Coupé, C., Hombert, J.-M., Marsico, E. & Pellegrino, F. (2013)

6

Investigations into Determinants of the Diversity of the World's Languages

Christophe COUPÉ, Jean-Marie HOMBERT,
Egidio MARSICO, François PELLEGRINO
Laboratoire Dynamique du Langage
(CNRS – University of Lyon)

Foreword

During 2011, we had the occasion to meet Professor William Wang in different meetings in Lyon, Hong Kong and Shanghai. During these encounters, results obtained by Quentin Atkinson and published earlier in the year (Atkinson 2011a) were a topic of hot discussion. Prof. Wang published a short paper in one of Southern China's main newspapers, where he defended the possibility of multi-regional evolution of languages, and more generally suggested not to pay only attention to the global at the expense of local phenomena. With this idea in mind, we decided to adapt the statistical approach employed by Atkinson in order to consider local ecological contexts as potential determinants of linguistic and sociolinguistic phenomena. We hope that the results we present will be seen as a step towards a better acknowledgment of the richness and complexity of language, both aspects Prof. Wang especially led us to study and value.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

1. Introduction: Identifying Determinants of Linguistic Phenomena

1.1 Phonemic diversity and the Out of Africa model

In a recent and heavily debated paper, Atkinson (2011a) put forward what he called a 'serial founder effect' in the distribution of the phonemic diversity of the world's languages. Relying on data from the WALS database (Dryer and Haspelmath 2011) and statistical analysis, he concluded that the size of the phonemic inventories of languages declined with distance from Africa, in the same way human genetic and phenotypic diversities do.

The concept of founder effect was borrowed from genetics, where a small group leaving a larger one only carries a portion of the genetic diversity of the latter. The relevance of this concept in linguistics has been attacked by linguists stating that no speakers were ever leaving a larger population with only a subset of the phonemes of its language (e.g., Maddieson, Bhattacharya, Smith and Croft 2011). Atkinson's core assumption was however different and perhaps not very convincingly described by the notion of founder effect. In fact, Atkinson relied on a previous study which found a correlation between the number of speakers of a language and the number of phonemes of this language (Hay and Bauer 2007). The reasoning, as we understood it, is as follows: if the Out of Africa mainly took place through the repetition of small groups leaving larger ones to go further away from the origin point of the migration, then these small groups, although initially leaving with the complete phonemic inventory of their language, could have later and gradually lost some of their phonemes. A gradual impoverishment would have thus resulted from the repetition of scissions of small populations from larger ones, with a spread from the origin point to the most recently colonized areas.

Atkinson's results were obtained on the basis of compiled quantitative data and statistical models. While some typologists have acknowledged that the approach was innovative with respect to typological issues (Jaeger, Graff, Croft and Pontillo 2011), many of its aspects came under criticism (Bybee 2011).[1] The use of a single quantitative measure of phonemic diversity was first contested (Maddieson et al. 2011), as well as the assimilation between the concepts of diversity and complexity (Ross and Donohue 2011). The positive correlation between the number of speakers and the size of the phoneme inventory of a language was rebutted on the basis of either specific examples of linguistic families (Rice 2011; Dahl 2011) or analyses of larger datasets of languages (Wichmann, Rama and Holman 2011; Donohue and Nichols 2011; Pericliev 2011). Its relevance during pre-Holocene times—when human populations were small and sparse—was also questioned (Sproat 2011). Ringe (2011) more specifically investigated the processes by which phonemes are lost or created in languages, and Trudgill (2011) insisted that, in general, migration does not lead to inventory reduction. Finally, specific assumptions behind the statistical models used by Atkinson were questioned, and the large number of small-sized linguistic families was identified as a limit preventing from investigating the potential weight of genetic groupings (Jaeger et al. 2011).

[1] In the fall of 2011, a half special issue of the journal Linguistic Typology dealt with Atkinson's proposal.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

The considerable amount of attention received by Atkinson's study showed that today's typology is not yet at ease with the use of quantitative and statistical approaches. It also highlighted the need for a better understanding of the relationship between social factors and linguistic ones.

Atkinson's study followed other works which focused on the relationship between the phonemic diversity of a language and its number of speakers. Trudgill stressed the contradicting hypotheses regarding the possible interaction between these two elements: on the one hand, a smaller population may have a smaller phonemic inventory, since a higher degree of interpersonal knowledge in tight social networks might result in a weaker need for phonological differentiation between words (Trudgill 2002). On the other hand, a smaller population size may imply easier adherence to linguistic norms, and therefore larger inventories (Trudgill 2004). Hay and Bauer (2007)'s study suggested a positive correlation between the number of speakers and the number of phonemes. However, the authors didn't propose any convincing explanation for it. Their results diverged from Pericliev (2004)'s statement of a lack of correlation between the number of speakers and the size of the consonantal inventory. Finally, a strong relationship was recently suggested between a range of sociolinguistic factors and the degree of complexity of morphological and syntactic structures (Lupyan and Dale 2010). Walker and Hamilton (2010) also stressed a link between social complexity and linguistic diversity during Austronesian and Bantu population expansions.[2]

In this context, we wish to stress the relevance of environmental factors to understand linguistic and sociolinguistic phenomena. In other words, we propose to go 'upstream' in the investigation of the cascade of factors that influence the current global linguistic situation. At the methodological level, we conduct statistical analyses on linguistic data and high-resolution biophysical and demographic datasets, which accurately describe the 'ecology' of today's languages.

In the remaining of this introductory section, we report studies which previously highlighted the impact of environmental factors. In the rest of this contribution, we introduce our sources of data and our statistical approach, before presenting and discussing the results.

1.2 Environmental factors and linguistic situations

Several of Atkinson's critics mentioned Nettle's pioneering works with respect to linguistic diversity across large geographic areas (Nettle 1999a). Nettle especially brought to light the role that ecological factors could play in influencing social aspects of language; he focused on

[2] Such studies often define linguistic complexity by counting elements, or by considering the occurrence of specific features. Finer-grained approaches exist—e.g., (Coupé, Marsico and Pellegrino 2009; Maddieson 2009) regarding phonological complexity, but they require data difficult to collect for a large number of languages.

the concepts of ecological risk and growing season to explain the geographic extension and population size of linguistic groups in tropical regions (Nettle 1996, 1998). His arguments can be summarized as follows: i) the yearly duration of the growing season of plants is related to the risk faced by people of suffering from difficulties of getting the resources they need for their subsistence—the ecological risk; ii) the greater this risk, the stronger the need for a geographically wide social network, that is for solid relationships with distant people who can help in case of local problems; iii) the wider the social network, the higher the degree of linguistic convergence, and therefore the larger the linguistic groups (both in size and spread).

Nettle also investigated the relationship between the social structure of a population and its linguistic features. He tried to model how social factors impact language evolution (Nettle 1999b), and more specifically how the size of a population influences the rate of linguistic change (Nettle 1999c; see also Ringe 2011). In doing so, he paved the way for further studies on the relationship between ecological, social and linguistic factors. In the last decade, several authors have indeed built on Nettle's original geographic and age-divided social structure, relying for example on tools from graph theory (Ke, Gong and Wang 2008).

Jacquesson's advocacy for a 'linguistics of the quasi-desert' provides another example of how a specific living place—deserts—may result in particular linguistic dynamics because of the social relationships between the inhabitants (Jacquesson 2001; 2003). Though based on different approaches, these studies highlight that environmental factors may impact the social structure of speakers, which may in turn weigh on their linguistic structures.

It makes good sense to postulate indirect relationships between the environment and linguistic structures through the mediation of social factors. The alternative of direct relationships is also worth considering and some proposals do exist, such as the acoustic adaptation hypothesis. It states that languages, as other animal vocal communication systems, are adapted to the ecological environment in which they operate. Quoting Maddieson (2011d), 'temperate environments with open vegetation would facilitate transmission of higher frequency signals better than warmer more densely vegetated environments. Hence, languages in the first setting will tend to be more consonant-heavy (more consonant contrasts, more consonant-heavy syllables), whereas those in the second are likely to be more vocalic and to have simpler syllable structures (and perhaps longer words)'. Maddieson (ibid) tested this hypothesis and found a significant correlation between a combined phonological measure of consonant inventory size and complexity of the syllable system, and the proximity to the temperate zones around 45° of latitude north or south.

All these studies call for further investigation of the influence of the environment, whether direct or indirect, on linguistic variables. In the next section, we introduce the material we considered to this end, and explain how we assembled different sources of data into a proper dataset.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

2. Material and Methods

Current studies on linguistic diversity at global scales take advantage of recent scientific advances: i) the availability of high-resolution global maps of biophysical, social, and linguistic data, ii) the development of computational and statistical tools to explore the structure of these data, iii) interdisciplinary efforts to bridge theories and tools from fields such as climatology, ecology, anthropology, geography, or linguistics. Following this trend, we gathered data from a variety of sources and disciplines. We first describe linguistic, then ecological and demographic data. We then report the procedures applied to them in the perspective of statistical tests.

