Précis of *Linguistic Diversity Through Data*

Damián E. Blasi
Radcliffe Institute for Advanced Study, Harvard University

Introduction: linguistic diversity between cognition and human history

Within the cognitive sciences, worldwide linguistic diversity data play (barring a few exceptions) a modest role in attempts to explain the structure and function of language. More often than not, linguistic diversity is brought up either quantitatively in the form of statements regarding how cross-linguistically prevalent a given linguistic feature is qualitatively as a lively picture of the astonishing plasticity of the linguistic phenotype.

This state of affairs is not haphazard: experimental and computational studies of language have become the absolute gold standard of evidence within this subset of the cognitive sciences - understandably so. Leaving aside (still ongoing) discussions on how far can inferences with observational data go in terms of causal guarantees, the biggest culprit is human history. Languages are not independent realizations of a pristine language experiment of sorts, and the titanic task of disentangling history from cognition and function in the structure of languages has been the major impediment for a bigger role of linguistic diversity in the cognitive sciences.

To start with, human history is not keen of latin squares, and instead it is fond of heavily biased distributions of the conditions we do get to see. A handful of language families (Indo-European, Austronesian, Atlantic-Congo, Sino-Tibetan and Nuclear New Guinea) monopolize the majority of the world's languages, and they do so because of reasons that are orthogonal to the fitness of their languages to whatever functions they carry on. The domestication of the sweet potato, the tameness of the llama and the adoption of exogamy are the main factors that account for the overabundance of specific language structures in the Pacific, the Andes and Australia (just to name a few cases.) This is perhaps why the utility of linguistic diversity data has been circumscribed to the few cases with non-trivial distributions, such as the relative ordering of the (transitive) verb and its object, the presence of tone, or the

morphology/syntax split in relation to how languages express tense. For most other interesting linguistic phenomena of relevance for the cognitive sciences, we find that either almost all languages have it (e.g. a morphosyntactic divide between nouns and verbs) or they almost certainty don't (e.g. strong constraints on clausal embedding depth), and it takes only a few families to go from one situation to the other.

But even in those cases where there is enough observed variation in the languages of the world, a number of practical and theoretical complications hinder straightforward inferences. The typical statistical repertoire of the cognitive scientist is -again, barring a few exceptions- not tailored to the statistical challenges of human history. Differences between items and individuals can be efficiently captured by group (random) effects with minimal statistical specification, but similarities between languages due to history are substantially more complex. Language areas -regions of the world that experience extensive borrowing of linguistic material- usually display non-trivial structure, so that e.g. a main river displays a clear uniform structural profile that fades away gradually as one goes into its tributaries, or they may indicate the extension of a forgotten empire that dominated over that region for which scarce archeological evidence has been found. Language families are not homogenous categories either, and their similarity is often modelled (in contemporary computational historical linguistics) in the fashion of phylogenetic similarities following a probability distribution over plausible genealogical trees.

But even if the appropriate statistical tools were imported to accommodate those covariance structures, the assumptions underlying the reconstruction of shared language history collide with the implicit assumption in the study of linguistic diversity by cognitive scientists. The foremost approach to the question of linguistic diversity in the cognitive sciences, language evolution studies and functional/typological linguistics involves some notion of *language fitness*. Linguistic structures frequently found across the globe are regularly taken to indicate species-wide preferences; for instance, towards a particular ordering of the elements of the verb and its arguments, a systematic configuration of the vowel space or a given lexical architecture mapping a semantic domain into its representation. Conversely, linguistic *rara* and *rarissima* are interpreted as extreme configurations of the design space of languages, and as such are regarded to involve dispreferred, suboptimal and/or unstable traits.

In contrast, similar observations on linguistic diversity have been garnered for the purpose of inferences about the history of populations, cultures, and languages. In this camp we find mostly historical linguists (and their more recent computational spin), quantitative anthropologists and experts on cultural evolution. Regular sound changes, vocabulary items and structural properties have been leveraged for reconstructing the movement and affiliation of human groups everywhere from Tasmania to the Caucasus and the Amazon, which has resulted in the identification of over 300 families of languages with origins as deep as 10,000 years before present. The models embraced in this tradition make heavy use of the hypothesis that *most* change in language is neutral, so frequent or infrequent linguistic traits are taken to reflect the breadth and reach of the human groups bearing those languages.

