | 1.00 | 0.91 | 0.67 | 0.66 |
| Iraqw | 0.51 | 0.52 | 0.51 | 0.00 |
| Jakaltek | 0.60 | 0.25 | 0.00 | 0.00 |
| Ju'hoan | 0.56 | 0.00 | 0.00 | 0.00 |
| Ket | 0.51 | 0.00 | 0.00 | 0.00 |
| Khmer | 0.58 | 0.25 | 0.00 | 0.00 |
| Khasi | 0.51 | 0.42 | 0.34 | 0.00 |
| Lango | 0.60 | 0.52 | 0.58 | 0.62 |
| Latvian | 0.91 | 1.00 | 0.87 | 0.28 |
| Malagasy | 0.68 | 0.62 | 0.67 | 0.68 |
| Maori | 0.51 | 0.44 | 0.36 | 0.00 |
| Maung | 0.63 | 0.57 | 0.62 | 0.64 |
| Maybrat | 0.68 | 0.36 | 0.00 | 0.00 |
| Mixtec (Chalcatongo) | 0.62 | 0.57 | 0.62 | 0.64 |
| Mangarrayi | 0.60 | 0.52 | 0.58 | 0.62 |
| Nahuatl (Tetelcingo) | 0.64 | 0.56 | 0.62 | 0.66 |
| Persian | 0.91 | 0.73 | 0.05 | 0.00 |
| Russian | 0.60 | 0.52 | 0.58 | 0.62 |
| Spanish | 0.71 | 0.54 | 0.42 | 0.43 |
| Swahili | 0.51 | 0.52 | 0.51 | 0.00 |
| Taba | 0.61 | 0.26 | 0.00 | 0.00 |
| Thai | 0.51 | 0.25 | 0.00 | 0.00 |
| Tiwi | 0.69 | 0.64 | 0.68 | 0.70 |
| Vietnamese | 0.60 | 0.35 | 0.00 | 0.00 |
| Wari' | 0.56 | 0.58 | 0.53 | 0.02 |
| Wich | 0.60 | 0.40 | 0.34 | 0.36 |
| Yoruba | 0.51 | 0.52 | 0.51 | 0.00 |

Table 18: Misclassification percentages for the average across all the best rules according recall, $F_1$, $F_{0.5}$ and $F_2$ for a given rule length using the restricted set of features. Only languages misclassified more than 50% for at least one rule length are displayed.

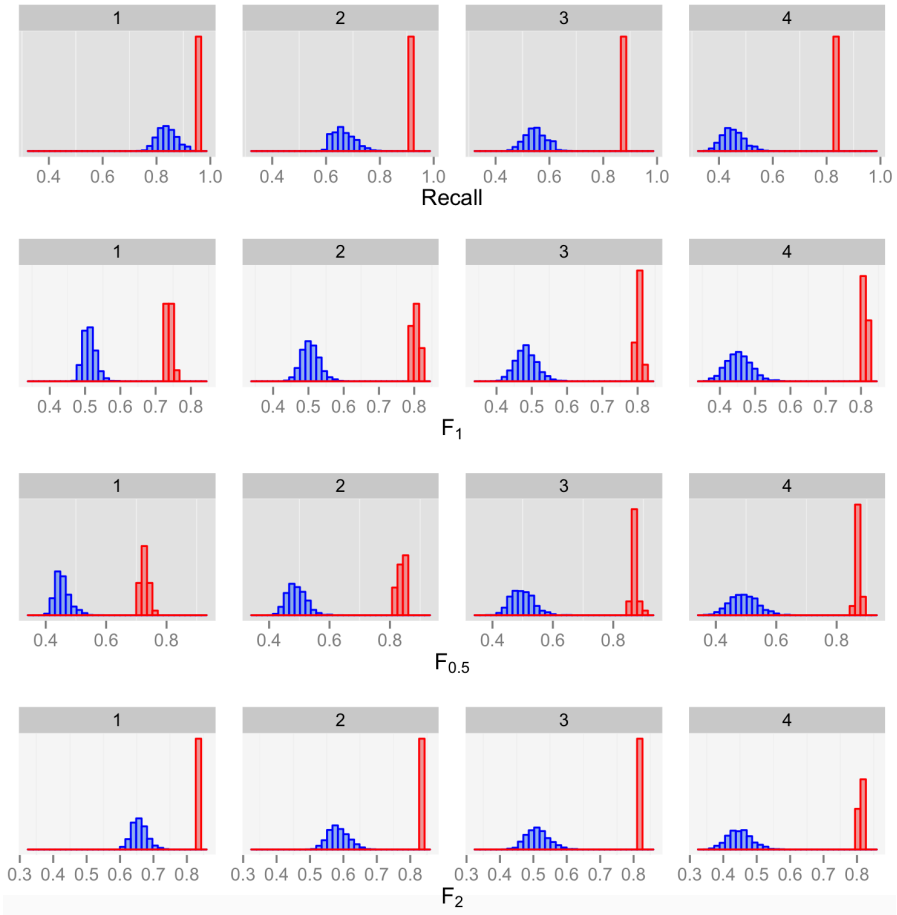


Figure 22: **Distribution of best rules for all features**

Distribution of the classification properties for the best rules in the empirical (red) and randomized (blue) datasets.

Two patterns arise. First, while the vast majority of languages belong to the European/Macro-Sudan biclan, they are actually underrepresented in the misclassified languages. Second, among the non-creole languages flagged incorrectly as creoles we find either the usual European or Macro-Sudan ancestors or languages typologically close to them.

It would be possible to argue that these results are a simple byproduct of the relative lack of versatility of the rule-based approach, and that the probabilistic approach should be exempt of this clear ancestral bias. However, the results are of a similar nature for that model as well, as it can be seen in 4.8 and 4.8.

These results show a clear pattern: the classifiers are picking up a group that can be better described as a combination of most creoles and the Indo-European and Sub-Saharan African languages, in contrast to most other non-creole languages as well as creoles with uncommon lexifiers or substrates.

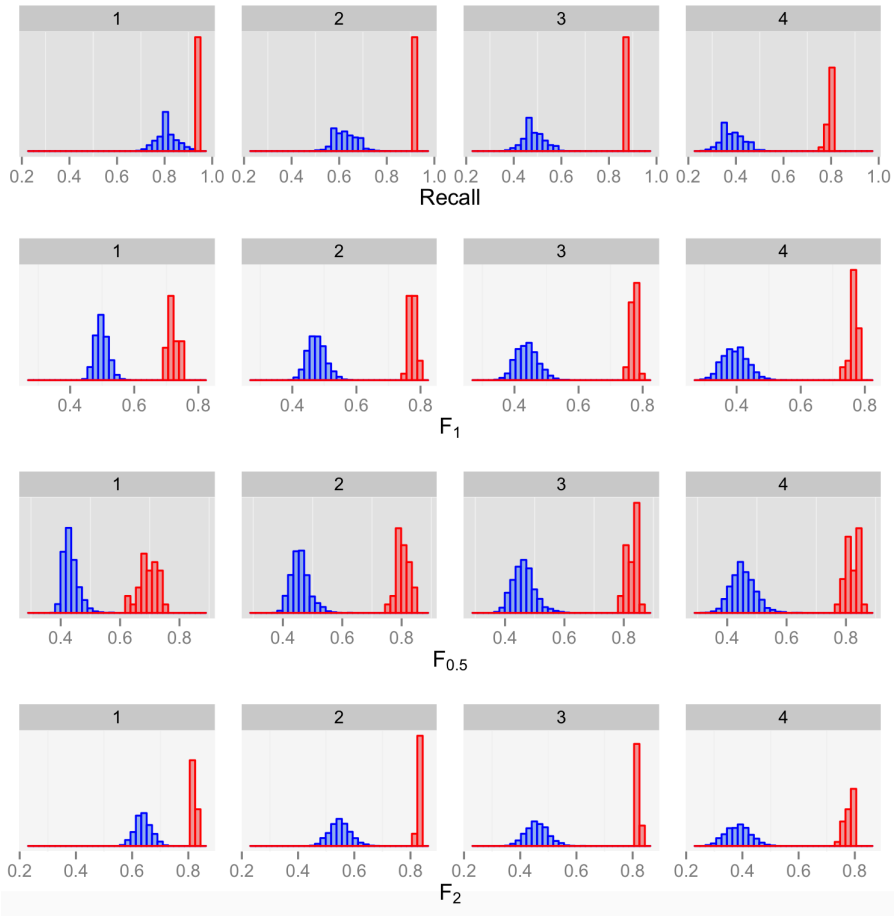


Figure 23: **Distribution of best rules for independent features**

Distribution of the classification properties for the best rules in the empirical (red) and randomized (blue) datasets.

| Language | Frequency |
|---|---|
| Korlai | 1.00 |
| Sri Lanka Portuguese | 1.00 |
| Ternate Chabacano | 0.73 |
| Zamboanga Chabacano | 1.00 |
| English | 1.00 |
| Ewe | 0.97 |
| Hausa | 1.00 |
| Indonesian | 1.00 |
| Ju—'hoan | 0.97 |
| Lango | 1.00 |
| Latvian | 0.97 |
| Russian | 1.00 |

Table 19: Misclassification percentages for random forest using all features. Only languages misclassified more than 50% for at least one rule length are displayed.

| Language | Frequency |
|---|---|
| Berbice Dutch | 0.87 |
| Ghanaian Pidgin English | 0.87 |
| Korlai | 1.00 |
| Sri Lanka Portuguese | 1.00 |
| Zamboanga Chabacano | 1.00 |
| English | 1.00 |
| Greek (Modern) | 1.00 |
| Hausa | 1.00 |
| Indonesian | 1.00 |
| Irish | 1.00 |
| Iraqw | 1.00 |
| Lango | 1.00 |
| Latvian | 1.00 |
| Persian | 1.00 |
| Russian | 1.00 |

Table 20: Misclassification percentages for random forest using the reduced set of features. Only languages misclassified more than 50% for at least one rule length are displayed.