2.1 Linguistic and sociolinguistic data

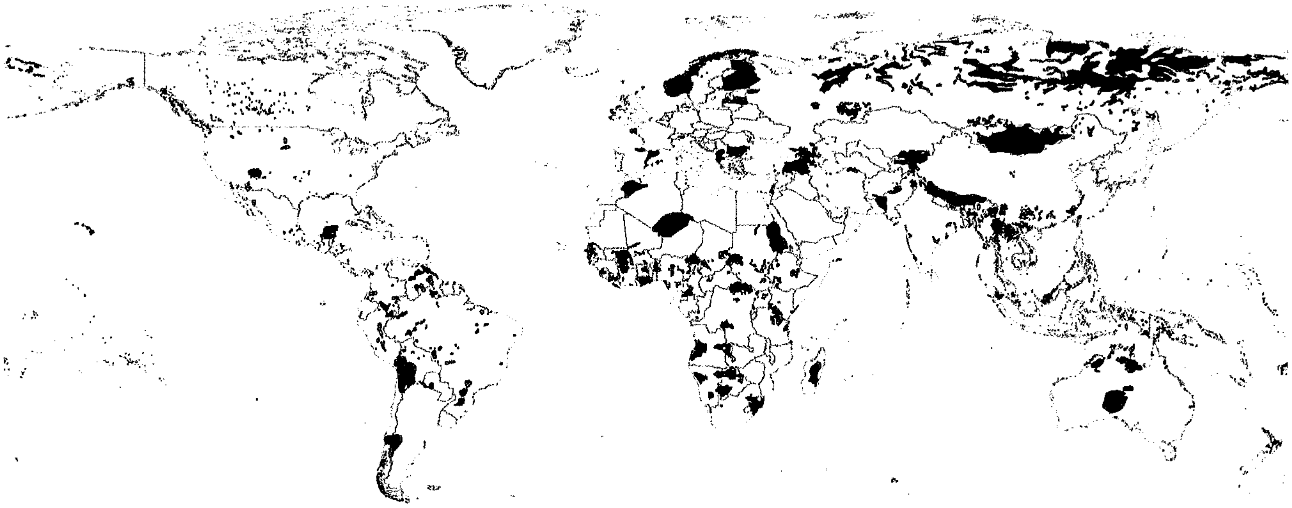
We started from Atkinson (2011a)'s data as they were available in the supplementary material of his publication. These data were extracted from the World Atlas of Linguistic Structure (WALS) (Dryer and Haspelmath 2011) and the Ethnologue (Lewis 2009). The dataset consisted in 504 languages, each of them being provided with geographic coordinates (latitude and longitude), a linguistic classification (at the family and gender levels), its estimated number of speakers and measures of vocalic, tonal and consonantal diversity in terms of numbers of phonemes. These phonological data were compiled by Maddieson in line with his efforts on the UPSID database (Maddieson 1984, 2011a, 2011b, 2011c; Maddieson and Precoda 1990). The three measures of diversity were standardized by Atkinson in order to take their average as a unified measure of phonemic diversity.

Additionally, we relied on the World Language Mapping System (WLMS). This digital dataset provided by Global Mapping International and SIL International (2012) contains geographic areas for the languages described in the Ethnologue at country-level, along with information such as the estimated number of speakers, linguistic family, etc. Such a dataset improves significantly over descriptions of languages as single geographic locations, and is useful to compute average values of environmental variables for the area where a language is spoken.

We merged data to shift from country-level to country-independent descriptions, deleted languages with no speakers, and obtained a dataset of 6,270 languages. In this dataset, we attempted at identifying the 504 previous languages. Most languages were readily found; a dozen corresponded to two languages in the WLMS, and were therefore split. Remaining languages could not be identified, and were therefore removed from the dataset. We furthermore discarded languages with more than 10 million speakers or less than 10 in order to minimize linguistic evolutions—expansions or language deaths—due to recent demographic and economic changes.

The previous processing led to a dataset of 460 languages. For each of them, we relied on the linguistic classification and measures of phonological diversity provided by Atkinson. From the WLMS dataset, we retained the numbers of speakers, and computed language areas with Quantum GIS (version 1.8). With the same software, a measure of linguistic density was

Figure 6.1 Geographic distribution of the 460 languages initially considered in the study



estimated by counting for each language how many other languages were overlapping or spoken less than 50 km from its area.

In the statistical models, the previous variables are described respectively with the codes Family, PhonemeDiv, NbSpeakers, LgArea and LgContact.

Figure 6.1 illustrates the areas of the 460 languages.

2.2 Environmental and demographic data

We took advantage of a number of freely available high-resolution datasets to obtain relevant data for our study.

We considered the WorldClim 1km global climate dataset (version 1.4, release 3) as a source of **climatic data** (Hijmans, Cameron, Parra, Jones and Jarvis 2005). Various variables were available and reminiscent of studies in ecology focusing on the determination of biomes from climatic factors (Holdrige 1947; Whittaker 1975; Kottek, Griser, Beck, Rudolf and Rubel 2006). For the sake of simplicity, we however only extracted the yearly average precipitations and temperatures.

Regarding **elevation**, we relied on the ETOPO 1' global elevation model (Amante and Eakins 2009), and used Quantum GIS to derive a surface rugosity index from elevation data.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Figure 6.2 Percentage of tree cover (light grey: 0% – black: 100%)

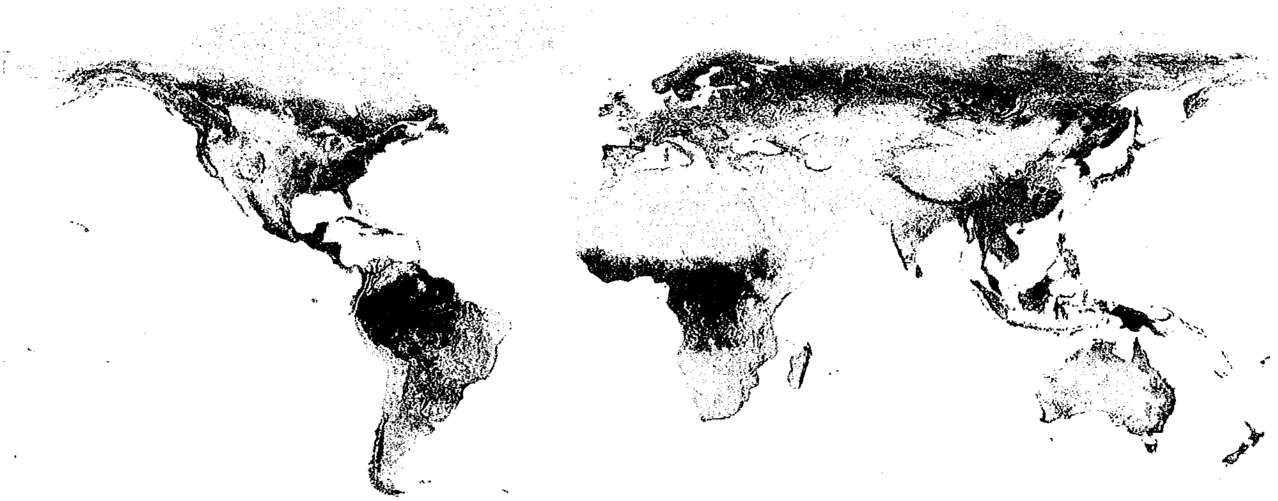
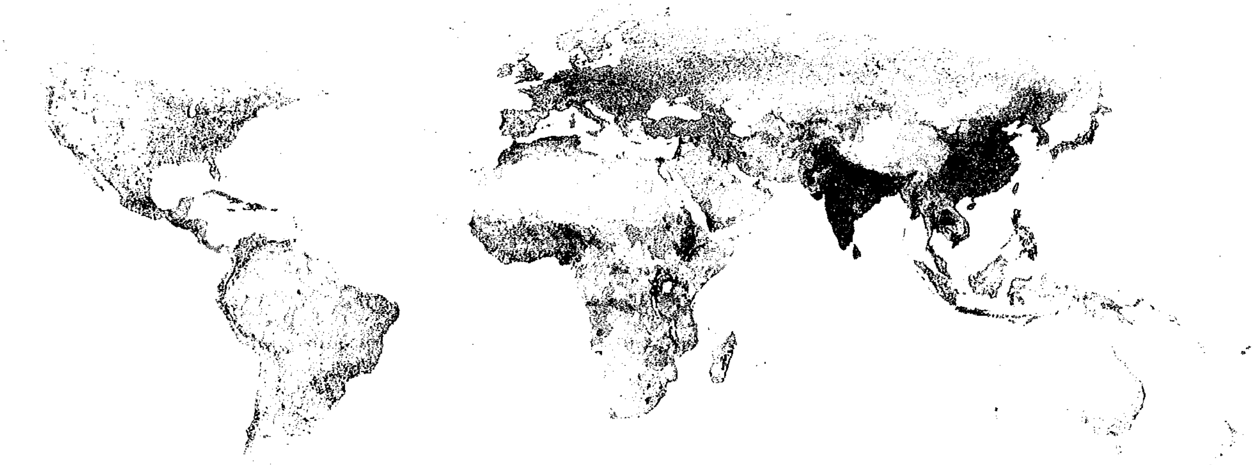


Figure 6.3 Population count (light grey: 1 pers./pixel – black: 36,455 pers./pixel; logarithmic scale)



Two variables related to **vegetation** were considered. The length of growing period (LGP) was obtained as an average 16-class value for the period 1901–1996, at a 5' resolution from the Food and Agriculture Organization and International Institute for Applied Systems Analysis (2000). Tree coverage was obtained as a percentage of ground coverage at a 1km resolution from the Global Map (version 1) dataset (Geospatial Information Authority of Japan, Chiba University and collaborating organizations 2008).

The Global Rural-Urban Mapping Project (version 1) finally provided us with figures of **human population**—number of persons per pixel of dataset—at a 1km resolution for the year 2000 (CIESIN, IFPRI, World Bank and CIAT 2011).

For each language of the dataset described in 2.1, we computed the average values of this set of variables over the area where it is spoken (for population density, we divided the total number of persons by the surface of the linguistic area). Due to lack of data, this was not possible for three languages located on small islands (Fijian, Kiribati, and Rapanui); we therefore removed these three languages from the dataset.

Figure 6.2 depicts the percentage of tree cover at a global scale, and figure 6.3 the logged population count.

In the statistical models, the previous variables are described respectively with the codes Prec, Tmp, Elv, Rug, LGP, Tree, PopDens.

