University of Groningen

Linguistic probes into human history
Manni, Franz

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
General introduction
This chapter is unpublished, please cite it as follows:

GENERAL INTRODUCTION

This dissertation includes five research-articles published between 2006 and 2016 in peer-reviewed publications (Manni et al. 2006, Nerbonne et al. 2007, Manni 2010, Rodriguez Diaz et al. 2016, Mennecier et al. 2016) but also an extensive and unpublished report (CHAPTER 6, Linguistic probes into the Bantu history of Gabon) that summarizes 12 years of research.

The different chapters correspond to research that has been conducted over a time-span that is probably longer than the average involved in a PhD thesis because I had no intention of obtaining a second PhD degree (the first was in 2001 in population genetics) until recently, that is until becoming (2013) a member of the Scientific Committee in charge of determining the contents of the new permanent exhibition of the Musée de l’Homme (Paris).1 I was responsible for the section presenting the linguistic diversity of the world. During this experience, the scholars I approached were surprised to learn that I did not already hold a PhD in linguistics. Their reaction did not surprise me, because I am well aware that academia involves sensitivity to rules ensuring that research meets the standards of the disciplines so that it might enjoy the recognition of peers. The formal credential of the degree indicates that the bearer has operated within the system and is familiar with the expectations. I think this is a good scheme and, now, I would like to redress my lack of scientific credibility in linguistics by submitting this work to a doctoral committee. My agenda is to further develop scientific inquiry in the frame of demography and linguistics, especially linking the demographic aspects of speech-communities to the sociolinguistic effects that demography influences. Sociolinguistics is about contact among groups, and population genetics too.

Over the years I have interacted with many linguists. When I sought collaboration in research about surname and genetic variability in European populations, I found in PROFESSOR JOHN NERBONNE an excellent scientific partner because he had the answers to many of my questions and because he was willing to focus on new research problems that only tangentially involved his own scientific interests, at least initially. He uses computational linguistic methods that parallel what is done in genetics and the collaboration turned out to be beneficial for both of us. Over the years, I have received the encouragement of PROFESSOR SERGE BAHUCHET, the director of the scientific department where I work (Hommes, Natures Sociétés, National Museum of Natural History, Paris). He happens to be a linguist too and, when I asked him, he enthusiastically accepted to co-supervise this doctoral work.

---

1 The museum was reopened to the public the 15th of October, 2016.
My focus on languages and dialects developed outside the discipline of linguistics, and, since the chapters concern research involving at least three sister disciplines, linguistics, genetics and demography, this introduction in meant to provide a broader perspective, and not only a summary and description of the chapters that follow. Here I address the rationale for multidisciplinary research (1.1 Why genetics and linguistics?), report on comparative research involving family names and dialects (1.2 What family names tell about a population?), and, finally, introduce methodological questions related to computational linguistic classifications that are relevant to cross-comparisons (1.3 How to assess the reliability of linguistic classifications?; 1.4 Lexical databases). I end by providing the traditional outline of the dissertation chapters. The last one is not only about discussing what I have learned from the different experiments and projects. It is meant to attract the attention of reader by presenting some additional results, based on the datasets that are analysed in the different chapters, in order to link them together and suggest new methodological research directions. By assessing empirical evidence with novel approaches, I have tried to develop a conceptual frame that is larger than the scope of each chapter.

1.1 WHY GENETICS AND LINGUISTICS?

1.1.1 Some thoughts about the emergence of the language faculty in a social context

The cohesion of human societies relies on common beliefs and practices that are inter-related with the environment and the lifestyle. There are simple inferences concerning both the technical skills and communication repertoires needed to ensure the minimal viability of a (small) human group, but the harmonious development of larger societies, with respect to both population-size and the geographical area occupied, needs a more complex organization in which the benefit of the group is paramount, perhaps to the detriment of the individual.

The size of the group seems to be the fundamental parameter determining when the shift from personal to common interest arises leading to the concept on the minimum viable population size for a sustainable population. It has been estimated that a minimum of 150-180 individuals is necessary (Moore 2003), a figure that matches observations of existing groups of hunter-gatherer populations. If a minimum size is necessary for reasons related to the survival during crises, epidemics, climatic adversity and to avoid levels of consanguinity that are too high, the maximal size of a population certainly depends on its ability to communicate in order to maintain the cohesion of the society and its effective functioning. In this framework rules and taboos are seen as the necessary architecture of the social system. Intuitively, societies without
the communication net that a language enables would rely on immediate personal contact for the organization of their social life.

This is the standpoint of Mark Turner (1998) in explaining the emergence of speech as related to its narrative function, that is, to its role in structuring human behaviours according to shared beliefs that emerge from stories that are told and repeated. In a different way, the French researcher in artificial intelligence Jean-Louis Dessalles (2014a; 2014b), has suggested the argumentative function as the main advantage enabled by the language faculty. He developed and modelled mathematically several arguments showing that human communication did not emerge as a form of cooperation, but as the ability of the speaker to be relevant and display personal qualities, two crucial aspects to achieve influence and lead a group. The argumentative function of language makes possible to solve social conflicts with a decreased degree of physical violence and menace. Actually, the two theories are complementary and point to the possibility of developing larger societies through speech, societies that can maintain cohesion beyond frequent and direct personal contact and outside necessity. While under a functionalist viewpoint languages would emerge just to improve the practical organization of life, there are exceptions showing that in small groups languages can be extremely simple, such as in the famous example of the Pirahã, a language of Amazonia that seems to lack recursion and has a very limited set of words and phonemes (Everett 2005). Actually the Pirahã people are monolingual hunter-gatherers, and their idiom is the only surviving variety of a language that went extinct. They live in a single small group of about 200 people; they all know each other, and they are all related. Direct observation and visual learning are able to substitute for the functions of a majority of the words. This is reminiscent of familiar jargon that is also limited but sufficient for everyday life, and if the group were larger, greater linguistic complexity would be expected. Instead of the attempted demonstration that Chomskyan theories are wrong (Everett 2012), the case of Pirahã might be brought to bear from the perspective of the social benefit that a more complex language gives to the cohesion of a larger social group.

Following Chomsky, the quantification of rumours and gossips in natural speech is relevant to the role of language communication in sustaining large social groups whose members have irregular direct contact. A recent study (Beersma and Van Kleef 2012) shows that 90% of conversations in a professional environment concern gossip about colleagues, and this estimation echoes a similar figure concerning human speech in general (Dunbar 1996). Spreading rumours and gossiping takes

---

2 This claim created a lively debate, still ongoing. See Nevins et al. (2009) and Everett (2009) for some elements about it.
place, by definition, in conversations where the people mentioned in the stories are not there. In a social context, rumours are very useful because they contribute to increase the cohesion: they remind the interlocutors of the existence of members of the society that might be far away or seldom encountered, in the same way that telling stories about dead persons functions to preserve memories and traditions over generations. The link existing between the population size and the language seems clear.