## 4.9 CONCLUSIONS

I have shown that creoles can be distinguished efficiently from non-creole languages on a typological basis, without the need of pre-selecting which variables would be relevant for that purpose. This is true for both a fixed-rule scheme as well as a probabilistic classifier. Surprisingly, their classification efficiency becomes only slightly impoverished when only features for which no ancestral association can be found are used.

However, the misclassification patterns reveal that, in spite of the previous pruning of the ancestry-related features, the structures that are being picked up are of a clearly genealogical nature, in particular related to the overwhelming contributors from Europe and Sub-Saharan Africa.

Knowing the rough ancestry of a creole does not perfectly predict the value of any particular feature, possibly due to the fact that the ancestry groups might fail to capture variation structured by the different sources creoles have at their disposition (due to limitations discussed before), or because of regular language change, which renders languages less similar to their ancestry through time. In principle, some of this variation unexplained by the ancestry in the creoles features might well be due to genuine innovation of grammatical structures related to creolization, and more generally our results do not preclude the existence of truly creole features (perhaps in domains that are not well covered by our data, like morphophonology) but they prove that the overwhelming majority of creole grammars are transmitted as in any other natural language, either by genealogy or contact. Crucially, our results bring into question the need of a transmission bottleneck and a pidgin phase in the history of the development of creoles to explain commonalities across creole languages. If, as we have shown, a substantial number of features are passed along from the ancestry to the creoles, then positing an intermediate pidgin stage that would have considerably reduced and simplified the ancestry features does not seem to be plausible, since it would remain to be explained why creoles faithfully continue word orders, copula patterns, ditransitive constructions, subject relative clauses, and indefinite pronoun patterns — just to mention some grammatical patterns — which are continued from their ancestral languages.

It should be stressed that these results are consistent with both genealogical and contact-related transmission. Some of the creoles have coexisted with some of their ancestral languages, which might have resulted in the ancestry-dependent features being a later development in the history of those languages.

Why such a complex human behaviour can be successfully transmitted even in the typical (intricate and multilingual) contact situations of creoles is still unclear. Whatever the underlying reason, put

together, our results speak about the astonishing resilience of language transmission.

# 5

# ECOLOGICAL PRESSURES ON SPEECH SOUNDS

## 5.1 HUMAN BEHAVIOUR IS FLEXIBLE

While non-human apes are restricted to a rather narrow set of ecological and geographical circumstances, the Sapiens species — and some the related species of the Genus Homo — have spread over the range of all other vertebrates (which accounts to most of the Earth's land) [211]. The efficient cause behind this evolutionary success story is, undoubtedly, the versatility of human behaviour, a true testimony of adaptiveness. Diet, sleep habits, age of reproduction, parental investment and basic societal organization vary dramatically from the *trepang* harvesters in the Philippines —who spend most of their lives in a boat on the sea— to the Andes cultivators of Maize and the professors of the University of Leipzig. Some of these dimensions of variation have compelling (or at least, suggestive) explanations in terms of evolutionary biology. To take one example among many, the average number of a woman's childbirth in a population is tightly correlated to chance of child death, which in its turn it is a good predictor of the overall investment of the parents on their offspring [175]. These concepts and ideas are (not innocently) the same that are used to explain the differences in other species' behaviour, from penguins to *E. Colli*.

Even more, aspects of human life that appear to be beyond the mundane influence of immediate or mediate bodily needs have been linked (with differing degrees of success) to the same causing factors. Religiosity, xenophobia, ethnocentrism and in-group solidarity have been argued to emerge with more intensity in regions and populations that are subject to large pathogen-stress [76]. Belief in moralizing high gods is stronger in populations that inhabit ecologically labile regions, which suggests that (at least some kinds of) religiosity might appear as adaptive strategies to cope with a changing and unpredictable context [29].

These changes in behaviour are sometimes paralleled, reinforced or made unnecessary by modifications in the biological basis of human populations. High-altitude settings pose the serious threat of hypoxia, but a special set of genes (that maximize oxygen-intake) have evolved independently in Tibet, the Ethyopian plateau and the Andes

[163]. Skin color is strongly associated with amount of exposure to UV radiation: darker tones resist better solar radiation, while fairer ones are better for the absorption of calcium in light-poor regions [115]. Populations that have practiced pastoralism for a long period of time have gained, in many cases, an appropriate set of genes that break down lactase efficiently even in adulthood [221].

Given this brief summary of how human behaviour and (in a lesser extent) human biology are connected to the surrounding pressures imposed by the environment, the question is: what happens with language, the human behaviour *par excellence*? Sadly — and leaving aside romantic and inaccurate accounts about the number of words for snow that Inuktitut has and the like — there has been almost no mainstream *linguistic* research in this direction in the last hundred years and until very recently.

If there is any suspect among the aspects of human language to exhibit the behavioural (and maybe biological) variation discussed before, that is the capacity to produce sounds. While presumably most of the operations concerning morphosyntaxis, semantics or phonology involve arrangements of firing neurons in the brain, the production and (the initial stages of) perception of speech depend on hardware that has been used for communication for at least the last half a million years[1]

## 5.2 THE WORLD-WIDE DISTRIBUTION OF SPEECH SOUNDS

The first step is, naturally, to describe the range of variation of the speech sounds that characterize the languages of the world. More precisely, we will study the distribution of *phones*: human speech sounds described in basic phonetic dimensions, like manner and place or articulation and voice.

The data for this cross-linguistic analysis come from the 2013 version of the PHOIBLE [165]. PHOIBLE gathers published descriptions of languages' phone inventories; the version used here covers about 1200 languages divided among most linguistic areas.

As with all the previously described linguistic databases, there are many factors that introduce noise in the data. Sometimes similar phones are collapsed under a similar label or are not described at all because they do not imply phonemic distinctions in a language, or the size of the corpus the description was based on was insufficient to capture the occurrence of low-frequency phones. Because of this, classes of phones – instead of individual phones – have been the usual

1 This is a thorny subject with which I do not want to engage here, but estimates on the origin of modern language range between 50k to 2M ybp [22, 52]. This is not necessarily informative about when speech became functional for the purpose of communication, though.

object of study for cross-linguistic comparison, and we will follow this direction here as well.

### 5.2.1 *Results*

Raw number of different phones is a classic measure that is often associated to the slippery concept of linguistic complexity (which we encountered in chapter 4). Some early proposals suggested that the distribution of this quantity is more or less normal, thus implying that large-scale variation in this dimension lacks an interesting structure and is probably due to neutral historical processes [178]. A log-normal model was predicted as well on the basis of a simple multiplicative model [125]: if languages make a productive use of their phonetic features, then — under the questionable approximation of features as combinatorial units, which is empirically inadequate — the overall distribution of number of phones could be approximated by the product $S = \Pi_i F_i$, where $F_i$ is a random variable that takes value 1 if the feature $i$ is not part of the language repertoire or any other positive number equivalent to the number of distinctions allowed by the feature. This leads to a log-normal distribution for the random variable $S$.

The empirical distribution of $S$ can be observed in Figure 24. Unsurprisingly, $S$ is right-skewed (as many other count data distributions) with a skewness of 1.75. The language with the largest phone inventory is the Khoe-San !Xóõ with 161 units and the smallest are Pirahã (a Mura isolate) and Rotokas (West Bougainville) with 11. While a more detailed parametric evaluation of the distribution could be done, the fact that $S$ is a coarse pooled empirical distribution that collapses historical and areal information makes any further formal exploration futile.

Most of what we have pointed out for segment distributions is also true for the most important partitions of phones —consonants and vowels— when taken separately. A more important relation among these two categories concerns the hypothesis of a linguistic trade-off. Following the idea that the overall number of phones is a faithful proxy for complexity, and that languages are expected to be roughly comparable in that dimension in order to achieve their main function, it has been conjectured that an increase in complexity in a particular subsystem would lead to a decrease of complexity in another [151, 191] — one often-cited instance of such phenomenon being, for instance, the relative word order rigidity in languages with little morphology.

Maddieson [146, 147] investigates correlational patterns between different phonological subsystems in an attempt to shed light on compensatory relations in phonological system, noting that regardless of whether the driving force is historical or communicative (or whether

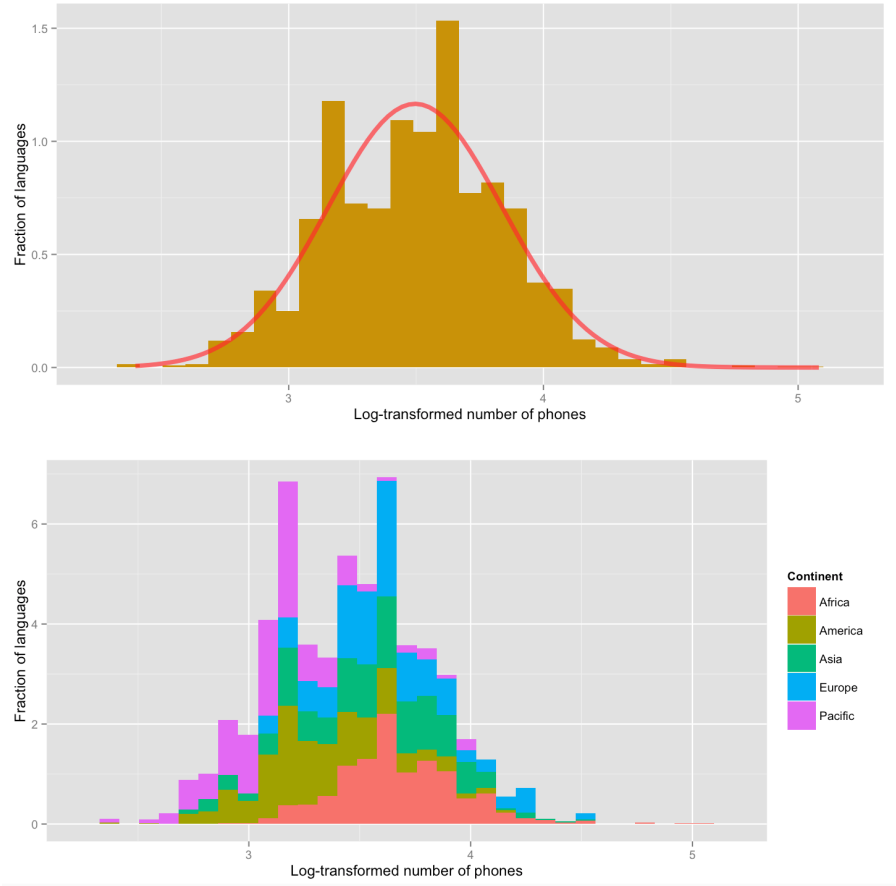


Figure 24: **Distribution of best rules for independent features**

Aggregated distribution of phone inventory size. Log-normal approximation (upper panel) and divided by continent (lower panel.)

if it exists at all), compensatory dependencies should emerge in a survey of phonological properties of the world's languages. Maddieson and Disner [149] examines suprasegmentals (tone and stress) in a set of 317 languages and reports that the "overall tendency appears once again to be more that complexity of different kinds goes hand in hand, rather than for complexity of one sort to be balanced by simplicity elsewhere."