Hence these traditions take different vantage points on the same object of study: language as a communication system embedded in human cognition and behavior, on the one hand, and language as a (mostly neutral) cultural complex that indexes human history. Cross-talk, while present, has been limited due to several factors. The methodological machinery used by the first camp draws from psychology and the cognitive sciences more in general, whereas the second has recruited numerous strategies taken from evolutionary biology and bioinformatics. The sets of journals where these communities publish have a modest intersection, which is simply the symptom of a more substantial situation: the average researcher in one side of the divide is in general not aware of the important findings and results from the other camp. This results in one side being generally naive about the patterns of history whereas the other ignores the many findings that point out to the fact that humans do not learn, use and transmit linguistic material in a neutral manner.

The thesis

My thesis comes as a response to the aforementioned gap. I attempted to bridge the gap between these two perspectives on linguistic diversity (cognitive/functional and historical) by exploiting the rich distributional patterns of languages through time and space (instead of regarding those as a statistical nuisance that should be controlled for.) More precisely, I evaluated, within a handful of domains, the pool of hypotheses about language structure concocted by cognitive scientists by studying how languages spread in space and change through time, ultimately aiming at pinning down the relative contribution of systematic functional/cognitive biases against a baseline of conservative transmission of culture blind to its content.

In addition to this conceptual backbone, a main goal of the thesis was to show that the nature of (cross-)linguistic data is unique, and as such it requires tailored methods that cannot be taken off the shelf from psychology or evolutionary biology. Hence every single chapter formulates, combines and introduces novel ways of thinking, exploring and inferring from data. Most chapters start with considerations about the geometry of the data at hand which (in addition to the research question) determine which inferential tools will be used in the end.

Finally, I wanted to make the most out of the handful of open databases built around the notion of capturing linguistic diversity, and I actively looked for collaborators that could supply the necessary expertise on the relevant domains. As a result, the thesis covers very diverse aspects of language, including grammar (Ch3, *dependencies in word order patterns*), phonology and phonetics (Ch5, *ecological pressures on speech sounds*), lexica and semantics (Ch2, *non-arbitrary sound-meaning associations*) as well as large-scale typological profiles (Ch4, *creoles as a typological group*).

Chapter 1 (Introduction) sets up the historical circumstances that gave rise to data science and the study of linguistic diversity, and outlines the challenges and promises of the interdisciplinary method embraced in the following chapters.

Chapter 2 (Sound-Meaning Associations) sets out to determine whether recurrent sound-meaning associations can be found in the largest basic vocabulary database available (covering over 2/3 of the world's languages) while at the same time asking whether their spatial and historical distribution can reveal something about their nature. A very conservative approach was taken, so that the statistical focus is on minimizing false positives and the dataset (wordlists of basic vocabulary) was built under the premise of the absence of sound symbolism - hence we were aiming at determining a lower bound to the presence of non-arbitrary sound-meaning associations.
Data: *Automated Similarity Judgment Program* (ASJP), open access available at asjp.clld.org
Results: a whopping 40% of all associations tested turned out to display a consistent pattern across most families and areas, including classic ones like high front vowels and 'small', liquids and 'round', labial consonants and 'breasts', and novel ones such as liquids and 'tongue', nasal consonants and 'nose', and labial consonants and 'leaf'. More strikingly, these associations do not seem to result from extraordinary phylogenetic or areal persistence, which suggests languages produce *de novo* such associations - in agreement with the experimental evidence showing that a main function of non-arbitrariness is to speed up conventionalization (after which it can be swept away by other pressures such as economy or systematicity.)

Chapter 3 (Dependencies in Word Order Patterns) revisits the old question of word order patterns (i.e. bundles of word order positions that are regularly found in tandem, e.g. verb-before-object and genitive-before-noun) by fleshing out novel empirical predictions out of the few competing theories available (which involve factors as diverse as biases in the human parser, grammaticalization and cultural evolution) and testing them in a causal inference framework with independent tests for spatial and genealogical effects.
Data: *World Atlas of Language Structures* (WALS), open access available at wals.info
Results: a causal graph displaying a recurrent set of disjoint noun phrase and verb phrase word order patterns (largely in agreement with the existing literature) emerges, and the relative lack of higher-order interactions beyond those clusters suggests that ease of processing or grammaticalization (*contra* accounts that claim no universal tendencies in word order) might be the strongest candidates to explain these findings.