1.1.2 The emergence of the language faculty and the peopling of the world from Africa

Chomsky’s hypothesis of a Universal Grammar, implying that the human mind disposes of innate structures allowing us to acquire, comprehend and use language, might be seen through the new prism of neuroimaging and genetics. Mirror-neurons and specific genes are involved in speech production, meaning that there is an inherited genetic base enabling language. This kind of research is currently in its infancy and must be linked to social cognition (see Fitch et al. 2010 for a challenging review), but it could well be that some DNA mutations determined a rearrangement in the neuronal nets of our brain leading to new cognitive abilities that made us better able to communicate.3 These genes were probably under positive selection because communication was granting to the offspring a higher degree of survival.

The rise of verbal communication and its complexity enabled us to distinguish space and time and facilitated the expression of symbolic thinking, together with the social advantages that have been listed in the preceding paragraph. The emergence of the language faculty is one of the plausible hypotheses that have been advocated to explain why and how the descendants of human groups having left the African continent some 50,000 to 100,000 years ago have been able to successfully colonize all the continents (Mellars 2006), leading to the almost complete replacement of pre-existing human populations (Homo erectus, Denisovan hominin, Homo floresiensis and Homo neanderthalensis). While interbreeding with these geographical species was possible, it seems to have contributed to less than 10% of the genome of modern humans (Green et al. 2010) either because of low inter-fertility or because of a larger population-size of the immigrants, maybe related to a more efficient use of language. Interestingly, a regulatory gene that all mammals share (FOXP-2) (see Takahashi et al. 2013) has been

3 By the way Chomsky himself argues that communication is secondary, and that language evolved because of the advantages to think. He sees language as a set of atomic elements and rules that allow the construction of more complex rules of thinking according to a criterion he calls computational efficiency, which is different from communicative efficiency.

For example see: http://www.u-plum.fr/actualites/232-conference-de-noam-chomsky
recently suggested to be necessary for proper development of speech and language in humans, but the sequencing of the genome of fossil DNA concerning *Homo neanderthalensis* has confirmed that the gene was present in the latter as well, meaning that it is well possible that Neanderthals could speak (Coop et al. 2008; Krause et al. 2007, Vargha-Khadem et al. 2005). Language is an important factor explaining human evolution.

1.1.3 *The recurrent interest of geneticists in the diversity of human languages*

The link between the emergence of an improved speech production and a better ability to migrate and to constitute viable and successful societies throughout the world, echoes an older and similar debate, in the 1980s, concerning the genetic evidence for the Out of Africa model, which seems supported by the extant worldwide linguistic diversity. As an answer to the critics of the evidence for a migration wave that led to the re-peopling of the world from Africa in “recent” times (Cann, et al. 1987; Stoneking and Cann 1989), Cavalli-Sforza et al. (1988) focused on cultural evolution and published, side by side, a phylogenetec tree based on human genetic diversity and another corresponding to a worldwide linguistic classification provided by Merrit Ruhlen (1987). He suggested that the similarities between the two classifications were the proof of synchrony between cultural and biologic diversity, implying that cultural divergence happened over a timeframe comparable to that of genetic differentiation. As diversity accumulates at a much faster pace in languages than in DNA, the fact that a correspondence was found had to be seen as a demonstration that all present human populations had a common and recent ancestor in Africa. This paper attracted wide attention and led to a schism between population geneticists and the community of linguists. In fact, the work of Ruhlen had stood outside accepted comparative linguistics by rejecting the “temporal ceiling” beyond which the comparative method fails, considered by some (Kaufman 1990; Nichols 1992) to lie at roughly 6,000 to 8,000 years ago. Ruhlen was comparing lexical items with supposedly close meanings showing resemblances in large linguistic groups called macrofamilies (1994), but these

---

4 Actually, in more general terms, the current view is that the evolution of human speech capabilities required neural changes rather than modifications of vocal anatomy. Macaques have a speech-ready vocal tract but lack a speech-ready brain to control it (Fitch et al. 2016).

were also explainable as chance-similarities estimated to be equally likely once one treats meaning correspondences as approximate as Ruhlen does (Boë et al. 2003).

For their part, many population geneticists did not realize that their continued high esteem for Ruhlen’s analyses was disqualifying their work in the eyes of a vast majority of historical linguists, perhaps also because a scholar as authoritative as Charles Darwin had foreseen a match between the two disciplines, formulating his ideas in what was to attain the status of an *Ipse dixit*:

>If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, had to be included, such an arrangement would, I think, be the only possible one. Yet it might be that some very ancient language had altered little, and had given rise to few new languages, whilst others (owing to the spreading and subsequent isolation and states of civilisation of the several races, descended from a common race) had altered much, and had given rise to many new languages and dialects. The various degrees of difference in the languages from the same stock, would have to be expressed by groups subordinate to groups; but the proper or even only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and modern, by the closest affinities, and would give the filiation and origin of each tongue (Darwin, 1859, 422–423).

>Setting aside all the theories that have been mentioned, does it make sense to compare genetic diversity and linguistic diversity? Why should a link be expected? As complex as they might have been, the majority of human societies have left few traces behind them: some artefacts, some bones, when the soil was not too acid to melt them, and lineages of offspring that have sometimes survived until the present. Ignoring aspects that might have outlived those societies such as useful tools and techniques, which are difficult to link to more abstract beliefs and myths, the cultural traits that defined such societies, and their symbols, have generally disappeared to a very large extent. Given the importance of communication, and as long as there is demographic continuity in a society, a language is likely to be maintained; unless external influences come into play, leading to bilingualism and language shifts, perhaps because of military threats or because of the attractiveness of the social and economic model associated with another language. Nevertheless languages do change over time, probably in relation to the size of the population speaking them, according to the tightness or looseness of the linguistic-net connecting the speakers (social strata, geographic distance), and depending also on sociolinguistic factors and, of course, on linguistic contact. Continuity normally exists so that we can explore the extent to which two peoples speaking re-
lated languages also share some genetic make-up. This is a likely tendency, but clearly does not hold in complete generality.

This is the very prudent position I embraced when I decided to initiate the multidisciplinary research reported on in this dissertation and, as I was familiar with population genetics research addressing surname diversity, I decided to compare the latter with the linguistic diversity that is found at national scales, that is the variability of dialects. The models suggest correlations at this scale, and the smaller scale should be more easily testable. Chapter 2 is partly about this question.