As a proof of concept, we test this with a simple mixed-effect model with WALS genera as random intercepts and slopes and scaled variables,

$$\text{consonants} = (\beta + \beta^{\text{genera}})\text{vowels} + \alpha + \alpha^{\text{genera}}$$

where $\beta$ and $\alpha$ are the fixed effects (slope and intercept respectively) and $\beta^{\text{genera}}$ and $\alpha^{\text{genera}}$ the corresponding random slopes and intercepts. This model suggests that number of vowels is a significant predictor of the number of consonants ($\beta = 0.15, P < 10^{-3}$), although the overall quality of the model does a poor job at explaining the vari-

ation: the marginal $R^2$ (the proportion of variance accounted for the inclusion of the fixed effects) is less than 0.02, whereas the conditional $R^2$ (with random effects included) is 0.56. Once again, the aggregated effect disappear when genealogical dependencies are taking into account.

## 5.3 THE ACOUSTIC ADAPTATION HYPOTHESIS

Within anthropology, some researchers have explicitly endorsed the idea that ecology affects the composition of phonic systems through modifying the conditions of vocal communication.

Fought et al. [79] and Munroe et al. [171] find that cold climate significantly predicts lower average proportions of CV syllables in words, and they submit that the reason is that people in warmer climates spend more time outdoors and therefore communicate at greater distances compared to people in colder climates. Aiming to take into account Galton's problem, a further study shows that this effect also holds within four language families [170]. Distal communication might also explain why languages spoken in warm areas tend to have more sonorous sounds (i.e. sounds that carry more energy). This idea is reinforced by a follow-up study by Ember and Ember [67], where it is found that additional hindering factors, including density of plant cover and mountainous terrain, predict total-sonority scores[2].

This idea has a long pedigree in animal communication studies, under the form of the acoustic adaptation hypothesis (AAH,[166, 199, 241]). The hypothesis states that animal communication systems are adapted to the climatic/ecological environment in which they operate, optimized for the transmission characteristics of the environment or other factors like phenotypic plasticity (e.g. Ziegler et al. [245]). Evidence for associations between the acoustic attributes of a habitat and vocalizations have been found in birds, prairie dogs and macaque monkeys [174].

Besides the obvious differences between the anthropological and the animal communication literature, the later is considerably more explicit when comes to explain why systems change in the direction of a better fit between acoustic signal and environment: typical Darwinian dynamics. While it seems straightforward to assume a scenario where more efficient signals improve mating or survival chances in non-human animals, how exactly languages develop more sonorous repertoires in the presence of communication obstacles is left undiscussed.

The pioneer study of Maddieson (summarized in Maddieson and Coupé [148]) makes an explicit link from the work of Fought et al. [79] and Munroe et al. [171] to the AAH. Maddieson uses a sample

2 And, we should add, other predictors outside of the explanatory scope of the acoustic adaptation hypothesis, like the amount and prevalence of extra-marital sex [68].

of about 450 languages and shows a correlation between the sum of consonant inventory size and a measure of syllabic complexity — both are normalized to give them equal weight and the result is called *phonological index* — and absolute latitude (a point demarcating a central location for the speaker population). This is used to vindicate the account laid out before: temperate environments with open vegetation facilitate transmission of higher frequency signals more than warmer and more densely vegetated environments, which entails that languages spoken close to the Equator tend to use more sonorous sounds (and specifically, more vowels) than those far from it.

### 5.3.1 *A stringer test of the AAH*

In the absence of a precise mechanism explaining how the AAH is expressed in language, the correlations discussed in the previous section are nothing but a first suggestive step forward.

Maddieson and Coupé [148] overcomes one of the most obvious limitations of previous studies, which was the modest size of the data used. Using either LAPSYD, [150] or the already introduced PHOIBLE should be, at least when dealing with coarse descriptions of inventories, the minimum standard.

The usage of latitude as a proxy for the actual ecological variables is not justified given the availability of high-resolution satellite information. Ultimately, we would like to test that a specific ecological factor or combination of factors is predictive of certain characteristics of the repertoires even after accounting for otherwise circumstantial correlates (like latitude). However, as we have pointed out before, which variable *precisely* triggers or explains the AAH in language is undetermined.

There is a last requirement that the testing of the AAH needs to satisfy: it should be able to outperform other competing or overlapping explanations. Among the theories that have been put forward to explain why the distribution of language sounds is the way it is by appealing to exogenous factors, two have gained notoriety and are regularly discussed: the population link (via absolute population and contact) and the Out-of-Africa scenario. We have briefly discussed the later: in analogy with population genetics, languages can be approximated as phone pools that are amenable of being subject to bottlenecks due to migration, thus leaving a trace of decreasing number of phones according to (migration) distance with respect to the putative origin of modern human languages, in Africa [7, 44]. The population link suggests that the number of people speaking a language in a given time has an impact on the rate the language will gain or lose phonemic distinctions [7, 107], although there has been work showing that the available evidence is weak and mostly an artifact of Galton's

Problem [60, 164]. In addition, the number of languages with which a language is in contact might impact its own repertoire (through loanwords, for instance.)

In certain manner, these three predictors (distance from Africa, population size and density of neighbors) are comparable to the AAH in that they are not fully-fledged testable mechanisms and that they strongly rely on the statistical fit of the data. If the AAH does not yield a better model than these other alternatives, then we can safely forget about deciding on its feasibility based on this type of data.

### 5.3.2 *Results*

In the absence of a single and well-defined ecological variable to test, we first summarize the variation in the ecological variables across the languages of the world via a simple principal component analysis. The variables used for this purpose encompass all of the factors that have been previously highlighted: annual mean temperature, tree coverage and precipitation. The first component of the decomposition (that explains 63% of the variance) gets positive projections of 0.62, 0.66 and 0.41 for tree coverage, precipitation and temperature, respectively. The behaviour of this component (which we will refer for short as "environment") can be linked to a coherent unique prediction: a unit increase in this dimension would lead to a reduction in the number of non-sonorous sounds.

As for the segment classes, we divide the individual elements of both PHOIBLE and LAPSYD into three encompassing categories: vowels, sonorants and obstruents (in order of decreasing sonority). The reason why we collapse several segments within these broad bins is twofold. First, determining relative sonority between individual segments that belong to these categories is not always straightforward, and second, the presence of most individual segments is likely to be tied to specific linguistic areas or lineages. Using these segment classes permits to predict a particular cline of the expected effect of the acoustic adaptation hypothesis: obstruents $>$ sonorants $>$ vowels.

In order to test this, we propose a generalized mixed effects model with a Poisson link

$$\begin{aligned}\log(\delta) \quad = \quad & \beta + \beta^{\text{Environment}} z_{\text{Environment}} + \beta^{\text{Dist. from Africa}} z_{\text{Dist. from Africa}} \\ & + \beta^{\text{Log. population}} z_{\text{Log. population}} + \beta^{\text{Contact int.}} z_{\text{Log. Contact int.}} \\ & + \alpha_{\text{Family}} + \alpha_{\text{Area}}\end{aligned}$$

where $\delta$ is the mean number of segments of a given class. Coefficients $\beta_{\cdot}$ are composed by a fixed effect plus a random slope depending on linguistic family. $\alpha_{\cdot}$ are random intercepts according to the relevant subscripted variable. All predictor variables have been

z-transformed for the sake of model convergence and coefficient comparability.

Segment data come from a combined dataset of PHOIBLE and LAPSYD consisting of a balanced set of 1400 languages. Logged population size and geographical distribution of languages was taken from Ethnologue [138]. The number of linguistic neighbors is measured in a circle centered in a given language with radius of 1000 km, and the out-of-Africa distance is measured in a similar fashion to Atkinson [7] by considering waypoints between continents. Parametric bootstrapping was employed on order to obtain confidence intervals for the main effects. Results can be observed in Figure 25.