2.3 A priori model of the interactions

A number of studies introduced in section 1 relied on detecting statistically significant correlations and relationships between variables in large sets of numerical or categorical data. Within such a methodological framework, two important issues need to be addressed. The first one is the well-known difference between correlation and causation. A correlation between two phenomena A and B in no way necessarily means a direct causal relationship between them. Alternative options are a third phenomenon C causing both A and B, multiple causal interactions between A, B and other phenomena, or chance. Conclusions regarding causality therefore need to be cautious. The second issue is related to spurious correlations occurring by chance, and so-called Type II errors when crossing many variables with each other in regression models in the hope of detecting significant correlations.

To prevent the previous traps, we defined an a priori model of the possible causal interactions underlying the relationships between our variables. This model divides the variables into four levels, with the key assumption that factors at a given level may influence factors at a higher level, but not vice-versa.

The first level comprises environmental factors, such as Elv, Rug, Tmp, Prec, LGP and Tree in our study, and could include additional ones such as distance to freshwater, biomes, etc.

The second level contains phenomena related to human activities yet independent from the

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

spoken languages. We only considered PopDens in this study, but age distribution, kin systems, etc. would fall into this category. The Out of Africa migration and its consequences also do.

The third level refers to languages with respect to their social aspects, but not their internal structures. It includes their geographical distribution, the social organization or history of speakers, but not features—whether phonological, morphological, syntactic, etc. NbSpeakers, LgArea and LgContact fall into this level.

Finally, the fourth level consists of variables describing the internal states of the language, its structures, and intrinsic complexity. Here we considered PhonemeDiv and Family, but any measure of complexity, the absence or presence of specific linguistic devices etc. could be considered.

Our assumption is that environmental variables (Level 1) are not determined by either social (Level 2), sociolinguistic (Level 3) or purely linguistic (Level 4) variables; non-linguistic social variables (Level 2) are not determined by sociolinguistic (Level 3) or purely linguistic (Level 4) variables. Importantly, we assume numbers of speakers, linguistic contacts and language areas (Level 3) to be independent of genetic relationships between languages (Level 4).

Table 6.1 summarizes the four levels and how the variables of our study fall into them.

This model can be contested on the basis that nowadays, human activities bear an impact on climate and other environmental phenomena; in some cases, linguistic facts weigh on social facts, such as in case of conflicts when spoken languages are politically utilized to displace populations, etc. However, in the vast arena of causal relationships to be unearthed, our approach reflects the assumption that some causal relationships are weaker than others, if not negligible, and can be discarded to better focus on the main phenomena.

Table 6.1 Categories of variables defined for statistical analysis

Level	Types of variables	Variables
1	Environmental	Elv, Rug, Tmp, Prec, LGP, Tree
2	Non-linguistic social	PopDens
3	Sociolinguistic	LgArea, LgContact, NbSpeakers, geographic locations of languages
4	Purely linguistic	PhonemeDiv, Family

2.4 Statistical processing of the data

We applied various statistical methods to our dataset of 457 languages, using the software R (version 3.0.0) and additional packages dedicated to statistical processing.

Some of the 12 variables (Prec, Tmp, Elv, Rug, LGP, Tree, PopDens, Family, PhonemeDiv, NbSpeakers, LgArea and LgContact) yielded distributions very distant from normality (PopDens, NbSpeakers and LgArea), and none of them passed the Shapiro-Wilk normality test. For each variable, we therefore estimated the best parameters for a Box-Cox transformation (Box and Cox 1964), relying on the boxcoxfit function of the geoR package (Ribeiro and Diggle 2001) and the boxcox function of the MASS package (Venables and Ripley 2002). Normality tests were still significant at .05 after transformations for all variables but LgArea. However, all distributions except Tree and LGP were much closer to normality and therefore potentially better adapted to linear regression.

Our results are all based on the application of linear regression models to various subsets of variables. Given the nature of our data, we considered simple fixed effects regressions as well as random effects and spatial regression.

In each caes, we started with a standard linear regression of a predicted (dependent) variable against predictors (independent variables)—e.g., predicting NbSpeakers against the predictors PopDens, Tree, and Elv. We looked in the outputs of the model for outliers (normalized residuals, i.e. predictions errors, higher than 2.5 or 3 in absolute value) with high leverage thanks to influence plots (which simultaneously display for each entry of the model the value of the studentized residual, the hat value, and Cook's distance). When such entries were detected, they were removed from the dataset before running the model once again. As explained below, we relied in some cases on stepwise regression with the stepAIC function of the MASS package, which performed model selection by Akaike information criterion (AIC) to identify the 'best' set of predictors for a predicted variable.

Following among others (Jaeger, Graff, Croft and Pontillo 2011), we tested the assumptions behind the application and validity of such a model, namely:

- The linearity of the relationships between the dependent and independent variables—with Ceres plots
- The normality of the error distribution—with quantile-quantile plots and Shapiro-Wilk tests
- The homoscedasticity of the error distribution, i.e. the absence of correlation with predictors—with plots of residuals versus fitted values and Breusch-Pagan tests
- The absence of strong multicollinearity between the predictors—with Variance Inflation Factors (VIF) and condition indexes (VIF should be lower than 5 for all predictors, and the largest condition index below 30)

In our experiments, we did not often end with residuals satisfying tests like the Shapiro-Wilk or Breusch-Pagan tests. This was however not surprising since we had several hundreds of observations. The various graphics always suggested near normality and good homoscedasticity.

Another important assumption in linear regression is the assumption of independence of the observations from each other. In time-series for example, observations at time t partly influenced by earlier observations violate this assumption. This leads to autocorrelation in the residuals, and failure to take this issue into account weakens the predictive quality of the model.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

In the case of languages, spatial proximities and their possible consequences as well as genetic relatedness violate the assumption of independence. Problems then lie in the possible differences between subgroups of languages, potentially generating artifacts such as Simpson's paradox (Simpson 1951). The issue was addressed by Hauer and Bauer (2007) by looking at the effect of the number of speakers at the family level, and also by considering family as a fixed effect. Atkinson also relied on a by-group approach, but replaced the former fixed effect by a more adapted random effect. Jaeger et al. (2011) stressed the relevance of mixed effects models for typological study, but also showed that the distribution of languages into families was too sparse to fully sort out the problem of genetic groupings in Atkinson's study.

In our analysis, we relied on two approaches to control for the spatial and genetic relationships between languages: in cases where we predicted variables of the 3rd level of our model—LgArea, LgContact, NbSpeakers—we assumed that genetic grouping was not significant, and only considered spatial relationships with spatial regression models. Two languages distant by less than 1,000 km were considered as neighbors in the definition of spatial weights. Following (Anselin 2005, 2007; Anselin, Syabri and Kho 2006), we applied Lagrange multiplier diagnostics for spatial dependence to our linear regressions to detect spatial autocorrelation and identify which model—spatial error or spatial lag model—could be applied to take it best into account.[3] The application of the correct model led to more robust regression coefficients and p-values. These p-values were computed, along with confidence intervals, with Markov Chain Monte Carlo (MCMC) approaches. More specifically, we used the pvalsfunc function of the spdep package.

In cases where we predicted the variable at the 4th level of our model—PhonemeDiv, we compared models with and without the categorical variable Family introduced as a random effect. This once again led to a better assessment of the various predictors.

As further explained in section 3, our analysis rested on the hierarchy between variables defined in our a priori model of interactions (see section 2.3). While we investigated predictions of variables at level 3 or 4 against variables of lower levels, we wished to consider the influence of factors at a given level while controlling for the influence of other factors. To this end, we inspected several definitions of 'sums of squares' in the regression (Myers, Montgomery, Vining and Robinson 2010), and the related Type I and Type III analyses. A Type III analysis evaluates the relationship between an independent variable and the dependent variable given that all other independent variables are included in the model. A Type I analysis is based on a hierarchical (or sequential) decomposition: whether an independent variable predicts a significant part of

[3] Lagrange multiplier diagnostics are usually analyzed as follows: non-robust versions of the statistics—LMerr and LMlag—are considered first. If LMerr is not significant and LMLag is, a spatial lag model should be considered to account for spatial autocorrelation. If LMerr is significant but LMLag isn't, a spatial error model is recommended. If, and only if, both non-robust statistics are significant, robust versions—RLMerr and RLMlag—are analyzed in the same way as non-robust statistics. If both robust tests are significant, the spatial model is chosen according to the most significant one.

the variance is computed given that all the independent variables listed above in the model are included, but not the variables listed below. To refer to our specific variables, we could thus estimate the influence of environmental variables on sociolinguistic variables with or without controlling for population density in the model, or the potential impact of tree coverage on phonemic diversity controlling for sociolinguistic factors.

The following conventions were adopted to present the results:

- Regarding levels of significance, '***' stands for p<0.001, '**' for p<0.01, '*' for p<0.05, '.' for p<0.1, and 'n.s.' for non-significant
- S-W stands for the Shapiro-Wilk test, B-P for the Breusch-Pagan test, Cond. for the largest condition index, VIF for the largest variance inflation factor. LM test stands for the LM test for residual autocorrelation in the spatial regression model.
- For the Shapiro-Wilk test, the Breusch-Pagan tests, Lagrange multiplier diagnostics and LM tests, only p-values are given.

3. Results

We summarize here the different models which were applied to our data, following the framework defined in section 2. Section 3.1 offers preliminary observations on the basis of graphical representations and coefficients of correlation between the variables. Section 3.2 revisits Nettle's proposal regarding LGP at a global scale, first without, then with, controlling for population density. Section 3.3 summarizes stepwise regressions for the three sociolinguistic variables NbSpeakers, LgArea and LgContact. Section 3.4 gives predictions of phonemic diversity against successively sociolinguistic variables, sociolinguistic variables and tree coverage, and sociolinguistic and environmental variables. Finally, section 3.5 reports investigations of Atkinson's hypothesis of a gradient of phonemic diversity from the origin of modern human populations.