1.1.4 Towards a wider Anthropology

A wider and synthetic Anthropology is becoming one of the aims of a large community of scholars and I would like to mention the remarkable efforts made by Professors Luigi-Luca Cavalli-Sforza and Marcus Feldman in promoting a trans-disciplinar approach to human and culture evolution. Their book, *Cultural transmission and evolution*, as well as many subsequent articles, constitute a landmark that has inspired similar efforts outside anthropology in its narrower sense. Later, the archaeologist Colin Renfrew also embraced such views, organizing a series of conferences that “forced” the debate among archaeologists, linguists, geneticists and demographers, later ensuring that the views be publicized in a series of high-quality books (Renfrew and Boyle 2000; Bellwood and Renfrew 2002; Forster and Renfrew 2006). By applying methods from the natural sciences (Bryant et al. 2005), while maintaining contact with the primary linguistic and cultural data, questions considered intractable have been addressed. To mention only the research of those with whom I have personally interacted, the research of Professor Russell Gray (Max Planck Institute for the Science of Human History) has made the subfield of historical linguistics less hesitant about sources of evidence that proceed through digital computing (see for example Gray and Atkinson 2003, Dunn et al. 2011, Bouckaert et al. 2012). Comparable efforts to develop a wider Anthropology have been undertaken by Professors Stephen Shennan and James Steele (UCL, UK), especially concerning demography and cultural macro-evolution (see for example Steele and Shennan 2009), and by Professors Mark Pagel (University of Reading) and Tecumseh Fitch (University of Vienna) concerning linguistics (see for example Mesoudi et al. 2011, Pagel et al. 2013, Grollemund et al. 2015). My research interests have been shaped by this incredibly large body of scientific research.

---

6 By inviting them to conferences and symposiums I organized or by co-editing special issues published in *Human Biology* (Wayne State University Press, Detroit, MI), a Journal I have edited from 2008 to 2013.
While a majority of the scholars I mentioned above are biologists, archaeologists, psychologists, meaning that they approached linguistics from a different background, unlike historical linguistics, dialectology has its own computational tradition. It started with Jean Séguy (1914–1973) and was further fostered by Hans Goebi (University of Salzburg—see Goebi 2006) and pursued by the scholars of the Department Alfa-Informatica at the University of Groningen. If dialectology has experienced this digital evolution earlier, it is probably related to the availability of larger corpora than those existing in historical linguistics. This is similar to the surname studies in population genetics, markers that are freely available in large amounts and that have made possible very detailed studies well before the advent of DNA sequencing technology (see Colantonio et al. 2003, Darlu et al. 2011 for a review).

1.2 What family names tell us about a population

The application of models based on surnames to infer the genetic structure of human populations relies on the parallel transmission of surnames and Y-chromosome DNA, at least in occidental naming practices, where a legitimate son has the surname and the Y-chromosome of his father. This is the reason why the variability of the surnames (i.e. their different types and their frequencies) can be quantified to estimate the consanguinity of different societies, without the need to undertake expensive laboratory DNA typing. This variability also enables the evaluation of population isolation, differentiation, and the directionality of migrations. Many population geneticists made major contributions to this field, including Crow, Cavalli-Sforza, Morton, Relethford, Lasker, and Barrai (see Lasker 1985, Colantonio et al. 2011). Surname methodologies have been applied to more than thirty societies, all around the world. The geographic scope ranges widely, from the household or village to a whole continent.

The confirmation that surnames indeed mirror genetic variability came with the advent of full genome sequencing, a technique allowing very deep inferences about regional and micro-regional genetic differences that can be explained by demographic factors that, in turn, can rely on historical and cultural processes. Family names of patrilineal descent have indeed proved to mirror a single locus on the Y-chromosome (King and Jobling, 2009). However, the temporal depth of surnames is limited (between 4 and ±30 generations)7 when compared to the scale of demographic processes inferred by molecular markers, and in any case variation in the Y-chromosome represents an extremely small amount of the total genetic variability.

---

7 About 20/25 generations in European Christian countries as the Roman Catholic Church started to register births and deaths after AD 1563.
In this context, why should anthropologists take into consideration surname information that, albeit easier to collect than DNA data, is sometimes tricky to interpret? One simple answer is that family names allow a retrospective look at human variation, which is hard to achieve with DNA studies because these require living populations, if they are to be conducted at a reliable scale. Historical documents often report surname information over several successive generations, and with a degree of polymorphism that (for the moment) is larger than the one available with DNA. The major strand in surname studies rests on the exploitation of databases that are increasing in size and exhaustiveness due to widespread digitization. In this respect, Pablo Mateos and Paul Longley’s UCL Worldnames database,8 which includes about 6 million surname-types registered in 26 different countries, constitutes an impressive quantity of information and a valuable tool for future research (Mateos 2011). Millions of different types of surnames are drawn from diverse sources, such as national electoral registers, telephone directories, or national online censuses. Moreover, these can be organized according to lexis, phonology (vowels, consonants, morphology) and based on surname type (derived from place names, professions, nicknames, or first names). The second major research direction, besides these attempts to draw from modern registers a large number of surnames in vast geographic regions, involves a focus on historical data that I am going to skip here, but see Darlu et al. (2001).

The large expansion of the available data, both in time and space, has led to the development of new methods and analytical tools. Among them, and now widely used, are automatic geographic representations of surname diversity which plot the variations of frequency of a given name, or a set of names, sharing some phonetic or grammatical features. Some recent statistical methods are also becoming established, such as Bayesian approaches to infer the origins of migrants (see contribution of G. Brunet in Darlu et al. 2012), Self-Organizing Maps to automatically identify surnames sharing the same geographical origin (Manni et al. 2010), or to identify ethno-cultural groups (Mateos et al. 2011). The purpose of Mateos was to create a ‘universal’ classification of forenames and surnames by ethnic group. 250,000 surname-types and 120,000 forename-types have been aggregated in 150 possible cultural categories. The classification is empirical, not based on scientific or ethnologic background and the methodology is based on a technique of cross occurrences between forenames and surnames (Mateos and Tucker 2008): with a given forename it is possible to retrieve related surnames and, from the latter ones, to retrieve corresponding forenames. Iteration after iteration, the database expands until a stage is reached where all retrieved surnames correspond to a same set of forenames and vice versa. In this way it is possible to iden-

---

8 http://worldnames.publicprofiler.org/
tify clusters of linked individuals corresponding to several isolated clusters existing in the population. They allow depicting contemporary migrations, multiculturalism and assimilation and the results have a direct interest for social anthropologists and population geneticists.