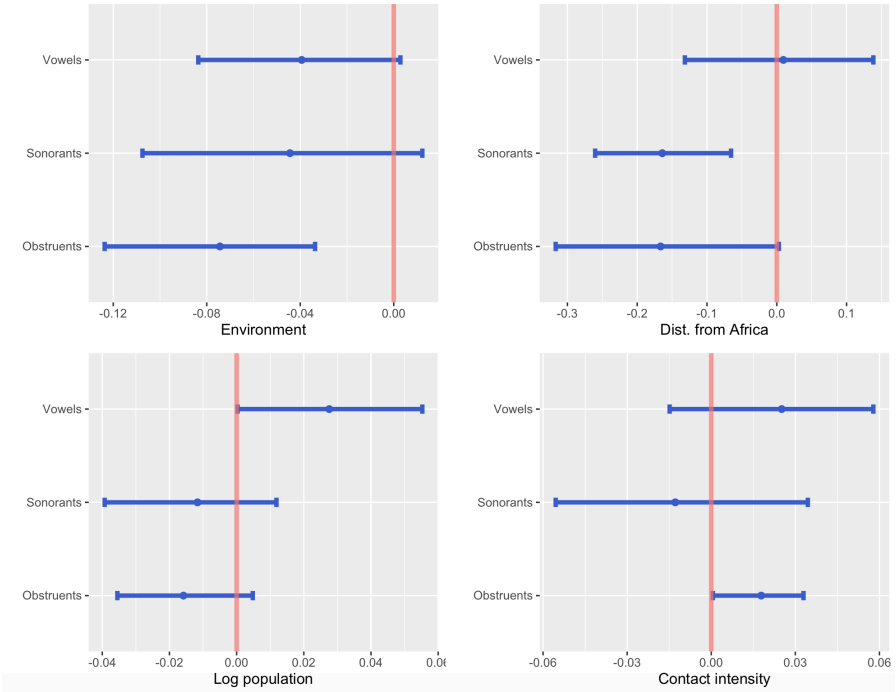


Figure 25: **Evaluation of the Acoustic Adaptation Hypothesis**

Each panel corresponds to the estimated coefficients (blue dots) and bootstrapped 95% confidence intervals for the segment classes (vowels, sonorants and obstruents) for each of the four predictors (environment, distance from Africa, log-transformed population and contact intensity).

For starters, the role of logged population and contact intensity can be disregarded as not relevant (since the confidence intervals include or are adjacent to zero for all categories.) While no effect of environment can be observed in vowels and sonorants, there is a clear effect on the obstruents ($\exp(-0.07) \approx 0.93$, which means that increasing 1 SD the environment variable leads to a decrease of about 7% in the number of obstruents). This is consistent with the cline proposed before: if there is any effect, it should be more visible in obstruents (and

then sonorants and finally vowels.) Interestingly, the out-of-Africa effect emerges as well: 1 SD in that direction leads to a decrease of about 14% of the number of sonorants a language has.

## 5.4 TONES AND HUMIDITY

In the previous section we considered the importance of the acoustic properties of the environment for the change and development of phonic systems, and we ended up with inconclusive results. Here we take a different perspective with a much more robust grounding, by considering whether ecology can impact in any meaningful way the anatomical basis of speech.

### 5.4.1 *Vocal folds hydration*

The larynx is the main organ responsible for the main frequencies of speech sounds of pulmonic nature [149]. It can be roughly described as a muscle and cartilage tube that connects the trachea with the bucal cavity. More or less in the middle of its length there are two flap-like pieces of tissue perpendicular to the main axis of the larynx. The space between these *vocal folds*, the *glottis*, can be open or closed at will, thus changing the aerodynamic properties of the air that emerges from the lungs. The most pervasive effect of vocal fold vibration is to confer *voice* to consonants, a buzz-like quality responsible for the differences between /b/ and /p/ or /g/ and /k/ in English. Other configurations of the glottis and the vocal folds are responsible for some other consonantal classes, like glottal stops — achieved through phonation while the glottis is closed.

Critically for our study, while the rate of vibration of the vocal folds is beyond the minimum threshold of human audition — 1 versus 20 Hz — it does impact on the perception of *pitch*. Pitch is relevant for prosody and intonation, which are frequently used to convey specific pragmatic meanings (e.g. between affirmative and declarative sentences in English). In those cases, pitch changes might occur over a span of several words or even phrases. More restrictively, differences in pitch or pitch countours within individual phones — mostly vowels — is used in some languages with the purpose of distinguishing lexical meanings. For instance, the difference between the Thai words *mâk* (ADV often, frequently) and *màk* (V ferment) is that the vowel of the first word involves rising pitch whereas the second has a decreasing pitch. When this is the case, these phonemic differences in pitch receive the name of *tones*.

Experimental evidence — mainly based on bovine and canine excissed larynxes — shows that vocal fold vibration decreases with dehydration. Considerable human clinical reports — of populations ranging from karaoke singers to school teachers — coincide on that

dehydration increases the PTP (phonation threshold pressure) and PPE (perceived phonation effort). These effects are summarized in Everett et al. [73].

There are two noticeable effects in phonation that result from dessicated air inhalation: jitter (imprecise pitch) and shimmer (varying amplitude). Males have a pitch range of about 100 Hz [130], and the minimal *phonemic* frequency distance between two tones is around 20-30 Hz [141]. It has been suggested that the limit on the number of level tonemes a language can have is five [149], but even in those cases secondary phonetic laryngeal features serve as cues to the tonemes [130].

Other things being equal, the most important factor determining hydration is the temperature of aspired air, which is directly dependent on climate and geography. On the evolutionary time scale, a sustained exposure to particular climatic conditions might have driven some of the phenotipic differences in nasal features observable in the world today — dry air in contact with the mucosal tissue gets warmer and more humid, so different volume/surface ratios might be the result of strong environmental pressures [183].

This converging evidence leads to the prediction that the development of languages with many tonemic distinctions will be hindered or at least not favoured in languages spoken in dry and/or cold regions.

It is timely to note that this is not the first proposal relating hydration and phonic systems. Everett [72] focused his attention on *ejectives*, relatively rare non-pulmonic consonants that are produced by releasing air from the bucal cavity in a sudden manner. Everett noticed that those consonants are found in a few regions of the world: the Ethiopian highlands, the Andes, the Caucasus, Tibet and the North American Cordillera, all places of considerable high altitude. He suggested that the reliance on ejectives, being non-pulmonic phones, helps preventing water vapor loss, which is more accentuated in high altitudes.

### 5.4.2 *Results*

To test this hypothesis we make use of the Phonotactics Database of the Australian National University (ANU), the largest collection of phonological and phonetic inventories to date [61].

Figure 26 is a map representing the distribution of languages with and without systems of complex tone, for the larger ANU database (3,756 languages). As we see in the figure, about 20% of the languages have complex tone (n = 629)

The ecological variables we chose to evaluate that more directly relate to our hypothesis are the mean humidity (MH) and mean average temperature (MAT).

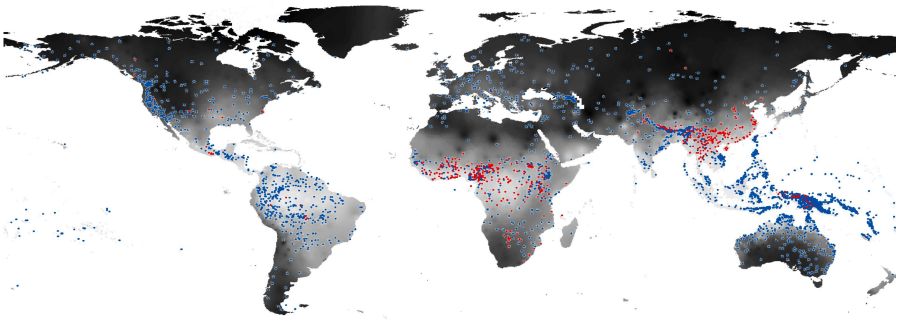

Figure 26: **Distribution of languages with and without complex tonal systems.**

Distribution of the languages in the Phonotactics Database of the Australian National University (ANU) [61]. Blue and red dots correspond to languages without and with complex tonal systems. The color of landmass maps on humidity: the whiter, the more humid.

Beyond the typical areal and genealogical correlations that we have encountered in the previous chapters, our testing scheme needs to account for the fact that most of the populations and languages of the world are circumscribed to the Equatorial belt. For most typological features, the majority of languages having a particular property will be trivially found in warm and humid places due to this demographic bias.

Second, it is important to remark that the claim is silent with respect to the development of complex tonal systems in humid and warm places — more precisely, we do not predict any trade-off between climate and the presence of tonemes beyond certain minimal ecological conditions. Ideally, we would like to determine such threshold from experimental data, but such an endeavour is well beyond the reach of the present study.

In order to address these concerns, we evaluated the hypothesis in the light of a resampling scheme. Summarily, if the most humid regions in which a sample of complex tonal languages can be found are systematically more humid than what it would be found for regions of languages with no tones or with a simple tonal system, then we can be confident that the hypothesis lives up to its most direct observational consequences.

More specifically, and given the fact that languages without complex tone outnumber languages with complex tone in a 6:1 relation, we sampled exactly the same number of languages coming from both groups in each iteration. In each run, different lower quantiles of humidity were computed on each set of languages, which were sampled in a genealogically balanced way (i.e., each family or isolate is represented by a single randomly chosen member in each set and each run).

The distributions of the languages with and without complex tone clearly differ. The 15th, 25th, 50th, and 75th MH percentiles of balanced samples of languages with complex tone have higher MHs in 89%, 88%, 43%, and 49% of the sample cases, respectively, when contrasted to the simulations' balanced samples of the same percentiles for the remaining languages. This is exactly in line with our predictions, as languages with complex tone overwhelmingly have higher MHs in the lower percentile ranges, suggesting clearly that such languages are extremely infrequent in very arid contexts, regardless of temperature. The 15th, 25th, 50th, and 75th MAT percentiles of balanced samples of languages with complex tone have higher MATs in 93%, 77%, 17%, and 19% of the sample cases, respectively, when contrasted to the equivalent simulations balanced samples of remaining languages. These results also pattern neatly in the direction of our prediction, as languages with complex tone are clearly particularly avoided in very cold regions but also in very hot, arid regions.

The distribution of the seven language isolates with complex tone is also restricted to warm and humid regions: Amazonia (Ticuna, Pirahã), sub-Saharan Africa (Laal, Banggime) and New Guinea (Damal, Lepki, Morori). In contrast, the 108 isolates without tone are spoken throughout the Americas, Eurasia, Africa, and Australia. A number are spoken in high latitudes in North America, South America, and Eurasia, as well as in numerous other arid regions. For both MH and MAT, there are clear disparities across isolates with and without complex tone, respectively. The average MH for isolates with complex tone is 0.017, whereas the average for other isolates is 0.013. This crossgroup disparity is significant (P = 0.02, MannWhitney). Similarly, the average MAT for isolates with complex tone, 23.7 C, is greater than the average for the remainder of isolates, 19.1 C, but not significantly so (P = 0.07, MannWhitney). These disparities align with the predictions of our account, despite the small number of isolates with much tonality.