3.1 Preliminary observations

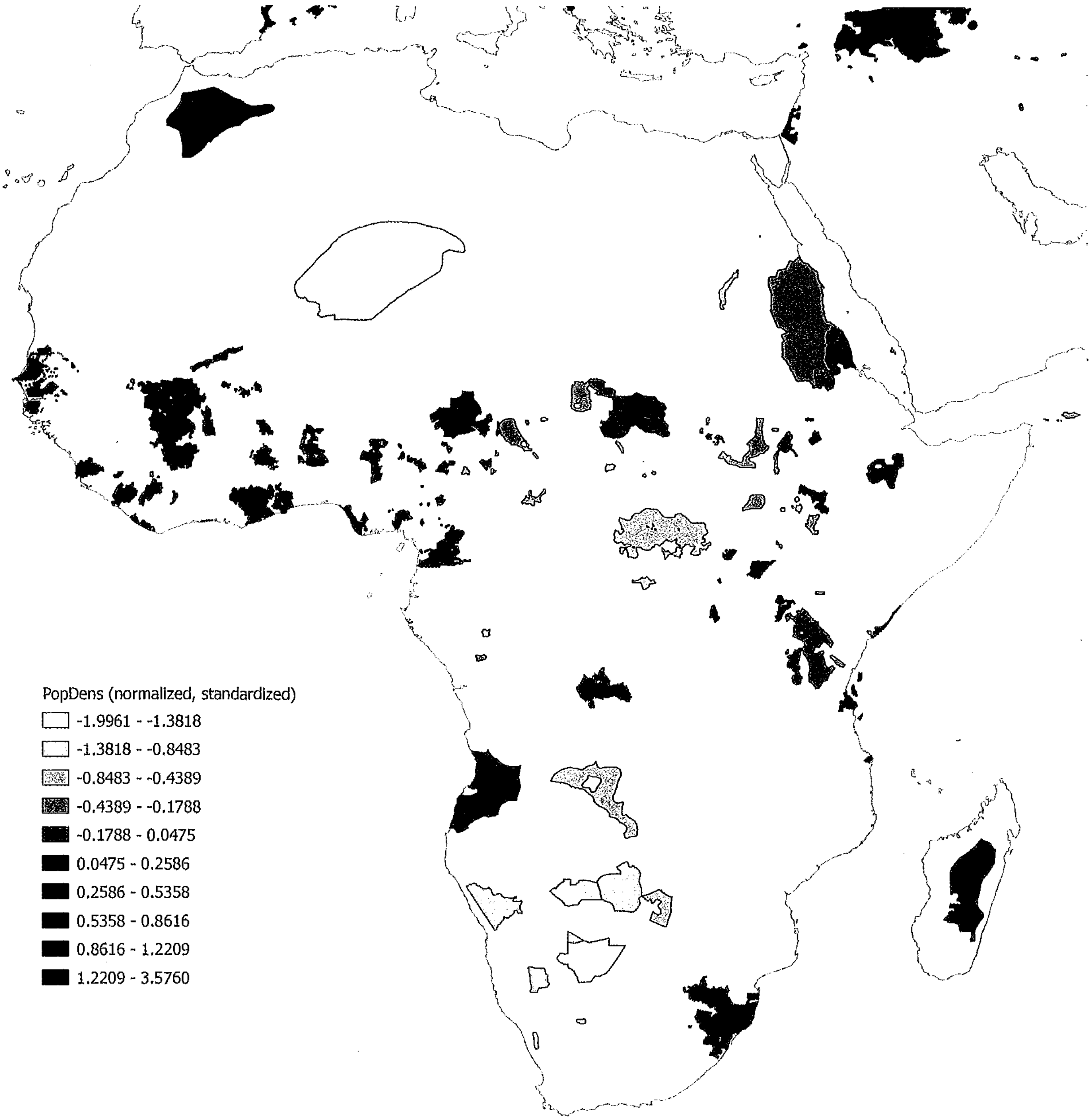
In order to graphically illustrate data distribution, we propose to look at African languages. Figures 6.4, 6.5 and 6.6 respectively display normalized and standardized PopDens, NbSpeakers and LgContact variables for the African languages of our dataset. Language areas can easily be seen on each map. The maps allow large scale patterns to be detected, and similarities and differences between languages to be noticed. For example, Figure 6.6 illustrates the high number of contacts for languages in the sub-Saharan region.

In addition to maps, coefficients of correlation between the variables can be computed and analyzed. Tables 6.2, 6.3, 6.4 and 6.5 provide Pearson's product-moment correlations for the variables, grouped according to the levels of our model.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

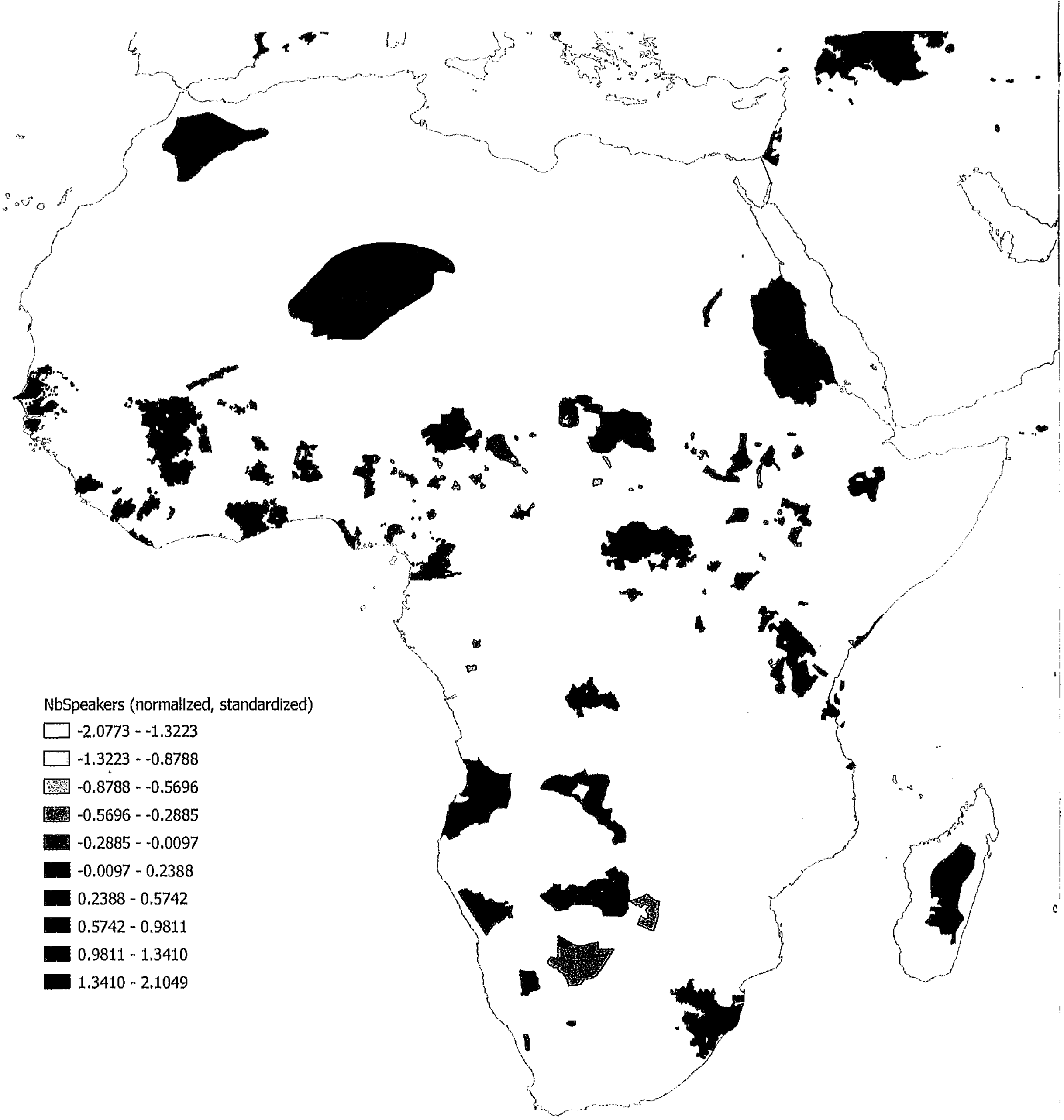
Coupé, C., Hombert, J.-M., Marsico, E. & Pellegrino, F. (2013)

Figure 6.4 Density of population for the African languages of the dataset



in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Figure 6.5 Number of speakers for the African languages of the dataset.



in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Figure 6.6 Language contact for the African languages of the dataset.