1.2.1 Surnames and dialects

Family names carry social and economic information that merits inclusion in several interdisciplinary approaches to human history. Historians, linguists, and geographers can play as active a role as biologists in surname studies and population analysis. Today, in an age of global migration (Castles and Miller 2009), the distribution of surnames remains far from random and has the potential to allow an intermediate level of access to the recent past and to smaller geographical scales, both of which are difficult to obtain otherwise: this is where parallel studies in dialectology become intriguing. By providing evidence of migration phenomena in different periods, it is possible to delineate past genetic isolates and population structures that have been modified or disappeared altogether, thus enabling demographic hypotheses to be tested linguistically (Falck et al. 2012).

By identifying the geographic origins in large corpora of surnames at national scales and aggregating the results, a double matrix (immigration and emigration) leads to the identification of provinces falling in four categories: 1) attractive provinces towards which immigration has been strong but from which emigration was weak, which is typical of important urban areas; 2) unattractive provinces, from which emigration has been considerable but which have failed to attract immigrants; these are usually economically poor areas where the surname-set of the population has remained closer to its initial make-up at the time of surnames’ introduction; 3) corridor provinces, in which immigration and emigration have been important phenomena leading to a considerable modification of the surname-set over the time; and 4) isolated self-sufficient provinces that have never really attracted immigration but from which people have not left; these often correspond to geographically isolated areas. These four classes of regions match alternative demographic phenomena, and it is likely that these had different impacts on regional linguistic variability.

As a hypothesis to be tested, one would expect to find higher dialectal diversity, and clearer geographical structure, in the provinces where the number of immigrants speaking external varieties has been low, since the immigrants could not influence the speech communities much (unattractive and isolated provinces). On the other hand, the areas that have been the target of massive immigration are expected to have lost linguistic variability, since some of the original local groups will have lost the
critical size needed to support a linguistic variety; in fact the identity-marking function of local varieties is less relevant where the number of allochthonous speakers is too large. This hypothesis suggests that demography and variationist linguistics should be able to collaborate profitably.

1.2.2 How to compare dialects and surnames?

The cross-comparison of surname and dialect variability (Manni and Barrai 2000, 2001) made me face the same kind of problems that other geneticists were confronted with, that is the impossible comparison of quantitative information, such as frequency-vectors and distance matrices summarizing surname variability, to the qualitative descriptions and maps that dialectologists often use. An obvious way to circumvent the problem was to compute maps of surname variability to be visually compared to existing dialect atlases that, sometimes, report appealing isoglosses. This is the reason why I turned to methods focused on the computation of genetic barriers: areas where the rate of change of a given variable is higher (Manni et al. 2004), boundaries making sense geographically and/or linguistically. But this kind of comparison was unsatisfactory because it is rather easy to lie with maps (Monmonier 1996).

As a population geneticist, I was familiar with DNA sequence alignment and aware of an interesting paper addressing, by sequence comparison, the time that toponyms of Arabic origin in Sicily needed to evolve into their current Italianized form, such as Moio Alcantara (Province of Messina) and al-maya al-kantara which means “the water of the bridge” (Barrai, 1993). Inspired by this example (Fig. 1.1) and with the help of Professor Barrai (University of Ferrara), I later developed an alignment algorithm to measure the linguistic diversity of some dialects, to evaluate the congruence between linguistic differences and surname/genetic diversity of the province of Ferrara (Manni and Barrai 2000, 2001). Linguistic distances were computed according to what we called unit cost model for insertion, deletions, point mutations, presence/absence of the article but not normalizing our alignments by their length (Fig. 1.2). We were unaware that Kessler (1995) had already applied this same Levenshtein method to Gaelic dialects and that the research group of John Nerbonne had largely improved it over the years (Nerbonne et al. 1996; Heeringa 2004). Actually the idea was “in the air” as the book Time Warps, String Edits and Macromolecules clearly shows (Sankoff and Kruskal, 1999).

When I contacted the research group of Groningen to ask for assistance, they were in the process of computationally measuring the diversity of Dutch dialects and, by an incredible coincidence, I was processing Dutch surnames. This circumstance and the methodological vicinity favoured the establishment of a collaboration that this
PhD dissertation witnesses. In this way the Levenshtein algorithm has represented the needed computational base to compare linguistic diversity with family name diversity. CHAPTER 4 is about this first collaboration that was quite dialectic because we were unsure about how to relate the patterns of surname and dialect variation because of instability in the clustering at finer scales. The hierarchical clustering of surnames had been tested by resampling techniques, while dialect Levenshtein distances had not. This uncertainty led to a phase of methodological progress directed to better assess the reliability of linguistic classifications (next section).

![Figure 1.1](image1.png)

**Figure 1.1** ‡ Early alignment of lexical data. Matrix of dots used by Barrai (1993) to estimate the homology between toponyms of Arabic origin in Sicily according to the original pronunciation and to the present ‘Italianized’ one. The diagonal visible among the dots corresponds to matching strings.

<table>
<thead>
<tr>
<th>Case</th>
<th>Penalty</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion/Deletion</td>
<td>1</td>
<td>Term: Baffi (Moustache)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loc. A: Ba ffi; Loc. B: Baffi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A) → ba ffi → .......</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) → ba. fi → .......</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Penalty:</strong> 1 (1 ins./del.)</td>
</tr>
<tr>
<td>Point mutation</td>
<td>1</td>
<td>Term: Carne (Flesh)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loc. A: Caran; Loc. B: Caran</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A) → caran → .... e.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) → caren → .... e.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Penalty:</strong> 1 (1 trans./transv.)</td>
</tr>
<tr>
<td>Presence/absence</td>
<td>1</td>
<td>Term: Ombelico (Umbilicus)</td>
</tr>
<tr>
<td>of the article</td>
<td></td>
<td>Loc. A: Umbregual; Loc. B: Al bigues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A) → umbregual... → .... e....--</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B) → al . . . b.igu.los → ....-i ..--</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Penalty:</strong> 5 (6 ins./del; 1 abs. art.; 1 transv.)</td>
</tr>
</tbody>
</table>

![Figure 1.2](image2.png)

**Figure 1.2** ‡ Description of the unit cost model through which linguistic differences were calculated in Manni and Barrai (2000, 2001).
1.3 HOW TO ASSESS THE RELIABILITY OF LINGUISTIC CLASSIFICATIONS?