## 5.5 CONCLUSIONS

In this chapter I have studied the overall distribution of human speech inventories and their relation to potential exogenous factors (like environment and demography).

The distribution of different segments is quite complex and cannot be captured satisfactorily through a parametric model, and the relation between segment classes is too weak as to shed light on potential mechanisms of segment gain or loss. This fact has an important consequence for the study of speech sounds as being behaviourally adaptive: if the observed distributions could be described by a simple parametric model, this could suggest that a simple process (like multiplicative growth or Brownian motion) is enough to understand

the variation in the number and quality of speech sounds used by the languages of the world.

Segment inventory size was tested against four proposed predictors in the literature: population size, contact, out-of-Africa distance and a number of ecological variables that hinder the propagation of soundwaves (like tree coverage and rainfall), summarized under a single dimension. I found no evidence for demographic factors, but the prediction born out by the acoustic adaptation hypothesis — less sonorous sounds will be less preferred in environmental contexts of strong soundwave degradation — is corroborated through a noticeable effect of environment on the number of obstruents. Puzzlingly, our models fail to reject the out-of-Africa effect proposed by Atkinson [7], although it is found to be restricted to sonorants only. Why this is the case is unclear at this point.

Finally, after a lengthy literature review on the physiology of the larynx and its role in speech production, I investigated whether the distribution of complex tonal languages could be restricted to areas of particular conditions of humidity, which turned out to be true.

All in all, these observational studies point out to the notion that *at least some aspects of speech* seem to covary with the environment, either indirectly (as in the case of the acoustic adaptation hypothesis) or directly (as in the distribution of complex tonal languages.)

There is one remaining important criticism that needs to be addressed in relation to this conclusion. Phonological systems have a strong areal flavour, which is most patent in the tones' study. From the perspective offered here, it must be asked why linguistic tone is so pervasively transferred across languages in some regions and why those regions tend to have arid borders. The account entertained here offers a natural explanation for this tendency and directly explains why the regions in which interlinguistic contact has led to pervasive use of tone —for example, sub-Saharan Africa and Southeast Asia— are warm humid regions. Put differently, it seems likely that tone spreads across languages more effectively via interlinguistic contact in regions with favorable ambient conditions, and very cold/dry regions apparently serve as barriers to the spread of (complex) tone. Both external and internal diachronic processes may be impacted to some degree by such ambient conditions, in ways that, at the least, merit serious consideration.

Particularly in light of the ubiquity of speech in the human experience, human sound systems are likely to be susceptible to ecological pressures. In cultures in which speech volume has been assessed quantitatively, it has been found that humans produce on average 15,000+ words per day [158]. So, if speakers rely on a sound pattern that is maladaptive (even in minor ways) in particular conditions, they do so ubiquitously. The analyses presented here suggest that

such an effect could have played a role in shaping the distribution of human speech sounds.

# Part IV

# GENERAL CONCLUSION

# 6

## A NEUTRAL APPROACH TO LANGUAGE

In this thesis I have approached a number of relevant issues on the nature of language by means of observational data and a diverse set of statistical methods. The individual conclusions of each of the four studies can be found in the respective chapters, so I am left now with the assessment of my initial goal for this piece of work and whether any general conclusions can be inferred.

I shall start with the obvious: an enterprise like the one I undertook is feasible. While more data and better methods are always welcomed, in all (or almost all) cases the marriage between statistics and linguistics led to clear results — not always groundbreaking, sometimes tentative but eventually new, interesting and challenging.

One of the two primary takeaway messages of this thesis is that statistical analysis and reliance on observational data are not biased towards supporting hypotheses coming from any specific theoretical camp. The unraveling of copious sound-meaning associations in unrelated vocabularies goes against classic textbook truths, but my findings on word order patterns are perfectly within the set of predictions that Greenberg produced over half a century ago. Linguists who believe languages are adaptive systems that accommodate to the needs and pressures operating over their users might find their message vindicated through the conclusions regarding tonal languages and humidity, but my conclusions on the origins of creole languages are absolutely compatible with that of the majority of generativists and in stark opposition to the idea of the previously mentioned community. This point cannot be stressed enough: choosing to look at languages through the lens of statistics does not imply taking sides in the (still active) rivalries found between schools of linguistics.

In close connection to this, the remaining important conclusion of this thesis involves reconsidering the role of observational data for the causal understanding of language.

It is clear that linguistics is not anymore a discipline about what we can *observe* or think about language. One the one hand, experiments in all the branches of linguistics grow in quality and ambition: we can intervene in the world and see, for instance, how humans react to non-canonical syntactic constructions or to communication in noisy conditions or how brains process subtle differences in prosody

or meaning. Experiments are the gold standard of science — observational data can be interesting but they provide ultimately only a correlational understanding of the world. On the other hand, computational models developed from simple parametric models into ambitious enterprises where researchers can explore what would happen in synthetic micro-worlds where counter-factual situations like effort-free communication, instantaneous gene-culture co-evolution or mating success tied to lexicon size are possible.

The forceful conclusion — one entertained by more than a few and rarely overtly expressed — is that the analysis of observational data belongs to the prehistory of the discipline, and that cause-effect relations can be only achieved either in experimental setups or in the rational conditions imposed by simulations. I beg to disagree.

As expressed before, languages do not always seem to go in parallel with whatever is best or preferred or easier from the point of view of the language users' needs or goals. For instance, while it is possible to envision scenarios where hypothetically simpler communication systems might replace more quirky ones, we see that creoles carry on, in a rather faithful fashion, the characteristics of their ancestral languages. Biases towards associating specific sounds with meanings in experimental conditions have been reported in almost every single domain of reference, yet we only find a handful of those associations robustly expressed in the languages of the world. The point here is that brains and behaviours of the humans who speak or sign a language are just but one (albeit important, of course) ingredient of language. We know embarrassingly little about the inception of language, but it is unlikely that language could be just a simple scaled-up picture of the psychology, preferences or communicative eagerness of a single human as captured in an experimental setup.

In addition, it seems that the number of factors that can potentially shape the distribution of the world's languages is larger than previously thought, and that their importance and effect are not constant across history or region. Humidity might determine the fate of tonal systems in extreme conditions, but it is unlikely to play any role in less than extreme circumstances. A few word order patterns are robustly attested in the regions of the world, although South America hosts a number of interesting exceptions — maybe the unusual migration and contact history has something to tell us about this. So, instead of the orderly and mechanistic world implicit in computer simulations, we are left with the impression that whatever we believe about what is common or important across languages has resulted from a large number of interacting and contingent forces.

Thus, experiments and simulations are excellent new ways of thinking about language, but they are far from being the silver bullet of linguistic research. The question of "what happens when languages develop in the presence of a myriad of potentially opposing forces

and under the influence of a large number of contingent historical, environmental and social events?" seems to have only one reasonable answer: look at the languages of the world. I tried to do my best to show that, thanks to the data and the methods of the 21st century, this is more feasible and worthy than ever before.

# BIBLIOGRAPHY


[1] George barnards obituary. https://www.theguardian.com/news/2002/aug/09/guardianobituaries.highereducation, 2002. Accessed: 2015-06-20.

[2] Mpi eva data and resources. http://www.eva.mpg.de/linguistics/past-research-resources.html?Fsize=0%25252C%2520%2540%252F, 2015. Accessed: 2015-06-24.

[3] Glottobank. http://glottobank.org/, 2015. Accessed: 2016-06-05.

[4] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[5] Ben Ambridge, Evan Kidd, Caroline F Rowland, and Anna L Theakston. The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(02):239–273, 2015.

[6] Anthony Rodrigues Aristar. On diachronic sources and synchronic pattern: An investigation into the origin of linguistic universals. *Language*, pages 1–33, 1991.

[7] Quentin D Atkinson. Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, 332(6027):346–349, 2011.

[8] Quentin D Atkinson and Russell D Gray. Curious parallels and curious connectionsphylogenetic thinking in biology and historical linguistics. *Systematic biology*, 54(4):513–526, 2005.

[9] Peter Austin et al. Word order in a free word order language: the case of jiwarli. *Forty years on: Ken Hale and Australian languages*, pages 205–323, 2001.

[10] R Harald Baayen, Douglas J Davidson, and Douglas M Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

[11] Peter Baker. Problems of sampling and definition. *Journal of Pidgin and Creole Languages*, 29(2):437455, 2014.

[12] Peter Bakker. Creole languages have nobut they do have. *Journal of Pidgin and Creole Languages*, 2015.

[13] Pierre J Bancel and Alain Matthey de lEtang. Brave new words. *New Perspectives Origins Language*, 144:333, 2013.

[14] K. Bankieris and J. Simner. What is the link between synaesthesia and sound symbolism? *Cognition*, 136:186–195, 2015.

[15] N. Bateman. On the typology of palatalization. *Lang Linguist Compass*, 5:588–602, 2011.

[16] Gareth J Baxter, Richard A Blythe, William Croft, and Alan J McKane. Utterance selection model of language change. *Physical Review E*, 73(4):046118, 2006.

[17] Sandra Beleza, Leonor Gusmao, Antonio Amorim, Angel Carracedo, and Antonio Salas. The genetic legacy of western bantu migrations. *Human genetics*, 117(4):366–375, 2005.

[18] B. K. Bergen. The psychological reality of phonaesthemes. *Language*, 80:290–311, 2004.

[19] Knut Bergsland and Hans Vogt. On the validity of glottochronology. *Current anthropology*, 3(2):115–153, 1962.

[20] B. Berlin. The first congress of ethnozoological nomenclature. *J Roy Anthropol Inst*, 12, S23–S44, 2006.

[21] Robert C Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242, 2011.

[22] Robert C Berwick, Marc Hauser, and Ian Tattersall. Neanderthal language? just-so stories take center stage. *Frontiers in psychology*, 4:671, 2013.

[23] Balthasar Bickel. Statistical modeling of language universals. *Linguistic Typology*, 15(2):401–413, 2011.

[24] Balthasar Bickel. Distributional biases in language families. *Language typology and historical contingency*, pages 415–444, 2013.

