Summary of the Dissertation

This thesis in linguistics includes five published articles and one study to appear, in which I review, test and use computational linguistic methods to classify linguistic varieties consisting of lexical items — the sort of material that is generally readily available from linguistic atlases and databases.

A minimum of 90 items is used in all the empirical studies reported in the dissertation, because this number has been recognized to be sufficient for aggregate analyses. To compare linguistic varieties and classify them, two methods that lead to the computation of a linguistic distance matrix are used. (i) One compares two basilectal varieties as the percentage of phonetic, morphologic, syntactical and lexical features with respect to which the two varieties agree according to the relativer Identitätswert (RIW). (ii) The other one is based on the Levenshtein distance (edit distance), which accounts for the phonological differences between two varieties through string comparison (of phonetic transcriptions). Words with the same meaning are aligned and, to determine the extent to which the two strings differ from each other, the number of substitutions, insertions, and deletions that are necessary to change one string into the other are calculated. The RIW method is used in one section only, while the Levenshtein method is used in the remaining chapters.

The studies reported respectively concern the classification of Dutch varieties from the Netherlands; languages and dialects from Spain; Bantu varieties from Gabon, Tanzania and neighbouring countries and, finally, Turkic and Indo-Iranian languages spoken in Kyrgyzstan, Tajikistan and Uzbekistan.

The dissertation aims at probing human history by interpreting linguistic differences together with other sources of evidence. In a multidisciplinary perspective aimed at providing a higher level of anthropological synthesis, linguistic diversity is used as a proxy for the cultural differences of corresponding populations and is then compared to the variability of family names (their number, frequency and geographic distribution) or to genetic differences based on molecular markers on the DNA. The comparison has required methodological developments to assess the stability of linguistic classifications, in particular concerning the use of the bootstrap method and the use of residual Levenshtein distances as a way evaluate the linguistic similarity due to geographic contact as opposed to the remaining similarity, presumably determined by historical (genealogical) relatedness. The relation between pronunciation differences and geographic distance suggests a new line of investigation showing how residual Levenshtein distances provide testable hypotheses about past linguistic convergence and divergence.
and, perhaps, addressing the influence that population growth and migrations have on linguistic variability. With respect to the latter, the analysis of family names enables the depiction of migrations which have taken place in historical times, and, allows us to distinguish regions that have received many immigrants from those that have remained demographically more stable. We conjecture that such migration patterns have influenced dialect and language contact. This is a novel perspective from which we may examine the effects of migration on language change.

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. A concluding discussion follows.

Chapter 2 (“Sprachraum and Genetics”, 2010) addresses 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 historically started with the comparison of maps.

Chapter 3 (“Projecting Dialect Distances to Geography”, 2007) is about clustering instability in dialectology and the application of the bootstrap method to linguistic data. When bootstrapping is impossible because original data is not available, an alternative approach consisting in adding random noise to the distance (or similarity) matrices is described: this is called noisy clustering.

Chapter 4 (“To What Extent are Surnames Words?”, 2006) is about comparing the distribution of surnames to the distribution of dialect pronunciations, which are clearly culturally transmitted, in the Netherlands. 19,910 different surnames, sampled in 226 locations, and 125 different words, whose pronunciation was recorded in 252 sites are analyzed. 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 dialects and surname variability, suggesting that surnames cannot be taken as a proxy for dialect 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.

Chapter 5 (“Footprints of Middle Ages Kingdoms are Visible in the Surname and Linguistic Structure of Spain”, 2016) is aimed at assessing whether the present-day geographical variability of Spanish surnames mirrors historical phenomena at the time of the names’ introduction (13th - 16th century). From the frequency distribution of 33,753
unique family names, the surname distances among the 47 mainland Spanish provinces have been measured and compared to of the relations among corresponding language varieties as a dialectometric analysis of phonetic, morphological, syntactical and lexical features portray them. Surname and linguistic variability suggest 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 quite homogeneous. We interpret this pattern as the long-lasting effect of political and demographic phenomena related to the southwards “reconquest” (Reconquista) of the territories ruled by the Arabs from the 8th to the late 15th century.

Chapter 6 (“Linguistic probes into the Bantu history of Gabon”, unpublished) concerns the cross-comparison of the linguistic and genetic diversity of Gabon (Africa) in order to contribute new perspectives to the scenarios about the early Bantu expansion related to the adoption of agriculture some millennia ago. Two independently obtained datasets have been processed accounting for a total of 126 different varieties. They lead to similar results, showing that the languages cluster into similar groups. The Levenshtein linguistic distances are fully compatible with other classifications based on shared vocabulary, where sharing is operationalized as the percentage of words (not) having the same historical origin. While the alternative method requires that experts label cognate words in different varieties, this coding is unnecessary with the Levenshtein method, making it simpler to use and, due the larger amount of information it takes into account (all the sounds in the words), more sensitive. Genetic data indicate a lack of differentiation between populations at level higher 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.

Chapter 7 (“A Central-Asian linguistic survey”, 2016) is related to a large research project aimed at describing and comparing the genetic and social differences of sedentary and semi-nomadic populations living in Central Asia. Studied language varieties (either Turkic or Indo-Iranian) come from 23 test sites corresponding to the major ethnic groups of Kyrgyzstan, Tajikistan and Uzbekistan (Karakalpaks, Kazakhs, Kyrgyz, Tajiks, Uzbeks, Yaghnobis). The measure of the phonological diversity obtained by applying the Levenshtein distance has been paralleled by the measure of linguistic contact as proportional to the number of borrowings, from one linguistic family into the other, according to a Precision/Recall analysis 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 to be 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 into Tajik society.

CHAPTER 8 (“General conclusions and new prospects”) 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 as related to contact or historical linguistics. If in the past linguistics has driven a considerable number of hypotheses about the anthropological diversity of human populations, today demographers and geneticists can provide an accurate and large-scale quantification of the demographic processes leading to linguistic contact, setting a new framework to understand linguistic differentiation in a wider perspective that might referred to as POPULATION LINGUISTICS.