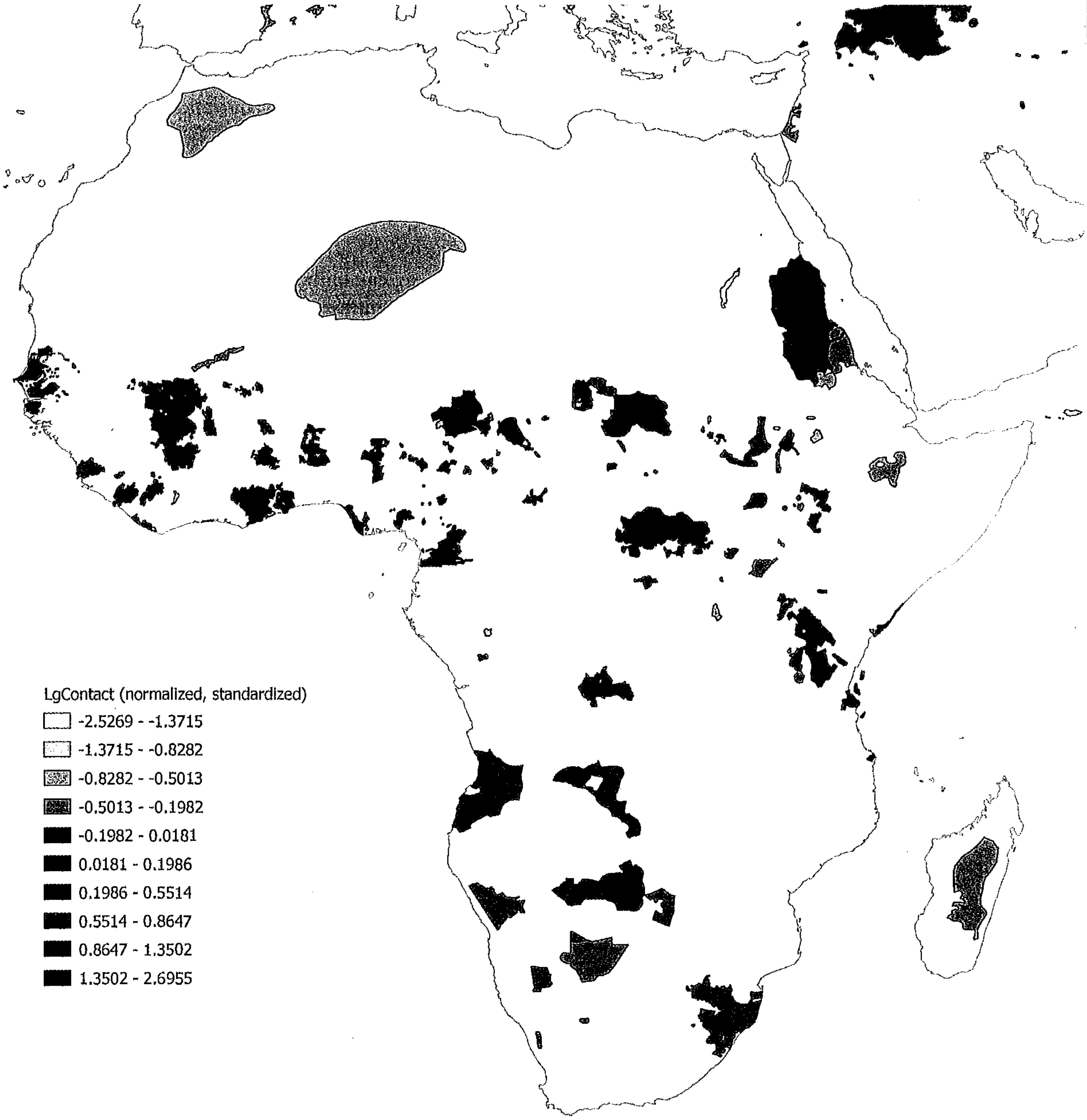


Table 6.2 indicates that environmental variables correlate with each other, positively or negatively, to a large extent. This was expected, as it is for example well known that rugosity increases and temperatures decrease as elevation increases. Prec, Tree and LGP correlate especially strongly and positively with each other; while this could also be expected—the more rain, the easier for plants to grow and the denser the tree coverage, correlations could have been weaker if one thinks of trees requiring only little water, or of heavy rains occurring yearly only during a few months. This suggests that Prec, Tree and LGP may play a similar role in a regression model, and should perhaps not be considered together to prevent strong multicollinearity.

In Table 6.3, PopDens correlates significantly and positively with Elv and Rug. This is surprising since we usually think of high and uneven lands as more difficult to inhabit, and since most of human population lives along the coast. However, since variables were normalized, the correlation may come from relationships primarily occurring at low altitudes. PopDens also correlates positively with Prec and LGP, which suggests higher densities of population in areas where plants grow easily, either for pastoralism, agriculture or gathering. NbSpeakers, LgArea and LgContact correlate with environmental factors, but one should wonder whether potential causal relationships are direct or indirect: if Rug for example causally impacts on LgContact, is it because rougher lands lead to denser populations, which in turn leads to more contacts, or because of another more direct causal mechanism?

Sociolinguistic variables also correlate with each other to a significant extent. Finally, PhonemeDiv correlates with most variables, suggesting a complex pattern of relationships and causal effects that require careful investigation.

Table 6.2 Pearson's correlations for environmental variables

	Tmp	Prec	Elv	Rug	Tree
Prec	.36***				
Elv	-.46***	-.22***			
Rug	-.46***	.15***	.67***		
Tree	.12*	.79***	-.15**	.21***	
LGP	.27***	.86***	-.16***	.18***	.79***

Table 6.3 Pearson's correlations between environmental variables and social and sociolinguistic variables

	PopDens	NbSpeakers	LgArea	LgContact
Tmp	.04 n.s.	-.02 n.s.	-.14**	.25***
Prec	.12*	-.15**	-.33***	.22***
Elv	.19***	.28***	.13**	.21***
Rug	.27***	.11*	-.16***	.13**
Tree	.01 n.s.	-.21***	-.30***	.16***
LGP	.14**	-.08.	-.30***	.25***

Table 6.4 Pearson's correlations for social and sociolinguistic variables

	PopDens	NbSpeakers	LgArea
NbSpeakers	.56***		
LgArea	-.17***	.52***	
LgContact	.28***	.41***	.08 n.s.

Table 6.5 Pearson's correlations between PhonemeDiv and other variables

	PhonemeDiv
Tmp	.04 n.s.
Prec	-.14**
Elv	.20***
Rug	-.02 n.s.
Tree	-.11*
LGP	-.12**
PopDens	.27***
NbSpeakers	.37***
LgArea	.08 n.s.
LgContact	.33***

3.2 Influence of the length of the growing period on sociolinguistic variables

In his approach to the linguistic diversity of Western Africa, Nettle (1996) considered the length of the growing period of plants (LGP) and the altitude to predict language areas and numbers of speakers.

We investigated whether Nettle's results could be reproduced at a global scale. We first ran regression models with Elv and LGP as predictors, and NbSpeakers, LgArea or LgContact as predicted variables.

From the three models, although the percentage of variance explained was low (around 7.5—12.5%), the following robust conclusions could be made:

- The higher, the more speakers, the larger the language areas and the more linguistic contacts;
- The longer the LGP, the smaller the language areas and the more linguistic contacts.

These results confirmed Nettle's hypothesis that when the ecological risk is high, language areas tend to grow. However, LGP did not have an impact on NbSpeakers as it had in Nettle's study. While this was potentially explained by different approaches to compute this figure, a positive correlation was nevertheless expected. Regarding language contacts, fewer contacts in case of higher ecological risk makes sense as a result of higher linguistic convergence.

The previous issue regarding LGP and NbSpeakers may be considered at the light of our a priori model and of the two levels of factors that may have an impact on sociolinguistic variables. In the former models as in Nettle's study, population density was not considered as a significant factor. It makes sense however to think of the impact of environmental factors with population density being taken into account, since significant correlations exist between PopDens and the sociolinguistic variables. Letting aside the effect of PopDens may thus prevent other effects to be visible.

To check this assumption, we ran three new regression models, with PopDens, Elv and LGP as predictors, and NbSpeakers, LgArea and LgContact as predicted variables. Tables 6.6, 6.7 and 6.8 provide the results.

Table 6.6 Regression of NbSpeakers against PopDens, Elv and LGP

Predicted variable: NbSpeakers									
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression			
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>ltl)	Estimate	Std. Error	Pr(>lzl)	
(Intercept)			4.95E-17	.038	1. n.s.	.007	.072	.921 n.s.	
PopDens	142.64	4.56E-41 ***	.549	.039	9.07E-38 ***	.444	.046	<2.2E-16 ***	
Elv	13.68	5.26E-06 ***	.151	.039	1.26E-04 ***	.132	.038	5.19E-04 ***	
LGP	7.93	4.93E-04 ***	-.136	.039	4.93E-04 ***	-.146	.048	.002 **	
Residuals	291.75								
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error			
.356	1087	5.23E-04	2.26E-06	1.34	1.08	AIC: 992			
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: 7.25E-04			
		<2.2e-16	2.58E-07	<2.2e-16	.002				
Deleted outliers: Squamish									

Table 6.7 Regression of LgArea against PopDens, Elv and LGP

Predicted variable: LgArea								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			-3.73E-16	.044	1. n.s.	.075	.082	.361 n.s.
PopDens	13.38	1.17E-04 ***	-.158	.046	5.79E-04 ***	-.104	.054	.055 .
Elv	12.79	1.64E-04 ***	.122	.046	.008 **	.118	.045	.009 **
LGP	28.54	2.47E-08 ***	-.257	.045	2.47E-08 ***	-.281	.061	4.62E-06 ***
Residuals	401.29							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error		
.114	1233	.016	.159	1.361	1.089	AIC: 1146		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .197		
		<2.2e-16	.002	<2.2e-16	.309			
Deleted outliers: Vanimo, Yapese								

Table 6.8 Regression of LgContact against PopDens, Elv and LGP

Predicted variable: LgContact								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			3.94E-16	.043	1. n.s.	-.777	.121	1.14E-10 ***
PopDens	36.00	1.64E-10 ***	.205	.044	5.51E-06 ***	.026	.049	.593 n.s.
Elv	11.51	2.42E-04 ***	.21	.045	3.32E-06 ***	.177	.039	5.53E-06 ***
LGP	27.51	1.94E-08 ***	.253	.044	1.94E-08 ***	.15	.057	.008 **
Residuals	380.97							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error		
.159	1210	.279	.004	1.353	1.085	AIC: 1031		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: 8.37E-05		
		<2.2e-16	5.25E-04	<2.2e-16	.003			
Deleted outliers:Tibetan, Yapese								

Adding PopDens as a predictor increased the linearity of the relationships between the predictors and the predicted variable (not shown). The percentage of variance also increased in all three cases.

Even with PopDens included in the model, Elv and LGP explain a significant percentage of the variance in the standard regression. Additionally, they significantly predict all three predicted variables when spatial relationships between languages are accounted for. PopDens only predicts NbSpeakers when spatial relationships are accounted for.