[25] Balthasar Bickel and Johanna Nichols. Oceania, the pacific rim, and the theory of linguistic areas. In *Annual Meeting of the Berkeley Linguistics Society*, number 2, pages 3–15, 2006.

[26] Balthasar Bickel, Alena Witzlack-Makarevich, Kamal K Choudhary, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PloS one*, 10(8): e0132819, 2015.

[27] Derek Bickerton. The language bioprogram hypothesis. *Behavioral and brain sciences*, 7(02):173–188, 1984.

[28] Cedric Boeckx. *Syntactic islands*. Cambridge University Press, 2012.

[29] Carlos A Botero, Beth Gardner, Kathryn R Kirby, Joseph Bulbulia, Michael C Gavin, and Russell D Gray. The ecology of religious beliefs. *Proceedings of the National Academy of Sciences*, 111(47):16784–16789, 2014.

[30] Claire Bowern and Quentin Atkinson. Computational phylogenetics and the internal structure of pama-nyungan. *Language*, 88(4):817–845, 2012.

[31] Melvin Joel Bradshaw. *Word order change in Papua New Guinea Austronesian languages*. PhD thesis, 2011.

[32] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16 (3):199–231, 2001.

[33] L. Bromham, X. Hua, T. G. Fitzpatrick, and S. J. Greenhill. Rate of language evolution is affected by population size. *P Natl Acad Sci USA*, 112(7):2097–2102, 2015.

[34] C. H. Brown, E. W. Holman, and S. Wichmann. Sound correspondences in the world's languages. *Language*, 89:4–29, 2013.

[35] Daniel Büring et al. Towards a typology of focus realization. *Information structure*, pages 177–205, 2009.

[36] L. Campbell and W. J. Poser. *Language Classification*. Cambridge University Press, New York, 2008.

[37] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[38] Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.

[39] Robert Chaudenson. *Des îles, des hommes, des langues: essai sur la créolisation linguistique et culturelle*. Editions L'Harmattan, 1992.

[40] Noam Chomsky. *Syntactic structures*. Walter de Gruyter, 2002.

[41] Guglielmo Cinque. Deriving greenberg's universal 20 and its exceptions. *Linguistic inquiry*, 36(3):315–332, 2005.

[42] Diego Colombo and Marloes H Maathuis. A modification of the pc algorithm yielding order-independent skeletons. *Preprint at, http://arxivorg/abs/12113295*, 2012.

[43] Bernard Comrie and Edward L Keenan. Noun phrase accessibility revisited. *Language*, pages 649–664, 1979.

[44] Nicole Creanza, Merritt Ruhlen, Trevor J Pemberton, Noah A Rosenberg, Marcus W Feldman, and Sohini Ramachandran. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences*, 112(5):1265–1272, 2015.

[45] William Croft. *Typology and universals*. Cambridge University Press, 2002.

[46] William Croft, Tanmoy Bhattacharya, Dave Kleinschmidt, D Eric Smith, and T Florian Jaeger. Greenbergian universals, diachrony and statistical analyses. *Linguistic Typology*, 15(2): 433–453, 2011.

[47] C. Cuskley and S. Kirby. Synaesthesia, cross-modality and language evolution. In J. Simner, editor, *Oxford Handbook of Synaesthesia*, pages 869–907. Oxford University Press, Oxford, UK, 2013.

[48] Federica Da Milano and Nicoletta Puddu. Trees, languages and genes: A historical path. In *Understanding Cultural Traits*, pages 341–355. Springer, 2016.

[49] Hal Daumé III and Lyle Campbell. A bayesian model for discovering typological implications. *arXiv preprint arXiv:0907.0785*, 2009.

[50] Aymeric Daval-Markussen. Is hmong njua a creole language? In *Workshop on Non-Indo-European Lexifier, Non-West African Pidgin and Creole Languages, Newcastle University, UK, June*, pages 10–11, 2010.

[51] Aymeric Daval-Markussen. First steps towards a typological profile of creoles. *Acta Linguistica Hafniensia*, 45(2):274–295, 2013.

[52] Dan Dediu and Stephen C Levinson. On the antiquity of language: the reinterpretation of neandertal linguistic capacities and its consequences. *Frontiers in psychology*, 4:397, 2013.

[53] Sally J Delgado. Pirate english of the caribbean and atlantic trade routes in the seventeenth and eighteenth centuries: Linguistic hypotheses based on socio-historical data. *Acta Linguistica Hafniensia*, 45(2):151–169, 2013.

[54] Jared Diamond and Peter Bellwood. Farmers and their languages: the first expansions. *Science*, 300(5619):597–603, 2003.

[55] M. Dingemanse. Advances in the cross-linguistic study of ideophones. *Language Linguist Compass*, 6:654–672, 2012.

[56] M. Dingemanse, F. Torreira, and N. J. Enfield. Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE*, 8:11, 2013.

[57] M. Dingemanse, D. E. Blasi, G. Lupyan, M. H. Christiansen, and P. Monaghan. Arbitrariness, iconicity, and systematicity in language. *Trends Cogn Sci*, 19(10):603–615, 2015.

[58] Robert MW Dixon. *Australian languages: their nature and development*, volume 1. Cambridge University Press, 2002.

[59] Robert MW Dixon. *Basic linguistic theory volume 2: Grammatical topics*, volume 2. Oxford University, 2010.

[60] Mark Donohue and Johanna Nichols. Does phoneme inventory size correlate with population size. *Linguistic Typology*, 15(2): 161–170, 2011.

[61] Mark Donohue, Rebecca Hetherington, James McElvenny, and Virginia Dawson. World phonotactics database. department of linguistics, the australian national university. *World phonotactics database. Department of Linguistics, Australian National University*, 2013.

[62] Matthew S Dryer. The evidence for word order correlations. *Linguistic Typology*, 15(2):335–380, 2011.

[63] Matthew S. Dryer. *Order of Demonstrative and Noun*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `http://wals.info/chapter/88`.

[64] Matthew S. Dryer. *Order of Subject, Object and Verb*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `http://wals.info/chapter/81`.

[65] Matthew S Dryer. On the six-way word order typology, again. *Studies in Language*, 37(2):267–301, 2013.

[66] Michael Dunn, Simon J Greenhill, Stephen C Levinson, and Russell D Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011.

[67] Carol R Ember and Melvin Ember. High cv score: Regular rhythm or sonority? *American Anthropologist*, 102(4):848–851, 2000.

[68] Carol R Ember and Melvin Ember. Climate, econiche, and sexuality: Influences on sonority in language. *American Anthropologist*, 109(1):180–185, 2007.

[69] Ansgar D Endress. Bayesian learning and the psychology of rule induction. *Cognition*, 127(2):159–176, 2013.

[70] Wichmann S. et al. *The ASJP database (version 16)*, 2013. URL `http://asjp.clld.org/Accessedon2013-07-02`.

[71] Nicholas Evans and Stephen C Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(05):429–448, 2009.

[72] Caleb Everett. Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PloS one*, 8(6):e65275, 2013.

[73] Caleb Everett, Damián E Blasi, and Seán G Roberts. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, 112(5):1322–1327, 2015.

[74] Ramon Ferrer-i Cancho. The placement of the head that minimizes online memory. *Language Dynamics and Change*, 5(1):114–137, 2015.

[75] Ramon Ferrer-i Cancho. Non-crossing dependencies: least effort, not grammar. In *Towards a theoretical framework for analyzing complex linguistic networks*, pages 203–234. Springer, 2016.

[76] Corey L Fincher and Randy Thornhill. Parasite-stress promotes in-group assortative sociality: The cases of strong family ties and heightened religiosity. *Behavioral and Brain Sciences*, 35(02): 61–79, 2012.

[77] W Tecumseh Fitch and Daniel Mietchen. 3 convergence and deep homology in the evolution of spoken. *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain*, page 45, 2013.

[78] M. Fort, A. Martin, and S. Peperkamp. Consonants are more important than vowels in the bouba-kiki effect. *Lang Speech*, 58 (2):247–266, 2015.

[79] John G Fought, Robert L Munroe, Carmen R Fought, and Erin M Good. Sonority and climate in a world sample of languages: Findings and prospects. *Cross-cultural research*, 38(1): 27–51, 2004.

[80] Michael C Frank and Joshua B Tenenbaum. Three ideal observer models for rule learning in simple languages. *Cognition*, 120(3):360–371, 2011.

[81] Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585, 2009.

[82] Michael C Frank, Sharon Goldwater, Thomas L Griffiths, and Joshua B Tenenbaum. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125, 2010.

[83] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[84] Francis Galton. I.statistics of mental imagery. *Mind*, (19):301–318, 1880.

[85] Murray Gell-Mann and Merritt Ruhlen. The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, 108(42):17290–17295, 2011.

[86] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[87] Edward Gibson, Steven T Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of crosslinguistic word-order variation. *Psychological science*, page 0956797612463705, 2013.

[88] David Gil. How complex are isolating languages. *Language complexity: Typology, contact, change*, pages 109–131, 2008.

[89] Daniel Gildea and David Temperley. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310, 2010.

[90] Susan Goldin-Meadow, Wing Chee So, Aslı Özyürek, and Carolyn Mylander. The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, 105(27):9163–9168, 2008.

[91] Jeff Good. Typologizing grammatical complexities: or why creoles may be paradigmatically simple but syntagmatically average. *Journal of Pidgin and Creole Languages*, 27(1):1–47, 2012.

[92] Matthew Gordon and Peter Ladefoged. Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406, 2001.

[93] Russell D Gray and Quentin D Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965):435–439, 2003.

[94] Russell D Gray, Alexei J Drummond, and Simon J Greenhill. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *science*, 323(5913):479–483, 2009.

[95] J. H. Greenberg, C. A. Ferguson, and Moravcsik E. A. (Eds.). *Universals of human language: phonology*, volume 2. Stanford University Press, Stanford, CA, 1978.

[96] Joseph H Greenberg. Universals of language. 1966.