A recurrent limitation of computational dialectology is that sometimes different clustering algorithms lead to conflicting classifications. In an effort to overcome the instability in clustering and get correct results, some research had been directed towards the definition of the optimal clustering algorithm (UPGMA, Ward’s, etc) to be adopted within dialectology, proceeding from an examination of the different biases of the different clustering methods and comparing these to the assumptions of linguistic models, but it is not always easy to theoretically justify why one algorithm is better than another, even if empirical experience can give some clues about it (see chapter 6 in Heeringa 2004). An alternative way to evaluate clustering for dialectology is to examine the frequent discrepancies between a dendrogram (the result of hierarchical clustering) and the projection of a distance matrix onto a two- or three-dimensional plot (obtained e.g. by multi-dimensional scaling or principal component analysis), assuming that the reduction to a small number of dimensions does not involve too much information loss. In routine analyses, it can happen that the clusters identified in the dendrograms are hardly visible, if at all, in the multivariate plot. On the other hand, well identified clusters in multidimensional plots always show up in dendrograms. These rules of thumb and clues were insufficient to definitively compare the geographic patterns of surnames and dialects, this is why we turned to the application of bootstrap to dialect data.

1.3.1 Efron’s (1979) bootstrap

The bootstrap method (Efron 1979) is a technique for obtaining standard errors and confidence limits in various statistics. The basic idea is to take the sample for which one hopes to estimate a parameter \( \pi \) using a sample statistic \( p \) and, once \( p \) is computed, if the sample is of size \( n \), one carries out a random resampling procedure by sampling with replacement, using the \( n \) items from the observed sample as a parent population. For each of these new (resampled) samples one estimates the desired parameter \( \pi \). Because we are sampling with replacement, most samples will contain two or more replicates of some variates appearing in the observed sample and, consequently, some other variates will be missing. It has been shown that the mean of the estimated statistic, from the bootstrapped sample, approximates the mean of the population, and that the standard deviation of such an estimate approximates the standard error of the statistic, as if we had repeatedly sampled from the unknown population without replacement. This is a very important result because it permits us to calculate standard errors for almost all statistics (Sokal and Rohlf 2001, pp. 823-25).
1.3.2 Felsenstein’s (1985) bootstrap

Considered to be an improvement over the Jackknife (Efron 1979), the bootstrap has been extensively used in biostatistical applications and has become popular in numerical taxonomic and phylogenetic research where it was introduced by Felsenstein (1985). In phylogenetic estimation, the branching sequence of \( t \) (animal, plant, etc.) species is estimated from a set of \( n \) characters, which vary in their states among species. The estimates are based, for example, on constructing the shortest tree, in terms of the amount of implied evolutionary change for a given data set. After constructing such a tree, the comparative biologist may conclude that species A and B are closer to each other than they are to other species—that is, that they form a monophyletic taxon. To test the reliability of such a taxon, Felsenstein suggested sampling \( n \) characters with replacement from the original data set, in order to create new data sets from which new minimum-length trees are constructed. The result is a number, say \( m = 100 \) trees. If 65% or more of the trees contain the taxon \( \{A, B\} \), then this branch of the tree is considered well substantiated. If, by contrast, only 30% of the bootstrapped trees show the set \( \{A, B\} \), little reliance can be placed on that portion of the taxonomic structure of the tree.

At this point it should be clear that bootstrap can be seen as a method to estimate the robustness of a given classification of taxonomic units in the sense that it tells if a given difference between them is widespread or not among the characters that describe such taxonomic units. Bootstrap is not a test of how accurate a tree is; it gives information about the stability of the tree topology (the branching order), and it helps in assessing whether the data is adequate to validate the topology.\(^9\) An implicit as-

\(^9\) A simple example will make the essence of bootstrap clearer, let’s imagine that we want to classify two persons according to the colour of their clothes (1. cap; 2. shirt; 3. trousers; 4. socks; 5. shoes). If the two individuals are dressed in completely different colours, when resampling the 5 characters that define them in new bootstrapped datasets we will find that all resampled datasets lead to the conclusion that they are different. This is obvious because each colour character that defines them (the colour of their caps, shirts, trousers, socks and shoes) is always different. In such case the difference between these two individuals is said to be supported by a 100% bootstrap score. Differently, if the two individuals have clothes identical in colour besides their shoes (one having brown shoes and the other having black shoes), we will find that the difference between them will be supported by a minority of the resampled data sets, since there are good chances that the only character that conveys a signal of difference (the shoes) will often be left-out in the resampling procedure, simply by chance. Only those datasets containing the colour of the shoes as one of the characters will support a difference between them, whereas those datasets not containing such character will not. If resampled datasets are numerous enough to allow each character to be sampled with equal probabilities, then the difference between the
sumption of the bootstrap technique is that different characters that define the items under classification bear equal weight in determining the classification.

1.3.3 Application of resampling techniques to dialectology

The application of bootstrap to dialectology allows us to assess the degree of reliability of a dialect classification, that is to test if the variability of the characters (words) that describe each dialect variety supports the final classification, and to what extent. As an hypothetical example, we can think of a group of dialect varieties that, according to scholarly traditions, are usually split into two groups (say a northern and a southern group) and investigate how many of the words that define such varieties exhibit a North/South differentiation by randomly resampling the set of words into many new auxiliary datasets, from which we then compute a corresponding number of distance matrices and, in turn, of dendrograms. If a majority of the words exhibits a North/South difference, then a majority of the trees obtained from the resampled datasets will show a major split between the northern and southern cluster, which could be intuitively summarized in a synthetic tree where the fork separating such clusters is substantiated by a percentage higher than 50%.

When using a resampling technique to generate new databases, we can make inferences about the strength of a given signal (North/South) between the words that constitute the database, since in the new resampled datasets some words will be over-represented while some others will be missing, meaning that the randomness of the resampling process gives a different random weight to the original elements constituting a database. To say it differently, once the initial weights are all equal, the resampling procedure randomly emphasizes, or diminishes, the weight of some words and tells if a split (North/South) is supported by only a few words of the database or by a majority of them.

1.3.3.1 Bootstrap consensus trees

By definition, dendrograms are constituted by branches that are topologically identified by the different splits (nodes) that partition the data into groups (clusters) of different size. A bootstrap consensus tree is a summary of a set of trees computed according to a clustering algorithm (UPGMA, Ward’s, etc.) from a set of distance matrices obtained from resampled databases. As said, visually a bootstrap consensus tree can be identified since its forks (nodes) are coupled with a score proportional to the number of times a given fork (node) appears in all the original trees it summarizes.

two individuals will be supported by a bootstrap score of ~20%, since they differ by one character over five.
There are at least two ways to compute a bootstrap consensus tree from a set of trees: 
a) strict consensus and b) majority-rule consensus. A strict consensus tree consists of 
all groups that occur 100% of the time, the rest being ignored. Less stringent, a major-
ity-rule consensus (MRC) tree consists of all groups that occur at least 50% of the time 
(Margush and McMorris 1981). We stress that in MRC trees there cannot be two con-
flicting splits supported by more than half of the trees in the same time. If they were, 
there would be at least one tree containing two conflicting splits at the same time, 
which is absurd.