The updated conclusions may therefore be suggested:

- The higher the population density, the more speakers
- The higher the elevation, the more speakers, the larger the language areas and the more linguistic contacts
- The longer the LGP, the less speakers, the smaller the language areas and the more contacts

Nettle's proposal regarding the role of ecological risk in sub-Saharan Africa therefore seems to be valid at a larger scale. The effect of PopDens on NbSpeakers also makes sense, while its independence from LgContact and LgArea is not surprising. The role of elevation is more difficult to explain. The relationship between LGP and LgContact suggests that an increase in ecological risk leads to reduced linguistic density and contacts, something which was not necessarily implied by an increase in numbers of speakers and language areas. More data on multilingualism would here be necessary to investigate the situation in more details.

3.3 Stepwise models for the regression of sociolinguistic variables against environmental and social factors

To further investigate the impact of environmental factors on sociolinguistic variables while controlling for population density, we ran stepwise regression models with PopDens as a fixed predictor in the regression. Tables 6.9, 6.10 and 6.11 give the results for NbSpeakers, LgArea and LgContact respectively.

We do not find LGP selected as a factor to predict NbSpeakers. However, Tree appears as a relevant factor and is highly positively correlated with LGP. It may thus stands for the previous effect of ecological risk. Both Elv and PopDens appear to have a strong impact on NbSpeakers, and rugosity is also selected in the model: other things being equal, the more rugosity the fewer speakers per language. It is interesting to note here that without controlling for other variables, rugosity was positively correlated, although not very strongly, with NbSpeakers. Once again, including other factors in the model modify the relationships between variables, which partially hinder analyses. The previous effect is complemented by the effect of rugosity as a predictor of LgArea: other things being equal, the rougher the ground, the smaller the areas of linguistic groups. We may postulate that on rough grounds, with increased difficulties to move across the land, linguistic groups are smaller both in terms of number of speakers and area.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Coupé, C., Hombert, J.-M., Marsico, E. & Pellegrino, F. (2013)

Table 6.9 Stepwise regression of NbSpeakers against PopDens and environmental variables

Predicted variable: NbSpeakers								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			7.23E-17	.036	1. n.s.	.024	.066	.718 n.s.
PopDens	142.64	6.10E-43 ***	.56	.038	1.61E-40 ***	.468	.045	<2.2.e-16***
Elv	13.68	2.83E-06 ***	.293	.054	9.49E-08 ***	.282	.059	1.83E-06 ***
Rug	19.2	3.38E-08 ***	-.21	.056	1.80E-04 ***	-.215	.062	5.49E-04 ***
Tree	5.56	.003 **	-.124	.041	.003 **	-.106	.044	.017 *
Residuals	274.93							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error		
.392	1042	.03	.006	2.772	2.332	AIC: 968		
Lagrange multiplier diagnostics		**LMerr**	RLMerr	LMlag	RLMlag	B-P: .005		
		<2.2e-16	1.67E-05	<2.2e-16	3.45E-04			
Deleted outliers: Comanche, Maricopa, Squamish								

Table 6.10 Stepwise regression of LgArea against PopDens and environmental variables

Predicted variable: LgArea								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			-3.50E-16	.042	1. n.s.	.088	.08	.272 n.s.
PopDens	13.38	6.50E-05 ***	-.129	.045	.005 **	-.122	.052	.019 *
Elv	12.79	9.36E-05 ***	.31	.063	1.40E-06 ***	.392	.075	1.64E-07 ***
Rug	40.1	1.07E-11 ***	-.356	.068	2.97E-07 ***	-.383	.076	4.08E-07 ***
Tmp	9.07	9.75E-04 ***	-.133	.051	.009 **	-.058	.075	.441 n.s.
Tree	9.36	8.09E-04 ***	-.164	.049	8.09E-04 ***	-.109	.053	.039 *
Residuals	371.29							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error		
.177	1209	.015	.001	3.093	2.642	AIC: 1131		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .141		
		<2.2e-16	8.72E-04	<2.2e-16	.136			
Deleted outliers: Ani								

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Table 6.11 Stepwise regression of LgContact against PopDens and environmental variables

Predicted variable: Lgcontact								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			4.41E-16	.04	1. n.s.	-.382	.099	1.18E-04 ***
PopDens	36.00	1.33E-11 ***	.169	.042	7.24E-05 ***	.047	.047	.316 n.s.
Tmp	25.16	1.21E-08 ***	.359	.047	1.33E-13 ***	.438	.077	1.39E-08 ***
Elv	43.09	1.75E-13 ***	.371	.047	2.22E-14 ***	.372	.051	3.89E-13 ***
LGP	14.32	1.48E-05 ***	.186	.043	1.48E-05 ***	.132	.055	.016 *
Residuals	337.42							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial error		
.253	1144	.249	9.02E-04	1.812	1.343	AIC: 999		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: 1.99E-04		
		<2.2e-16	1.90E-09	<2.2e-16	.641			
Deleted outliers: Soqotri, Tibetan, Yapese								

A dense tree cover also negatively impacts on the area of languages. This once again may be related to ecological risk. PopDens also negatively impacts on LgArea: the higher the population density, the smaller the language areas.

Finally, as in 3.2, PopDens does not influence LgContact when spatial proximities between languages are taken into account, while Tmp, Elv and LGP all significantly predict the number of contacts with positive Type III coefficients.

Disentangling the various effects is difficult given the correlations between the selected variables. Exploring the data with stepwise regression is more hazardous than testing whether some models support hypotheses made a priori, like the role of ecological risk. All in all and on safe grounds, it may be noticed that nearly 40% of the variance of NbSpeakers is explained by PopDens, Elv, Rug and Tree. Although the two other models have weaker R^2, they also support the idea that local environmental factors should not be forgotten when considering sociolinguistic factors, even if these factors may be predominantly determined by non-linguistic social factors like population density.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

3.4 Predictions of phonemic diversity by environmental, social and sociolinguistic factors

If we now turn to PhonemeDiv and the 4th level of our explanatory model, we may first investigate the effects of sociolinguistic variables. Table 6.12 summarizes a regression model with PhonemeDiv as predicted variable and NbSpeakers, LgArea and LgContact as predictors, while Table 6.13 gives the result of a non-spatial regression with Family as random effect.

Table 6.12 Regression of PhonemeDiv against sociolinguistic variables

Predicted variable: PhonemeDiv								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			.014	.027	.619 n.s.	-.014	.023	.558 n.s.
NbSpeakers	26.08	4.23E-17 ***	.228	.035	3.37E-10 ***	.084	.031	.007 **
LgArea	3.53	.001 **	-.079	.032	.015 *	-.04	.028	.158 n.s.
LgContact	5.99	3.32E-05 ***	.128	.03	3.32E-05 ***	.091	.026	5.62E-04 ***
Residuals	154.31							
Adj. R^2	AIC	S-W	B-P	Cond.	VIF	Model: spatial lag		
.182	806	.024	.037	2.132	1.689	AIC: 690		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .714		
		<2.2e-16	.926	<2.2e-16	5.99E-09	LM test: 1.21E-06		
Deleted outliers: Usan								

Table 6.13 Regression of PhonemeDiv against sociolinguistic variables with Family as random effect

Predicted variable: PhonemeDiv			
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-.143	.050	.004 *
NbSpeakers	.034	.040	.392 n.s.
LgArea	-.042	.032	.194 n.s.
LgContact	-.074	.030	.013 *
		Std. Dev.	
Family		.357	
Residuals		.455	
cor(fitted, predicted)^2: 0.573, AIC: 701			
Deleted outliers: Usan			

While spatial error models were predominant in earlier analyses, Lagrange multiplier diagnostics suggest a spatial lag model to account for the geographic proximities between languages. Including a Family random effect to account for the relationships between languages suggest that the effects of NbSpeakers and LgArea on phoneme diversity are rather problematic, since they lose their significance when Family is added. While LgArea is not a significant predictor in the spatial regression, NbSpeakers is; it is therefore difficult to conclude whether the random effect masks a real effect of NbSpeakers on PhonemeDiv, or whether this effect is an artifact when the groupings of languages into families are not considered. This incertitude is not surprising given the debates mentioned in section 1.1.

Interestingly, LgContact maintains a significant effect as predictor even when spatial or genetic relationships are considered. The idea that more linguistic contacts lead to higher phonemic diversity through borrowing from neighboring languages is therefore supported by our statistical tools.

Adding PopDens to sociolinguistic factors does not significantly increase the quality of the regression, and we therefore discard this factor in subsequent models.

Building on the previous conclusions regarding sociolinguistic and social variables, we also tested the acoustic adaptation hypothesis by considering the variable Tree as a predictor of both consonant diversity—ConsDiv – and vowel diversity—VowelDiv—, along with sociolinguistic variables.

ConsDiv and VowelDiv are defined by a limited range of values: five values for ConsDiv, which relate to the five categories 'Small', 'Moderately Small', 'Average', 'Moderately Large' and 'Large' for consonant inventories in the WALS, and three values for VowelDiv, which relate to the three categories 'Small', 'Average' and 'Large' used to classify vocalic inventories. Because of these specific distributions, linear regression is a rather poor choice of statistical model. To address this issue, we took advantage of the UPSID database, which contains the numbers of vowels and consonants of 451 languages overlapping the languages of our dataset. We identified the intersection between the two sets and built a new database of 319 languages, for which we not only had diversity figures but also and more accurately the numbers of consonants—ConsNb – and vowels—VowelNb. We then applied regression models to predict, on the one hand, ConsDiv and VowelDiv with our initial dataset of 457 languages and, on the other hand, ConsNb and VowelNb with our second dataset of 319 languages. Each time, the predictors were the sociolinguistic variables and Tree.