[97] Thomas L Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

[98] Tom Güldemann. 1 the" macro-sudan belt": Towards identifying a linguistic area in northern sub-saharan africa 1. 2008.

[99] Matthew L Hall, Victor S Ferreira, and Rachel I Mayberry. Investigating constituent order change with elicited pantomime: A functional account of svo emergence. *Cognitive science*, 38(5): 943–972, 2014.

[100] Harald Hammarström. Ethnologue 16/17/18th editions: A comprehensive review. *Language*, 91(3):723–737, 2015.

[101] Harald Hammarström and Mark Donohue. Some principles on the use of macro-areas in typological comparison. *Language Dynamics and Change*, 4(1):167–187, 2014.

[102] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 2.7. jena: Max planck institute for the science of human history. *Online: http://glottolog. org*, 2016.

[103] M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie. *The World Atlas of Language Structures*. Oxford University Press, Oxford, UK, 2005.

[104] Martin Haspelmath. Comparative syntax. *The Routledge handbook of syntax*, pages 490–508, 2014.

[105] John A Hawkins. *Word order universals: Quantitative analyses of linguistic structure*. New York: Academic Press, 1983.

[106] John A Hawkins. *Cross-linguistic variation and efficiency*. OUP Oxford, 2014.

[107] Jennifer Hay and Laurie Bauer. Phoneme inventory size and population size. *Language*, 83(2):388–400, 2007.

[108] H. Haynie, C. Bowern, and H. LaPalombara. Sound symbolism in the languages of australia. *PLOS ONE*, 9, 2014.

[109] L. Hinton, J. Nichols, and Ohala J. J., editors. *Sound Symbolism*. Cambridge University Press, New York, 2006.

[110] C. F. Hockett. The origin of speech. *Sci Am*, 203:89–96, 1960.

[111] Clare J Holden, Andrew Meade, and Mark Pagel. Comparison of maximum parsimony and bayesian bantu language trees. 2005.

[112] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, V. Müller, and D. Bakker. Explorations in automated language classification. *Folia Linguist*, 42:331–354, 2008.

[113] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[114] M. Imai and S. Kita. The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philos T Roy Soc B*, 369:1651, 2014.

[115] Nina G Jablonski and George Chaplin. Human skin pigmentation as an adaptation to uv radiation. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8962–8968, 2010.

[116] T Florian Jaeger. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446, 2008.

[117] T Florian Jaeger, Peter Graff, William Croft, and Daniel Pontillo. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, 15(2):281–320, 2011.

[118] G. Jäger. Support for linguistic macrofamilies from weighted sequence alignment. *P Natl Acad Sci USA*, 112(41):12752–12757, 2015.

[119] Gerhard Jäger. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155–204. Brill, 2014.

[120] R. Jakobson. Why "mama" and "papa". In *Perspectives in Psychological Theory, Dedicated to Heinz Werner*, pages 124–134. Kaplan B and Wapner S, International Universities Press, New York, 1960.

[121] Roman Jakobson. Why mama and papa. *Selected writings*, 1: 538–545, 1962.

[122] Silke Janitza, Carolin Strobl, and Anne-Laure Boulesteix. An auc-based permutation variable importance measure for random forests. *BMC bioinformatics*, 14(1):1, 2013.

[123] Mark Jobling, Matthew Hurles, and Chris Tyler-Smith. *Human evolutionary genetics: origins, peoples & disease*. Garland Science, 2013.

[124] N. Johansson and J. Zlatev. Motivations for sound symbolism in spatial deixis: a typological study of 101 languages. *The Public J Sem*, 5:3–2, 2013.

[125] John S Justeson and Laurence D Stephens. On the relationship between the numbers of vowels and consonants in phonological systems. *Linguistics*, 22(4):531–546, 1984.

[126] John S Justeson and Laurence D Stephens. Explanations for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists*, volume 3, pages 2372–76, 1990.

[127] Richard S Kayne. *The antisymmetry of syntax*. Number 25. Mit Press, 1994.

[128] Kathryn R Kirby, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, Carlos A Botero, Claire Bowern, Carol R Ember, et al. D-place: A global database of cultural, linguistic and environmental diversity. *PLoS One*, 11(7):e0158391, 2016.

[129] W. Köhler. *Gestalt Psychology*. Liveright, New York, 1929.

[130] Jianjing Kuang. The tonal space of contrastive five level tones. *Phonetica*, 70(1-2):1–23, 2013.

[131] Randy J LaPolla. Pragmatic relations and word order in chinese. *Word order in discourse*, 30:297, 1995.

[132] Claire Lefebvre. *Creole genesis and the acquisition of grammar: The case of Haitian Creole*, volume 88. Cambridge University Press, 2006.

[133] Julie Anne Legate and Charles D Yang. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 18(1-2): 151–162, 2002.

[134] Christian Lehmann. Thoughts on grammaticalization. Seminar für Sprachwiss. der Univ., 2002.

[135] Winfred P Lehmann. A structural principle of language and its implications. *Language*, pages 47–66, 1973.

[136] Stephen C Levinson and Russell D Gray. Tools from evolutionary biology shed new light on the diversification of languages. *Trends in cognitive sciences*, 16(3):167–173, 2012.

[137] M Paul Lewis, Gary F Simons, and Charles D Fennig. Ethnologue: Languages of the world. dallas, texas: Sil international. online version, 2015.

[138] P. Lewis, G. Simons, and C. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, 2013.

[139] Jeffrey Lidz and Annie Gagliardi. How nature meets nurture: Universal grammar and statistical learning. *Annu. Rev. Linguist.*, 1(1):333–353, 2015.

[140] Dina Lipkind, Gary F Marcus, Douglas K Bemis, Kazutoshi Sasahara, Nori Jacoby, Miki Takahasi, Kenta Suzuki, Olga Feher, Primoz Ravbar, Kazuo Okanoya, et al. Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, 498(7452):104–108, 2013.

[141] Chang Liu. Just noticeable difference of tone pitch contour change for english-and chinese-native listeners. *The Journal of the Acoustical Society of America*, 134(4):3011–3020, 2013.

[142] G. Lockwood and M. Dingemanse. Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Front Psychol*, 6, 2015. doi: http://dx.doi.org/10.3389/fpsyg.2015.01246.

[143] Jonathan B Losos. Convergence, adaptation, and constraint. *Evolution*, 65(7):1827–1840, 2011.

[144] V. U. Ludwig, I. Adachi, and T. Matsuzawa. Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (pan troglodytes) and humans. *Proc Natl Acad Sci USA*, 108(51):20661–20665, 2011.

[145] P. F. MacNeilage and B. L. Davis. On the origin of internal structure of word forms. *Science*, 288(5465):527–531, 2000.

[146] Ian Maddieson. Correlating phonological complexity: data and validation. 2005.

[147] Ian Maddieson. Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. 2005.

[148] Ian Maddieson and Christophe Coupé. Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America*, 138(3):1838–1838, 2015.

[149] Ian Maddieson and Sandra Ferrari Disner. *Patterns of sounds*. Cambridge university press, 1984.

[150] Ian Maddieson, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. Lapsyd: lyon-albuquerque phonological systems database. In *INTERSPEECH*, pages 3022–3026, 2013.

[151] André Martinet. Économie des changements phonétiques, 1956.

[152] Colin P Masica. *Defining a linguistic area: South Asia*. Orient Blackswan, 2005.

[153] Yaron Matras, April McMahon, and Nigel Vincent. *Linguistic areas: convergence in historical and typological perspective*. Springer, 2006.

[154] D. Maurer, T. Pathman, and C. J. Mondloch. The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Sci*, 9:316–322, 2006.

[155] Luke Maurits and Thomas L Griffiths. Tracing the roots of syntax with bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 111(37):13576–13581, 2014.

[156] John McWhorter. The worlds simplest grammars are creole grammars. *Linguistic typology*, 5(2/3):125–166, 2001.

[157] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 1995.

[158] Matthias R Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard B Slatcher, and James W Pennebaker. Are women really more talkative than men? *Science*, 317(5834):82–82, 2007.

[159] Natalia R Mesa, María C Mondragón, Iván D Soto, María V Parra, Constanza Duque, Daniel Ortíz-Barrientos, Luis F García, Iván D Velez, María L Bravo, Juan G Munera, et al. Autosomal, mtdna, and y-chromosome diversity in amerinds: pre-and post-columbian patterns of gene flow in south america. *The American Journal of Human Genetics*, 67(5):1277–1286, 2000.

[160] Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber. *The atlas of pidgin and creole language structures*. Oxford University Press, 2013.

[161] P. Monaghan, M. H. Christiansen, and N. Chater. The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. *Cognitive Psychol*, 55:259–305, 2007.

[162] P. Monaghan, M. H. Christiansen, and S. A. Fitneva. The arbitrariness of the sign: learning advantages from the structure of the vocabulary. *J Exp Psychol Gen*, 140:325–347, 2011.

[163] Lorna G Moore. Human genetic adaptation to high altitude. *High altitude medicine & biology*, 2(2):257–279, 2001.

[164] Steven Moran, Daniel McCloy, and Richard Wright. Revisiting population size vs. phoneme inventory size. *Language*, 88(4): 877–893, 2012.

[165] Steven Moran, Daniel McCloy, and Richard Wright. Phoible online. *Leipzig: Max Planck Institute for Evolutionary Anthropology*, 2014.

[166] Eugene S Morton. On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, pages 855–869, 1977.

[167] Salikoko S Mufwene. *The ecology of language evolution*. Cambridge University Press, 2001.

[168] Salikoko S Mufwene. Sla and the emergence of creoles. *Studies in Second Language Acquisition*, 32(03):359–400, 2010.

[169] Roger Mundry. Statistical issues and assumptions of phylogenetic generalized least squares. In *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 131–153. Springer, 2014.

[170] Robert L Munroe, Ruth H Munroe, and Stephen Winters. Cross-cultural correlates of the consonant-vowel (cv) syllable. *Cross-Cultural Research*, 30(1):60–83, 1996.