The procedure to compute a MRC tree is quite simple and consists of three 
steps: 1) computation of all splits; 2) removal of all splits supported by less than half 
the trees; 3) computation of the consensus tree containing the splits. The majority-rule 
can be extended to display some other splits that are supported by scores lower than 
50% From now on a majority-rule extended consensus tree will be referred to as 
MREC tree. Rephrasing Hillis and Bull (1993), the bootstrap value is a count or per-
centage of how often each branch was present in exactly the same topology in all the 
resampled trees, so it gives an impression of how much the tree topology could 
change if, for example, you’d reconstruct it using a different set of words.

1.3.3.2 Adoption of a cut-off value
We said that the nodes of a MREC bootstrap tree can be supported by values varying 
from 1% to 100% indicating how stable (robust) their topology (branching order) is. In 
biology there are many rules of thumb about how to interpret the bootstrap values, 
and a score of 95% is sometimes taken as the minimum since it reminds the 95% con-
fidence interval used in statistics. Nevertheless a score of 70% is often cited as a cut-off 
for a ’reliable’ branching (see also Hillis and Bull 1993). When applying bootstrap pro-
cedures to dialectological data we cannot rely on any available rule since, to our 
knowledge, this dissertation includes the first papers using bootstrap to assess the 
robustness of dialect variants. When choosing a cut-off value for dialect data, implicit-
ly, we are making a decision about the minimal difference that should exist between 
two dialect variants, if they are to be considered different. In other words: If we record 
100 words in two dialects and it appears that these two are perfectly identical apart 
from 10 words that are different (meaning a low bootstrap score in the fork separating 
them), will we still consider these variants as distinct or would we decide that, actually, we 
are dealing with one single variety and that the observed difference has to be attributed to 
noise? This is a very thorny question since dialectologists have been keen to catalogue 
new variants and to extend the sampling grid. The measuring of the noise related to 
variation within a same village, or town, in order to establish which percentage of 
variation has to be interpreted as the sampling-error has not received much attention,
probably the first paper to provide a mean and a standard deviation in dialectological measurements is Nerbonne et al. (1996).

1.4 The fuel: Lexical Databases

1.4.1 The number of items

All the research included in the dissertation is based on the comparison of lexical items – the sort of material that is generally readily available from linguistic atlases and databases. We generally try to include about 100 items, because this number has been recognized to be sufficient for aggregate analyses. In fact the words that are processed are in general composed of 5 segments, meaning that the Levenshtein method, involving sequence comparison, can extract a more robust signal from a short list of lexical items (Heeringa et al. 2002) than earlier dialectometric procedures (see below).

I will comment on two aspects of the dialectometric approach adopted by Nerbonne, Heeringa and others in Groningen. First, they insist, following Goebl (2006) and others, that reliable indication of the relations among dialects can only be gleaned from the examination of larger aggregates of comparable dialect material. We recall here Grimm’s (1819) adage that “each word has its own history” (see also Kirk et al. 1985), which we take to be true if we understand it to mean that any single word or dialect feature may be misleading with respect to the inferences it suggests about the relations among dialects (see Nerbonne 2009 where the argument for using aggregates is elaborated on). Some words and features are capricious in some aspects of their distribution.

The insistence on examining aggregates is shared by Goebl (2006 and elsewhere), but the Groningen direction has tried to simplify processing and to squeeze more information out of dialect atlas collections by employing sequence comparison (the Levenshein algorithm) extensively. Séguy’s and Goebl’s work tried word lists as categorical data where items might be the same or different. So the aggregate measure of similarity between settlements (represented by word lists) was the fraction of identical elements, and the distance the inverse, i.e. the fraction of different elements.\textsuperscript{10}

Sometimes the words were examined to extract a single, simpler feature, e.g. whether the stress vowel was pronounced high in the mouth (as an [u]) or low (as an [a] or [d]). The extracted feature was then compared categorically, but this comes at the cost

\textsuperscript{10} I am ignoring Goebl’s work on weighting schemes, but it is orthogonal to the point.
of losing information in the other elements in the word’s pronunciation, and at the cost of manually extracting the relevant feature. The Levenshtein algorithm compares the entire sequence of sounds, thereby incorporating more information, and, because it is automated, obviating the need for manual pre-processing. Finally, because it produces a numerical measure of the difference, it is more sensitive than earlier dialectometric procedures.

By using consistency measures, Heeringa (2004, p. 176) has demonstrated that pronunciation-based classifications tend to remain stable after processing a minimal number of items of about 30, and that the signal that a 100-word list delivers is not significantly different from the one of a 200-word list and, finally, that what matters is the choice of the words that are used instead of their number. Categorical comparisons require three to four times as much data to reach stability.

1.4.2 The choice of the words

With the Levenshtein method we measure pronunciation distances. While a random list of words is as likely to mirror the phonology of a language than a more specific wordlist, we stress that we are not comparing phonological repertoires but, each time, the pronunciations of a given word in variety A to the pronunciation of the word corresponding to the same concept in variety B, in a pairwise fashion. If shared vocabulary is large, for example because of extensive borrowing, computed linguistic distances will be lower than in the opposite case. If this bias does not seem to be a serious issue in dialectology because borrowing is not distinguishable in close varieties, it can lead to a systematic error in assessing linguistic diversity when analyzing more distantly-related varieties and languages.

In this direction, the lexicostatistics work of Swadesh (1955) is relevant because it was intended to identify lexical items less likely to be borrowed, words that constitute a basic vocabulary that is expected to have emerged first in every language, because necessary (like ‘water’, ‘fire’, ‘head’, etc.). Basic vocabulary has probably remained stable over the time, meaning that it was not borrowed from another speech community because each group of speakers has its own, long established, words for these concepts. Basic core vocabulary was needed in lexicostatistics to better address the historical phylogeny of languages (see Dyen 1975).