Predictions of vowel diversity are inconclusive with respect to the effect of Tree whatever the model or dataset used. Predictions of either ConsDiv or ConsNb prove more interesting, with a very significant effect of Tree in the different models. Tables 6.14 and 6.15 give the details of the regressions for the prediction of ConsNb. In this model, a barely significant positive effect of NbSpeakers is observed on the number of consonants (again, with other predictors included in the model) when Family is included as a random effect.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Table 6.14 Regression of ConsNb against NbSpeakers, LgArea, Lgcontact and Tree on the basis of a subset of 319 languages

Predicted variable: ConsNb								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)			-6.63E-13	.052	1. n.s.	-.02	.044	.65 n.s.
NbSpeakers	13.54	8.47E-05 ***	.276	.072	1.45E-04 ***	.17	.061	.006 **
LgArea	1.15	.246 n.s.	-.173	.063	.007 **	-.143	.054	.008 **
LgContact	10.18	6.29E-04 ***	-.118	.062	.057 .	-.056	.052	.278 n.s.
Tree	25.03	1.22E-07 ***	-.307	.057	1.22E-07 ***	-.198	.049	5.89E-05 ***
Residuals	268.09							
Adj. R²	AIC	S-W	B-P	Cond.	VIF	Model: spatial lag		
.146	831	.087	.002	2.364	1.912	AIC: 766		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .003		
		<2.2e-16	.9	<2.2e-16	1.46E-04	LM test: .037		
Deleted outliers: Archi, Juhoan, NorthernHaida								

Table 6.15 Regression of ConsNb on the basis of a subset of 319 languages with Family as a random effect

Predicted variable: ConsNb			
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-.006	.108	.954 n.s.
NbSpeakers	.156	.078	.0454 *
LgArea	-.140	.059	.0189 *
LgContact	-.0121	.055	.823 n.s.
Tree	-0.148	.054	.0068 **
		Std. Dev.	
Family		.829	
Residuals		.637	
cor(fitted, predicted)²: 0.673, AIC: 768			
Deleted outliers: Archi, Juhoan, Northern Haida			

The following conclusions may be drawn from the previous regressions:

- The more contacts between languages, the higher the phonemic diversity;
- The less dense the tree coverage, the higher the number of consonants.

The first conclusion is in line with earlier comments on phonemic diversity. The second conclusion is in agreement with Maddieson (2011d)'s hypotheses and findings (see section 1.2). Our models support the idea that in densely vegetated environments, an impeded transmission of higher frequency acoustic signals leads to a reduced number of consonants in languages.

Finally, for explanatory purposes, we ran a stepwise regression of PhonemeDiv against sociolinguistic and environmental variables. Tables 6.16 and 6.17 summarize the results.

Table 6.16 Stepwise regression of PhonemeDiv against sociolinguistic and environmental variables

Predicted variable: PhonemeDiv								
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression		
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>Itl)	Estimate	Std. Error	Pr(>Izl)
(Intercept)			.014	.026	.601 n.s.	-.015	.023	.509 n.s.
NbSpeakers	26.08	1.49E-18 ***	.25	.035	6.93E-12 ***	.107	.032	8.20E-04 ***
LgArea	3.53	7.87E-04 ***	-.161	.033	1.77E-06 ***	-.078	.03	.009 **
LgContact	5.99	1.34E-05 ***	.138	.031	1.07E-05 ***	.094	.027	5.36E-04 ***
LGP	7.45	1.26E-06 ***	-.174	.045	1.35E-04 ***	-.131	.039	8.79E-04 ***
Rug	1.49	.029 *	-.195	.041	2.40E-06 ***	-.103	.036	.004 **
Elv	4.6	1.31E-04 ***	.173	.041	3.12E-05 ***	.113	.036	.002 **
Tree	2.06	.01 *	.117	.045	.01 *	.108	.04	.006 **
Residuals	138.72							
Adj. R²	AIC	S-W	B-P	Cond.	VIF	Model: spatial lag		
.258	764	.12	.071	3.525	3.032	AIC: 671		
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .188		
		.	.451	.	1.56E-11	LM test: 5.57E-06		
Deleted outliers: Ndut								

Table 6.17 Stepwise regression of PhonemeDiv against sociolinguistic and environmental variables, with Family as a random effect

Predicted variable: PhonemeDiv			
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-.141	.048	.003 **
NbSpeakers	.055	.041	.178 n.s.
LgArea	-.063	.034	.064 .
LgContact	.076	.031	.014 *
LGP	-.108	.046	.018 *
Rug	-.081	.043	.059 .
Elv	.088	.041	.032 *
Tree	.104	.043	.016 *
		Std. Dev.	
Family		.334	
Residuals		.456	
cor(fitted, predicted)²: 0.569, AIC: 718			
Deleted outliers: Ndut			

Several environmental variables significantly increase the quality of the model even with sociolinguistic variables included in the model. LGP, Elv and Tree preserve their effect both in the spatial regression and when Family is included, along with LgContact. One should not be surprised that the coefficient for Tree is positive in the three regressions, although this seems contradictory to the previous results regarding the acoustic adaptation hypothesis. The discrepancy can indeed be explained by the inclusion of LGP in the model, and the strong positive correlation between this factor and Tree. Once again, a Type III coefficient reflects the inclusion of all factors in the model, and the positive impact of Tree may be seen as a partial counterweight to the negative impact of LGP. This again stresses the difficulty of interpreting outputs of stepwise regressions.

While it is difficult to think of an explanation for the impact of LGP and Elv, we reach the more general conclusion that environmental factors play a role in phonetic/phonological processes; the acoustic adaptation hypothesis is one of possibly several causal patterns.

3.5 Phonemic diversity at the light of local factors and large-scale migrations

Our last attempt consisted in reconsidering Atkinson's global causal effect—the Out of Africa migration—on phonemic diversity at the light of our previous results. More specifically, we wondered whether the effect of the distance from a potential origin of the migrations would still be preserved once local factors had been included in a regression model. We therefore started from our last stepwise regression of PhonemeDiv, and added the distance from the most likely origin point found by Atkinson.

As illustrated by Tables 6.18 and 6.19, the distance from Atkinson's most likely origin point is a strongly significant factor. Its effect is preserved in spatial regression and with the inclusion of Family as random effect. The hypothesis that Atkinson's proposal regarding a phonemic gradient could have been explained by a coincidental distribution of local factors is therefore not supported by our models. We furthermore compared the distance from the most likely origin point to other distances from locations in the center of the various areas considered by Atkinson (Africa, Europe, Asia, North America, South America, and Oceania). We found that the distance from the best origin point was the best predictor when other local factors were considered. As a conclusion, the factors we considered in our study did not invalidate Atkinson's proposal.

Table 6.18 Regression of PhonemeDiv against sociolinguistic and environmental variables, with Atkinson's hypothesis included in the model

Predicted variable: PhonemeDiv									
	Type I Anova		Coefficients - standard regression			Coefficients - spatial regression			
	Sum Sq.	Pr(>F)	Estimate	Std. Error	Pr(>It|)	Estimate	Std. Error	Pr(>Iz|)	
(Intercept)			.493	.059	4.90E-16 ***	.283	.063	8.42E-06 ***	
NbSpeakers	26.08	2.81E-21 ***	.096	.037	.01 **	.055	.034	.113 n.s.	
LgArea	3.53	2.74E-04 ***	-.091	.032	.004 **	-.062	.03	.037 *	
LgContact	5.99	2.42E-06 ***	.054	.03	.075 .	.06	.028	.033 *	
LGP	7.45	1.56E-07 ***	-.09	.043	.035 *	-.102	.04	.01 *	
Rug	1.49	.018 *	-.158	.038	3.68E-05 ***	-.103	.035	.004 **	
Elv	4.6	3.42E-05 ***	.185	.038	1.43E-06 ***	.13	.036	2.56E-04 ***	
Tree	2.06	.005 **	.157	.042	2.06E-04 ***	.143	.039	2.86E-04 ***	
Dist	21.16	7.50E-18 ***	-3.78E-05	4.21E-06	7.50E-18 ***	-2.29E-05	4.63E-06	7.50E-07 ***	
Residuals	117.55								
Adj. R²	AIC	S-W	B-P	Cond.	VIF	Model: spatial lag			
.37	686	.562	.154	5.689	3.172	AIC: 648			
Lagrange multiplier diagnostics		LMerr	RLMerr	LMlag	RLMlag	B-P: .353			
		7.87E-11	.411	4.11E-14	8.25E-05	LM test:	9.37E-04		
Deleted outliers: late									

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Table 6.19 Regression of PhonemeDiv against sociolinguistic and environmental variables, with Atkinson's hypothesis included in the model and Family as a random effect

Predicted variable: PhonemeDiv			
	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-.408	.102	1.E-4 ***
NbSpeakers	.024	.040	.546n.s.
LgArea	-.047	.033	.151 n.s.
LgContact	.042	.030	.167 n.s.
LGP	-.082	.044	.063 .
Rug	-.099	.041	.015 *
Elv	.104	.040	.009**
Tree	.126	.042	.003 **
Dist	-3.40E-05	6.07E-06	<1E-04 ***
		Std. Dev.	
Family		.250	
Residuals		.455	
cor(fitted, predicted)²: 0.557, AIC:715			
Deleted outliers: late			

4. Conclusion

We have investigated in this paper the effects and interactions of a number of non-linguistic and linguistic factors. To this end, we have put special care in the preparation of the dataset, and tried to apply meaningful statistical tests. The complexity and intricacies of the possible approaches make the exploratory analysis of the data difficult. However, we found statistical effects that, we believe, can be related to reasonable causal relationships:

- Nettle's hypothesis regarding the impact of ecological risk on sociolinguistic variables such as the number of speakers or the area of a language is supported by our models;
- Similarly, the acoustic adaptation hypothesis also finds support in our study of consonantal diversity;
- To some significant extent, elevation and rugosity also predict sociolinguistic and linguistic variables, although we did not find a simple explanation of these effects;

- More generally, it seems fair to assume that environmental factors do have causal effects on sociolinguistic and linguistic parameters, given the consistent effects repeatedly found in our models.