[171] Robert L Munroe, John G Fought, and Ronald KS Macaulay. Warm climates and sonority classes: Not simply more vowels and fewer consonants. *Cross-Cultural Research*, 2009.

[172] D. J. Napoli, N. Sanders, and R. Wright. On the linguistic effects of articulatory ease, with a focus on sign languages. *Language*, 90(2):424–456, 2014.

[173] Donna Jo Napoli and Rachel Sutton-Spence. Order of the major constituents in sign languages: implications for all language. 2014.

[174] Daniel Nettle. Language and genes: A new perspective on the origins of human cultural diversity. *Proceedings of the National Academy of Sciences*, 104(26):10755–10756, 2007.

[175] Daniel Nettle. Ecological influences on human behavioural diversity: a review of recent findings. *Trends in ecology & evolution*, 24(11):618–624, 2009.

[176] Frederick J Newmeyer. On the reconstruction of proto-world word order. *The evolutionary emergence of language*, pages 372–388, 2000.

[177] Frederick J Newmeyer. *Possible and probable languages: A generative perspective on linguistic typology*. Oxford University Press, 2005.

[178] Johanna Nichols. *Linguistic complexity: a comprehensive definition and survey*. Oxford University Press, 2009.

[179] Johanna Nichols and Tandy Warnow. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820, 2008.

[180] Johanna Nichols, Alena Witzlack-Makarevich, and Balthasar Bickel. The autotyp genealogy and geography database: 2013 release. *Zurich: University of Zurich*, 2013.

[181] A Nicolaidis, Kosmas Kosmidis, and Panos Argyrakis. A random matrix approach to language acquisition. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(12):P12008, 2009.

[182] A. Nielsen and D. Rendall. The sound of round: evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon. *Can J Exp Psychol*, 65:115–124, 2011.

[183] Marlijn L Noback, Katerina Harvati, and Fred Spoor. Climate-related variation of the human nasal cavity. *American journal of physical anthropology*, 145(4):599–614, 2011.

[184] S. Nordhoff, H. Hammarström, R. Forkel, and M. Haspelmath. Glottolog 2. 1, 2013. URL `http://glottolog.orgAccessedon2013-07-02.`

[185] B. V. North, D. Curtis, and P. C. Sham. A note on the calculation of empirical p values from monte carlo procedures. *Am J Hum Gen*, 71:439–441, 2002.

[186] J. B. Nuckolls. The case for sound symbolism. *Annu Rev Anthropol*, 28:225–252, 1999.

[187] L. C. Nygaard, A. E. Cook, and L. L. Namy. Sound to meaning correspondences facilitate word learning. *Cognition*, 112:181–186, 2009.

[188] M. Pagel, Q. D. Atkinson, A. S. Calude, and A. Meade. Ultra-conserved words point to deep language ancestry across eurasia. *Proc Natl Acad Sci USA*, 110:8471–8476, 2013.

[189] Marco Patriarca and Teemu Leppänen. Modeling language competition. *Physica A: Statistical Mechanics and its Applications*, 338(1):296–299, 2004.

[190] Judea Pearl. *Causality*. Cambridge university press, 2009.

[191] François Pellegrino, Christophe Coupé, and Egidio Marsico. Across-language perspective on speech information rate. *Language*, 87(3):539–558, 2011.

[192] Julian M Pine and Elena VM Lieven. Slot and frame patterns and the development of the determiner category. *Applied psycholinguistics*, 18(02):123–138, 1997.

[193] Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. Do young children have adult-like syntactic categories? zipfs law and the case of the determiner. *Cognition*, 127(3):345–360, 2013.

[194] S. Pinker. *Words and Rules: The Ingredients of Language*. Perseus Books, New York, 1999.

[195] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[196] M. I. Proctor. *Gestural characterization of a phonological class: The liquids*. PhD dissertation, Yale University, 1995.

[197] V. S. Ramachandran and E. M. Hubbard. Synaesthesia—a window into perception, thought and language. *J Consciousness Stud*, 8:3–34, 2001.

[198] J. Reilly, J. Hung, and C. Westbury. *Non arbitrariness in mapping word form to meaning: cross-linguistic formal markers of word concreteness*. Cognitive Sci, 2016. doi: 10.1111/cogs.12361.

[199] Douglas G Richards and R Haven Wiley. Reverberations and amplitude fluctuations in the propagation of sound in a forest: implications for animal communication. *American Naturalist*, pages 381–399, 1980.

[200] Jan Rijkhoff. On the (un) suitability of semantic categories. *linguistic Typology*, 13(1):95–104, 2009.

[201] Seán G Roberts, James Winters, and Keith Chen. Future tense and economic decisions: controlling for cultural evolution. *PloS one*, 10(7):e0132145, 2015.

[202] M. Ruhlen. *On the Origin of Languages: Studies in Linguistic Taxonomy*. Stanford University Press, Stanford, CA, 1994.

[203] Wendy Sandler, Irit Meir, Carol Padden, and Mark Aronoff. The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2661–2665, 2005.

[204] F. D. Saussure. *Cours de linguistique générale*. ed Bally C, Sechehaye A, Riedlienger A (Payot, Paris, 1916.

[205] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[206] Marieke Schouwstra and Henriëtte de Swart. The semantic origins of word order. *Cognition*, 131(3):431–436, 2014.

[207] Yajuan Si and Jerome P Reiter. Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013.

[208] Jeff Siegel. *The emergence of pidgin and creole languages*. Oxford University Press, 2008.

[209] Jae Jung Song. *Word order*. Cambridge University Press, 2012.

[210] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[211] Craig Stanford, John S Allen, and Susan C Antón. *Biological anthropology: the natural history of humankind*. Pearson, 2016.

[212] S. A. Starostin and Bronnikov Y. *Languages of the World Etymological Database*. Available at Part of the Tower of Babel - Evolution of Human Language Project, 1998. URL `http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl`.

[213] Lydia Steiner, Peter F Stadler, and Michael Cysouw. A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1):89–127, 2011.

[214] Stephen M Stigler. *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press, 2002.

[215] William C Stokoe. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

[216] Sabine Stoll. Studying language acquisition in different linguistic and cultural settings. *The Routledge Handbook of Linguistic Anthropology*, page 140, 2015.

[217] K. Strimmer. fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(12): 1461–1462, 2008.

[218] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1, 2007.

[219] M. Swadesh. Towards greater accuracy in lexicostatistic dating. *Int J Am Linguist*, 21:121–137, 1955.

[220] Morris Swadesh et al. What is glottochronology. *The origin and diversification of languages*, pages 271–284, 1972.

[221] Dallas M Swallow. Genetics of lactase persistence and lactose intolerance. *Annual review of genetics*, 37(1):197–219, 2003.

[222] Daniel Swingley. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1536):3617–3632, 2009.

[223] Douglas Taylor. Language shift or changing relationship? *International Journal of American Linguistics*, 26(2):155–161, 1960.

[224] Stefan Th. Gries. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1):95–125, 2015.

[225] Sarah Grey Thomason and Terrence Kaufman. *Language contact*. Edinburgh University Press Edinburgh, 2001.

[226] Andrei Nikolaevich Tikhonov. Regularization of incorrectly posed problems. SOVIET MATHEMATICS DOKLADY, 1963.

[227] Michael Tomasello. The cultural roots of language. *Communicating meaning: The evolution and development of language*, pages 275–307, 1996.

[228] Michael Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2009.

[229] Olga M Tomic. *Balkan Sprachbund morpho-syntactic features*, volume 67. Springer Science & Business Media, 2006.

[230] Russell S Tomlin. *Basic Word Order (RLE Linguistics B: Grammar): Functional Principles*, volume 13. Routledge, 2014.

[231] H. Traunmüller. Sound symbolism in deictic words. In *Tongues and Texts Unlimited. Studies in Honour of Tore Jansson on the Occasion of his Sixtieth Anniversary*, pages 213–234. Janson T, Aili H, af Trampe P (Stockholms Universitet, Institutionen för klassiska språk, Stockholm, 1994.

[232] M. Urban. Conventional sound symbolism in terms for organs of speech: a cross-linguistic study. *Folia Linguist*, 45:199–214, 2011.

[233] Robert D Van Valin Jr. *Exploring the syntax-semantics interface*. Cambridge University Press, 2005.

[234] Viveka Velupillai. *Pidgins, Creoles and mixed languages: an introduction*, volume 48. John Benjamins Publishing Company, 2015.

[235] Theo Vennemann. Categorial grammar and the order of meaningful elements. *Linguistic studies offered to J. Greenberg*, 3:615–34, 1976.

[236] M. S. Vitevitch and P. A. Luce. Phonological neighborhood effects in spoken word perception and production. *Annu Rev Linguistics*, 2:75–94, 2016.

[237] Scott I Vrieze. Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, 17(2):228, 2012.

[238] Robert S Walker and Lincoln A Ribeiro. Bayesian phylogeography of the arawak expansion in lowland south america. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1718): 2562–2567, 2011.

[239] S. Wichmann, E. W. Holman, and C. H. Brown. Sound symbolism in basic vocabulary. *Entropy*, 12:844–858, 2010.

[240] Søren Wichmann. On the power-law distribution of language family sizes. *Journal of Linguistics*, 41(1):117–131, 2005.

[241] Matthew R Wilkins, Nathalie Seddon, and Rebecca J Safran. Evolutionary divergence in acoustic signals: causes and consequences. *Trends in ecology & evolution*, 28(3):156–166, 2013.

[242] Charles Yang. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327, 2013.

[243] Chien Yuehchen and Sanada Shinji. Yilan creole in taiwan. *Journal of Pidgin and Creole languages*, 25(2):350–357, 2010.

[244] Damian H Zanette. Self-similarity in the taxonomic classification of human languages. *Advances in Complex Systems*, 4 (02n03):281–286, 2001.

[245] Lucía Ziegler, Matías Arim, and Peter M Narins. Linking amphibian call structure to the environment: the interplay between phenotypic flexibility and individual attributes. *Behavioral Ecology*, page arr011, 2011.

[246] Klaus Zuberbühler. Linguistic capacity of non-human animals. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3):313–321, 2015.