1.4.1 Swadesh wordlists

Swadesh word lists are processed in two chapters of the dissertation. In Chapter 6 (concerning Bantu languages), the word-lists we addressed correspond almost sys-
tematically to the core vocabulary. We analyzed a database assembled at the Musée Royal de l’Afrique Centrale, Tervuren (Bastin et al. 1999) and, since this institution has been very active in developing lexicostatistical approaches, the wordlists were based on Swadesh lists. In Chapter 7 (concerning Central Asian languages), we process Swadesh word lists too, but in this case we did not have to rely on a existing database, as we did the fieldwork. The reason to use this kind of word list was to provide novel linguistic material in a form that is comparable to available literature, largely based on Swadesh lists. If, during the fieldwork we noticed that such items offered advantages because speakers pronounced them without hesitation, we had to exclude some of them, either because polysemic or because not adapted to that specific ecological context.

This is similar to what Hombert (1990) did when he designed the wordlist to be used in the Atlas Linguistique du Gabon (ALGAB). To anticipate two tangential findings that will not be discussed in the final chapter of the dissertation, we noted inconsistencies between the classification of the two Gabon datasets we processed (Bastin et al. 1999 and the ALGAB, respectively based on an average of 89 and 132 items). Instead, the classification of 32 Tanzanian Bantu languages did not change when processing a 1400-word list or a subset of it consisting in 92 Swadesh concepts (see Chapter 6 concerning both examples).

While several articles have speculated on the effectiveness and the appropriateness of the Swadesh list in historical linguistics (Kessler 2001; McMahon and McMahon 2005; Holman et al. 2008), the Loanword Typology project\(^\text{11}\) coordinated by Haspelmath and Tadmor (2009) identified 113 concepts that are the most stable ones (among 1460 lexical items) in about 50 languages:\(^\text{12}\) 44 of them (22%) are included in the 200-word list of Swadesh. This result is interesting because i) it shows that the Swadesh list is less stable than previously assumed, ii) it suggest that Swadesh lists

\(^{11}\) http://wold.clld.org/

\(^{12}\) The list of the 113 more stable concepts according to Haspelmath and Tadmor (2009) is the following (when the item is underlined it appears in the first part [concepts #1-100], when underlined in italics it appears in the second part [concepts #101-200] of the Swadesh list): To walk, you, yesterday, black, the back, the nose, the tongue, to kill, the rib, the eyelash, to go out, when?, long, the root, to hear, wide, to bring, I, to rise, today, the head house, the bone, this, the foot, the toe, few, to fart, the day after tomorrow, to stand, stinking, to blow, to listen, sometimes, up, behind, bright, to borrow, the clay, that, the day before yesterday, the itch, to hollow out, he/she/it, to flow, raw, the nit, the woman, the house, to go, the bark, to carry, the fire, to speak or talk, you (pl.), to meet, the wood, the night, to come, to throw, the flea, to lie down, to follow, red, new, the fog, there, the flesh, the sun, the ear, the lip, the knee, the breast, the navel, the liver, to cough, to spit, to bite, to sleep, the thunder, to shiver, here, far.
can be also used to efficiently measure contact (borrowings), iii) it points to the noise that can explain the inconsistencies we noted concerning CHAPTER 6, and iv) demonstrates that our attempt to measure borrowing between Turkic and Indo-Iranian languages, as an indirect measure of population contact and gene-flow, was not biased by the use of a word list that is markedly conservative (see CHAPTER 7). On a personal note, if a new wordlist, more ‘conservative’ than the Swadesh’s, had to be designed, many of the concepts that Haspelmath and Tadmor (2009) find resistant to borrowing are likely to be of difficult elicitation in fieldwork conditions, because ambiguous or difficult to be defined by the elicitor when the speakers of the language to be documented are predominantly monolingual (for example: ‘behind’, ‘the day before yesterday’, ‘the day after tomorrow’, ‘today’, ‘the house’, etc.). The book about wordlists is still open.

1.5 OUTLINE OF THE DISSERTATION

The order of the chapters corresponds to their focus. CHAPTERS 2 and 3 give introductory and methodological elements, CHAPTERS 4 and 5 report comparative studies involving surname and linguistic variability in the Netherlands and in Spain; CHAPTER 6 and 7 address wider linguistic contexts: Bantu and Central-Asian languages. The discussion follows.

1.5.1 CHAPTER 2: Sprachraum and Genetics (Manni 2010)

In this chapter I address the viewpoint of a geneticist with respect to genetic and linguistic cartography in order to provide an historical and methodological background reviewing the steps that led some population geneticists to co-operate with linguists, a collaboration that started with the comparison of maps. The mapping, in genetics, has a quite recent tradition, and is not very accurate because genetic sampling is generally uneven and relies, to a wide extent, on available published material. I review the first attempts to obtain maps about the genetic variability of populations, to later focus on the discontinuities in the genetic landscape, which are the “boundaries”, the barriers to gene flow (Manni et al. 2004). It is generally difficult to explain the geographical distribution of genes because genetic variability arises over time-scales that often are much deeper than historical times and because they depend not only on geographical features but also on cultural divides that have changed or disappeared without leaving traces. To provide evidence of some successful attempts to identify and explain
these barriers I present some results concerning the surname and linguistic variability in the Netherlands (see Chapter 4).

1.5.2 Chapter 3: Projecting Dialect Distances to Geography (Nerbonne et al. 2007)

This second methodological chapter addresses clustering instability in dialectology and concerns the application of the bootstrap method to linguistic data. When bootstrapping is impossible because original data is not available, an alternative approach consists in adding random noise to the distance (or similarity) matrices during repeated clustering: this is called noisy clustering. We demonstrate that noisy clustering can parallel a bootstrap test but it has a major weakness: it is impossible to know how much noise has to be added to emulate a given cut-off (see section 1.3.3.2). The article is quite short and is complemented by section 1.3 of this introduction. This work is important because Chapters 4-7 heavily rely on bootstrap trees to establish reliable classifications of dialects and languages.

1.5.3 Chapter 4: To What Extent are Surnames Words? (Manni et al. 2006)

Our focus in this paper is the analysis of surnames. We compare the distribution of surnames to the distribution of dialect pronunciations, which are clearly culturally transmitted. Because surnames, at the time of their introduction, were words subject to the same linguistic processes that otherwise result in dialect differences, one might expect their geographic distribution to be correlated with dialect pronunciation differences. In this paper we concentrate on the Netherlands where two official languages are spoken, Dutch and Frisian. We analyze 19,910 different surnames, sampled in 226 locations, and 125 different words, whose pronunciation was recorded in 252 sites. We find that, once the collinear effects of geography on both surname and cultural transmission are taken into account, there is no statistically significant association between the two, suggesting that surnames cannot be taken as a proxy for dialect variation, even though they can be safely used as a proxy for Y-chromosome genetic variation. We find the results historically and geographically insightful, hopefully leading to a deeper understanding of the role that local migrations and cultural diffusion play in surname and dialect diversity.