Additionally, that a higher degree of linguistic contact leads to a higher phonemic diversity suggests that one should look at Atkinson's so called founder effect in a reverse fashion: the gradient may not be the result of languages gradually losing their phonemes along migratory routes, but rather the consequence of languages staying behind the waves of migration gradually gaining phonemes with the increase in linguistic density - itself due to the slow increase in population density.[4] On the contrary, human groups at the front row of the migrations were much less exposed to other populations and languages. Given the small size of our ancestors' communities before farming, this reading seems more convincing to us, regardless of the statistical issues raised against Atkinson's proposal.

Various improvements can be brought to our approach: other factors could be considered, as well as other statistical models, such as ordinal logistic regressions to better fit the specific distributions of measures of diversity. The acoustic adaptation hypothesis could also be assessed more thoroughly with the testing of more specific aspects of phonetic systems, for example the presence or absence of specific classes of consonants or vowels. All in all, the recent availability of large sets of data providing a wealth of information on the environment of human populations calls for further explorations, and will likely both broaden and strengthen the field of linguistics.

REFERENCES

Amante, C., and B. W. Eakins. 2009. ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis. *NOAA Technical Memorandum NESDIS NGDC-24*.

Anselin, L. 2007. *Spatial Regression Analysis in R. A Workbook*. GeoDa Center for geospatial analysis and computation, https://geodacenter.asu.edu/learning/tutorials.

Anselin, L., I. Syabri, and Y. Kho. 2006. GeoDa: An introduction to spatial data analysis. *Geographical Analysis* 38 (1): 5–22.

Atkinson, Q. D. 2011a. Phonemic diversity supports serial founder effect model of language expansion from Africa. *Science* 332:346–349.

Atkinson, Q. D. 2011b. Linking spatial patterns of language variation to ancient demography and population migrations. *Linguistic Typology* 15 (2): 321–332.

[4] Another factor influencing phonemic diversity may be the amount of within-population variability, although its relationship with population size is still debated (Trudgill 2004).

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

Boë, L.-J., P. Bessière, and N. Vallée. 2003. When Ruhlen's 'mother tongue' theory meets the null hypothesis, *Proceedings of the XVth International Congress of Phonetic Sciences.* Spain, Barcelona.

Bowern, C. 2011. Out of Africa? The logic of phoneme inventories and founder effects. *Linguistic Typology* 15 (2): 207–216.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B* 26:211–252.

Bybee, J. 2011. How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. *Linguistic Typology* 15 (2): 147–153.

Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI), The World Bank and Centro Internacional de Agricultura Tropical (CIAT). 2011. *Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Count Grid.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-count. Accessed 24–01–2013.

Coupé, C., and J.-M. Hombert. 2005. Polygenesis of linguistic strategies: A scenario for the emergence of language. In Minett, J. & Wang, W.S. (eds), *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics.* Hong Kong, City University of Hong Kong Press, 153–201.

Coupé, C., E. Marsico, and F. Pellegrino. 2009. Structural complexity of phonological systems. In Pellegrino, F., Marsico, E., Chitoran, I. & Coupé, C. (eds), *Approaches to Phonological Complexity*, Phonology & Phonetics Series Vol. 16. Berlin, New York: Mouton de Gruyter, 141–169.

Dahl, O. 2011. Are small languages more or less complex than big ones? *Linguistic Typology* 15 (2): 171–175.

Donohue, M., and J. Nichols. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15 (2): 161–170.

Dryer, M. S., and M. Haspelmath. (eds.) 2011. The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at http://wals.info/.

Fabrigar, L. R., D. T. Wegener, R.C. MacCallum, and E.J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4:272–299.

Food and Agriculture Organization (FAO) and International Institute for Applied Systems Analysis (IIASA). 2007. Length of growing period, 1901–1996. Global agro-ecological zones. In H. von Velthuizen et al. (eds.), *Mapping Biophysical Factors that Influence Agricultural Production and Rural Vulnerability*. FAO & IIASA.

Freedman, D. A., and W. S-Y. Wang. 1996. Language polygenesis: A probabilistic model. *Anthropological Science* 104:131–137.

Geospatial Information Authority of Japan, Chiba University and collaborating organizations 2008. *Global Map V.1 (Global version)*. http://www.iscgm.org/browse.html

Global Mapping International and SIL International. 2012. World Language Mapping System. Language area and point data for Geographic Information Systems (GIS). http://www.worldgeodatasets.com/language/.

Hay, J., and L. Bauer. 2007. Phoneme inventory size and population size. *Language* 83:388–400.

Hijmans, R. J., S. E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25:1965–1978.

Holdridge, L. R. 1947. Determination of world plant formations from simple climatic data. *Science* 105:367–368.

Jacquesson, F. 2001. Pour une linguistique des quasi-déserts. In A.M. Loffler-Laurian (ed.), *Etude de Linguistique Générale et Contrastive. Hommage à Jean Perrot*. Paris: Centre de Recherche sur les Langues et les Sociétés, 199–216.

Jacquesson, F. 2003. Linguistique, génétique et la vitesse d'évolution des langues. *Bulletin de la Société de Linguistique de Paris* 98 (1): 101–122.

Jaeger, T. F., P. Graff, W. Croft, and D. Pontillo. 2011. Mixed effects models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15 (2): 281–320.

Ke, J., T. Gong, and W. S-Y. Wang. 2008. Language change and social networks. *Communications in Computational Physics* 3 (4): 935–949.

Kottek, M., J. Griser, C. Beck, B. Rudolf, and F. Rubel. 2006. World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift* 15 (3): 259–263.

Lewis, M. P. (ed.). 2009. *Ethnologue: Languages of the World, Sixteenth Edition*. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Lupyan, G., and R. Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5 (1): 1–10.

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge, MA: Cambridge University Press.

Maddieson, I. 2005. Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. *UC Berkeley Phonology Lab Annual Report*, 259–268.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3

Maddieson, I. 2009. Calculating phonological complexity. In Pellegrino, F., Marsico, E., Chitoran, I., and Coupé, C. (eds), *Approaches to Phonological Complexity*, Phonology & Phonetics Series vol. 16. Berlin, New York: Mouton de Gruyter, 85–109.

Maddieson, I. 2011a. Consonant Inventories. In: Dryer, Matthew S., and Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, chapter 1. Available online at http://wals.info/chapter/1. Accessed on 2011–12–03.

Maddieson, I. 2011b. Tone. In: Dryer, Matthew S., and Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, feature 13A. Available online at http://wals.info/feature/13A. Accessed on 2011–12–03.

Maddieson, I. 2011c. Vowel quality inventories. In: Dryer, Matthew S., and Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, feature 2A. Available online at http://wals.info/feature/2A. Accessed on, 2011–12–03.

Maddieson, I. 2011d. Phonological complexity and linguistic patterning. *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong, China, 17–21 August 2011.

Maddieson, I., T. Bhattacharya, D.E. Smith and W. Croft. 2011. Geographical distribution of phonological complexity. *Linguistic Typology* 15 (2): 267–279.

Maddieson, I., and K. Precoda. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74:104–111.

Myers, R. H., D. C. Montgomery, G. G. Vining, and T. J. Robinson. 2010. *Generalized Linear Models: With Applications in Engineering and the Sciences*. John Wiley & Sons Inc.

Nettle, D. 1996. Language diversity in West Africa: An ecological approach. *Journal of Anthropological Archaeology* 15:403–438.

Nettle, D. 1998. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology* 17:354–74.

Nettle, D. 1999a. *Linguistic Diversity*. Oxford: Oxford University Press.

Nettle, D. 1999b. Using social impact theory to simulate language change. *Lingua* 108 (2–3): 95–117.

Nettle, D. 1999c. Is the rate of linguistic change constant? *Lingua* 108:119–36.

Pellegrino, F., C. Coupé, and E. Marsico. 2011. A cross-language perspective on speech information rate. *Language* 87 (3): 539–558.

Pericliev, V. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8:376–83.

Pericliev, V. 2011. On phonemic diversity and the origin of language in Africa. *Linguistic Typology* 15 (2): 217–221

Ribeiro Jr., P. J., and P. J. Diggle. 2001 geoR: A package for geostatistical analysis. *R-NEWS* 1 (2). ISSN 1609–3631.

Rice, K. 2011. Athabaskan languages and serial founder effects. *Linguistic Typology* 15 (2): 233–250.

Ringe, D. 2011. A pilot study for an investigation into Atkinson's hypothesis. *Linguistic Typology* 15 (2): 223–231.

Ross, B., and M. Donohue. 2011. The many origins of diversity and complexity in phonology. *Linguistic Typology* 15 (2): 251–266.

Ruhlen, M. 1994. *The Origin of Language: Tracing the Evolution of the Mother Tongue.* New York: John Wiley & Sons.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Ser. B* 13:238–241.

Sproat, R. 2011. Phonemic diversity and the out-of-Africa theory. *Linguistic Typology* 15 (2): 199–206.

Trudgill, P. 2002. Linguistic and social typology. In J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change.* Oxford: Blackwell, 707–28.

Trudgill, P. 2004. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8:305–20.

Trudgill, P. 2011. Social structure and phoneme inventories. *Linguistic Typology* 15 (2): 155–160.

Vautard, R., J. Cattiaux, P. Yiou, J.-N. Thépaut, and P. Ciais. 2010. Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nature Geoscience* 3:756–761.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* 4th ed. Springer.

Walker, R. S., and M. J. Hamilton. 2010. Social complexity and linguistic diversity in the Austronesian and Bantu population expansions. *Proceedings of the Royal Society—Biological Sciences* 278:1399–1404.

Whittaker, R. H. 1975. *Communities and Ecosystems*, 2nded. New York: Macmillan.

Wichmann, S., T. Rama, and E.W. Holman. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15 (2): 177–197.

Wang, William S-Y. 2011. Voices out of Africa? *Sunday Morning Post (South China Morning Post)*, May 29.

in Eastward Flows the Great River, Festschrift in Honor of Professor William S-Y. Wang on his 80th Birthday,
Peng G. & Shi F. (Eds), City University of Hong Kong Press, Hong Kong
ISBN: 978-962-937-216-3