Chapter 4 (Creoles as a Typological Group) provides a rigorous and fleshed out test on the status of creole languages as displaying a unique linguistic profile resulting from their peculiar context of emergence. Within the cognitive sciences, creoles are often invoked as languages that have innovated structure across the board thus revealing linguistic biases that are overridden in normal situations of language acquisition and transmission.
Data: *Atlas of Pidgin and Creole Structures* (APiCS), open access available at apics-online.info
Results: while creole languages can be statistically differentiated from non-creoles (as taken from a worldwide balanced sample), convergent patterns in the data reveal that creoles overwhelmingly continue the typological traits of their ancestor languages in a fashion that reveals some systematicity in the process (e.g. word order patterns usually derive from lexifiers, whereas the coding of the valency of the verb can be put in relation to the substrates),

revealing once more that language acquisition is surprisingly robust and efficient, even in extreme circumstances. These results leave no room for the presumed massive loss (and posterior innovation) of structure during creole genesis that is embraced in the cognitive sciences.

Chapter 5 (Ecological Pressures on Speech Sounds) draws inspiration from the well established literature on ecological factors shaping the space of vocalizations in animal communication systems to ask whether there are any detectable traces of ecological pressures on human speech systems. Ecological adaptations have been systematically sidelined by more cognitive-oriented researchers under the assumption that the main pressure shaping human sound systems involves the joint optimization of perceptual spread on the auditory dimension with as little overlap between phonemes as possible.
Data: PHOIBLE, open access available at phoible.org
Results: the observed distributional patterns of phonological systems are consistent with at least two biases rooted ultimately in ecology (the avoidance of complex tonal systems in dry regions and the comparatively small number of stops in parts of the world with dense vegetation)

Importance

The final set of issues tackled in the thesis was largely determined by (i) cross-linguistic data availability, (ii) reachable experts open to collaboration and (iii) potential to advance our understanding about the cognition-function/history divide. Fortunately, these criteria yield a handful of core questions on the nature of language (as reviewed above), including the prevalence of the same sound-meaning associations across completely independent languages, the origin of linguistic structure in the extraordinary creole languages, the subtle patterns of adaptation of human speech sounds to ecology, and the perennial question of whether word order patterns are functionally (or historically) associated in bundles.

The three chapters that led to publications in major outlets[1], including articles in the *Proceedings of the National Academy of Sciences* (two papers), *Trends in Cognitive Sciences* and *Nature Human Behaviour*. The interest on this research transcended the academic audience and the work in my thesis has been featured in many outlets of over 60 countries, including *Nature News and Views*, *The Economist*, *Scientific American*, *L.A. Times*, *The Guardian* and others.

Novelty

On choosing the main questions of my thesis I followed the guide of well-established big questions with a substantial academic history, so I cannot claim novelty in that domain. But at the end of the day, when the analysis of empirical data is put at the center of science-making, how is that those questions are parametrized in a statistically sound manner is as important as the question itself (I submit.) In this regard, every chapter of my thesis aims at faithfully

[1] See the full list of publications in the thesis

translating verbal hypotheses into testable and robust statistical models with total explicitness. Sometimes this resulted in discerning two or more effective hypotheses under the same label, finding a lack of discussions in the experimental literature on what is the appropriate unit of measurement of an effect, or determining clear weaknesses in the type of evidence discussed so far to test a theory. The task of painstakingly building statistical models capable of making justice to expert knowledge all while striking the right balance between statistical identifiability and richness of expression was by far the most demanding one during my years as a Ph.D. candidate.

Methodologically, I believe my thesis introduced for the first time a number of statistical tools in the context of the discussion of linguistic diversity - including the inference of causal structure through the PC algorithm and its derivatives, some measures of multi-information, PCA regression and local FDR, among others.

Interdisciplinary contribution

Given my non-standard academic path (which spans degrees in statistical physics and computer sciences and almost a decade of continued affiliation in linguistic, psycholinguistic and anthropological departments) and the nature of my research, I deem all of my scientific production (including my thesis) to be interdisciplinary.

I ignore what is the best predictor of impact across disciplines beyond the standardized (and faulty) citation metrics, although my personal experience dictates that different scientific communities display substantially contrasting values in this regard (e.g. being published in high IF journals is often a motive of disdain in many traditional linguistic groups.) Barring this issue, I would suggest that perhaps the best estimator of interdisciplinary impact is the ecology of papers (and disciplines) citing the work derived from the thesis. This research has been picked up by four partially overlapping communities: cognitive sciences, language evolution, cultural evolution and comparative linguistics. Yet a cursory look at the articles citing this work reveals that the impact goes far beyond, including journals of developmental psychology, animal communication, education, phonetics, physics of social systems, economics, anthropology, musicology and data sciences.