1.5.4 Chapter 5: Footprints of Middle Ages Kingdoms are Visible in the Surname and Linguistic Structure of Spain (Rodriguez Diaz et al. 2016)

To assess whether the present-day geographical variability of Spanish surnames mirrors historical phenomena at the time of the names’ introduction (13th - 16th century),
we have analyzed the frequency distribution of 33,753 unique surnames (tokens) occurring 51,419,788 times, according to the list of Spanish residents of the year 2008. From family-names we measured surname distances among the 47 mainland Spanish provinces (from which we infer consanguinity) and compared these distances to the relations among corresponding language varieties spoken in Spain. A dialectometric analysis of the first volume of the Linguistic Atlas of the Iberian peninsula (ALPI) started in 2009 in the laboratory of Hans Goebl (University of Salzburg, Austria). Original data have been analyzed by Goebl to identify phonetic, morphologic, syntactical and lexical features. Each feature has been processed separately in 375 working maps corresponding to 532 sampling points. From this we computed a final 47 x 47 similarity matrix (one sample point for each continental province) according to the relativer Identitätswert, (RIW, Goebl 2006). This index measures the similarity between two basilectal varieties as the percentage of items on which the two varieties agree.

The comparison of the two bootstrap consensus trees, accounting for surname and linguistic variability, suggests a similar picture; major clusters are located in the east (Aragón, Cataluña, Valencia), and in the north of the country (Asturias, Galicia, León). Remaining regions appear to be considerably homogeneous. We interpret this pattern as the long-lasting effect of the surname and linguistic normalization actively led by the Christian kingdoms of the north (Reigns of Castilla y León and Aragón) during and after the southwards reconquest (Reconquista) of the territories ruled by the Arabs from the 8th to the late 15th century, that is when surnames became transmitted in a fixed way and when Castilian linguistic varieties became increasingly prestigious and spread out. The geography of contemporary surname and linguistic variability of Spain in fact does correspond to the political geography at the end of the Middle Ages. The synchrony between surname adoption and the political and cultural effects of the Reconquista have permanently forged a Spanish identity that subsequent migrations, internal or external, did not deface.

1.5.5 Chapter 6: Linguistic Probes into the Bantu History of Gabon

In this extensive unpublished chapter we have compared the linguistic and genetic diversity of Gabon (Africa) in order to contribute new elements to the scenarios concerning the early Bantu expansion related to the adoption of agriculture. Two independently obtained datasets have been processed (Bastin et al. 1999; ALGAB—see Hombert 1990) accounting for a total of 126 different varieties consisting in Swadesh word lists. They lead to similar results, showing that the languages cluster into a comparable number of groups. The Levenshtein linguistic distances we computed are fully compatible with the classification of Grollemund et al. (2015) based on shared
vocabulary, where sharing is operationalized as the percentage of words (not) having the same historical origin. This coding is unnecessary with the Levenshtein method, making it simpler to use and, for the larger amount of information it accounts for, more sensitive.

We have tried to make the genetic dataset more representative of the 17 ethnic groups studied on the genetic side, by filtering-off all the DNA donors that were born outside the areas typically inhabited by their respective ethnolinguistic communities. The new results confirm the lack of genetic differentiation, which is even wider than previously observed. The linguistic cartography of our classifications shows well delimited areas that might be related to early waves of Bantu migrants that crossed Gabon in the early stages of their dispersal from Cameroon and Nigeria.

1.5.6 Chapter 7: A Central-Asian Language Survey (Mennecier et al. 2016)

In the frame of a large research project aimed at describing and comparing the genetic and social differences of sedentary and semi-nomadic populations living in Central Asia, we documented language varieties (either Turkic or Indo-Iranian) spoken in 23 test sites by 88 informants belonging to the major ethnic groups of Kyrgyzstan, Tajikistan and Uzbekistan (Karakalpaks, Kazaks, Kyrgyz, Tajiks, Uzbeks, Yaghnobis). The recorded linguistic material concerns 176 words of the extended Swadesh list. Phonological diversity is measured by the Levenshtein distance and displayed as a consensus bootstrap tree and as multidimensional scaling plots. Linguistic contact is measured as the number of borrowings, from one linguistic family into the other, according to a Precision/Recall analysis further validated by expert judgment.

Concerning Turkic languages, the results do not support regarding Kazakh and Karakalpak as distinct languages and indicate the existence of several distinct Karakalpak varieties. Kyrgyz and Uzbek, on the other hand, appear quite homogeneous. Among the Indo-Iranian languages, the distinction between Tajik and Yagnobi varieties is very clear-cut, despite the endangered status of the latter language whose speakers are in the process of being assimilated in the Tajik society.

More generally, the degree of borrowing is higher than average where language families are in contact in one of the many sorts of situations characterizing Central Asia: frequent bilingualism, shifting political boundaries, ethnic groups living outside the “mother” country. The latter case is of special interest because it systematically involves varieties that form clusters more coherent than those spoken inside the “mother” country (Kyrgyz of Uzbekistan vs. Kyrgyz of Kyrgyzstan, Tajiks of Uzbekistan vs. Tajiks of Tajikistan, Uzbeks of Tajikistan vs. Uzbeks of Uz-
We suggest that this phenomenon — that emigrant varieties are found to be more similar to one another — might be explained by $i)$ the use of a common set of borrowed words and, also, $ii)$ by a shared decreased exposure to a linguistic norm that is levelling them in the mother-country. All the trends we measured are attested by the availability of several speakers per village that allowed correcting for inter-individual variation.

1.5.7 Chapter 8: General conclusions and new prospects

By way of reflection on what has been learned, the final chapter provides a wider methodological discussion about the Levenshtein distance, discussion based on the empirical assays included in the dissertation and on what they show about its specificities in measuring linguistic difference. I will first review the findings about the relation between pronunciation differences and geographic distance, before suggesting a new line of investigation showing how residual Levenshtein distances can provide testable hypotheses about past linguistic convergence and divergence and, perhaps, addressing the influence that population growth and migrations have on linguistic variability. To do so, I will focus on family names: markers that enable the depiction of migrations occurred in historical times, that is, concerning European countries, in the last five centuries. Family names, appropriately processed, make it possible to distinguish the regions that received many immigrants from those that have remained demographically more isolated, aspects that underlie dialect and language contact. The last section develops a perspective from which we may examine the effects of migration on language change.
References:


