Linguistic probes into human history

Manni, Franz

General conclusions and new prospects

This chapter is unpublished, please cite it as follows:

Manni F. 2017. General conclusions and new prospects. In: *Linguistic probes into human history* (Chapter 8). PhD dissertation, Groningen dissertations in linguistics n° 162. ISBN 978-90-367-9872-3. Groningen: University of Groningen.

GENERAL CONCLUSIONS AND NEW PROSPECTS

The different chapters of the dissertation largely overlap in aims and methods, forming a coherent assemblage of empirical studies meant to shed light over the peopling phases of different countries and areas: *i)* CHAPTERS 4 and 5 concern the mapping and the classification of Dutch dialects and Spanish languages with respect to the surname differences of the two countries; *ii)* CHAPTER 6 is about the classification of the Bantu languages spoken in Gabon in connection to population genetics inquiry; *iii)* CHAPTER 7 addresses the classification of several languages spoken in Central Asia and the quantification of borrowing from each-other, with the aim to provide migration- and population-contact hypotheses that population geneticists can test. Methodological questions, inherent to every cross-disciplinary effort, have also been addressed in this thesis work, they concern the rationale for genetic and linguistic comparisons (see CHAPTER 2), the assessment of the robustness of linguistic classifications (CHAPTER 3) and, more generally (besides CHAPTER 5), the use of the Levenshtein distance to measure pronunciation distances. Each chapter ends with a specific *conclusions* section that remains valid today, some years after the publication of corresponding articles,[1] and there is no need to repeat here to what has been said.

Nonetheless, it is worthwhile to tie the various pieces together and to reflect on what we have learned. This final chapter provides a wider methodological discussion about the Levenshtein distance, because the empirical assays included in the dissertation enlighten about its specificities in measuring linguistic difference. This is the focus of Section 8.1 In Section 8.2 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.[2] Family names, appropriately processed, make possible to distinguish the regions that received many immigrants from those that have remained demographically more isolated, aspects that underlie dialect and language contact. Finally Section 8.3 develops a perspective from which we may examine the effects of migration on language change.

---

[1] Besides CHAPTER 6 that is unpublished.

[2] Surnames became fixed starting with the XVI century, when the Roman Catholic Church made compulsory, for every parish, to keep a register listing newborns and dead people.

8.1 THE ESSENCE OF THE LEVENSHTEIN DISTANCE

8.1.1 The Levenshtein distance and the feature system

To computationally measure the difference between two pronunciations Nerbonne *et al.* (1996) adopted the Levenshtein distance, an edit distance consisting in optimally aligning two text strings and counting the number of operations needed to pass from one string to the other; for example one step is required to go from 'ABA' to 'ACA', that is replacing B with C. This first implementation of the method was very simple, with 1/0 costs and no features describing segments. The results were very encouraging and, soon thereafter, Nerbonne and Heeringa (1997) based distance functions on binary features, as Gildea and Jurafsky (1996) had done. Phonetic segments were represented by binary vectors in which every entry stood for a single articulatory feature, thus enabling the distinction of a large number of phonetic segments. Later applications have tested other feature systems that are closer to the IPA features, especially concerning vowels. Other experiments concerned normalizing by the total length of the alignment, forbidding consonants to align with vowels,[3] and treating transpostions (swaps) in a special manner The calculations reported in CHAPTER 4 are based on a unit-cost model normalized by the length of the alignment, while those concerning CHAPTERS 6 and 7 are based on gradual segmental distances not normalized by the length (see Tab. 1).

The issue of segmental similarity was a major focus of the work in Groningen for about ten years, culminating in Heeringa (2004) which devotes one 52-page chapter to the question of how to measure similarity of two phonetic segments by comparing three phonetic feature systems: first, a 7-feature system borrowed from 'The Sound Pattern of English'; second, the ones of Vieregge *et al.* (1984) and Cucchiarini (1993) respectively for vowels and consonants; and third, a system developed for phonetic segment characterisation by Almeida and Braun (1985, 1986). The more sensitive representations failed to lead to significant improvements when measured against the perception judgments of dialect speakers (Heeringa 2004, p. 186) in the frame of the

[3] To deal with syllabicity, the Levenshtein algorithm may be adapted so that only vowels may match with vowels, and consonants with consonants, with several exceptions: [j] and [w] may match with both consonants and vowels, [i] and [u] with both vowels and consonants, and central vowels with both vowels and sonorant consonants. So the [i], [u], [j] and [w] align with anything, the [ə] with syllabic (sonorant) consonants, but otherwise vowels align with vowels and consonants with consonants. In this way unlikely matches (e.g., a [p] with an [a]) are prevented. This approach was first applied to Sardinian dialects (Bolognesi and Heeringa 2002), and then to Dutch (Heeringa and Braun 2003) in a validation study.

aggregate comparison of full dialect varieties. Other distances exist when the task is to focus with very high precision on the phonological difference of single words (see next section).

Table 8.1 ▸ Alignment of the word 'rabbit' with a unit-cost model (*Kaninchen* in German and *konijn* in Dutch). Five differences are found over a total length of the alignment of nine positions. When a normalization by the length is applied a difference of 5/9 = 0.55 % is found. From Heeringa (2004), p. 131.

k a n i n ç ə n	German	k a n . i n ç ə n
k o n ɛ i n	Dutch	k o n ɛ i n . . .
		0 1 0 1 0 0 1 1 1 = 5
		1 2 3 4 5 6 7 8 9

8.1.2 The Levenshtein distance measures intelligibility

Given the large use that is made of the Levenshtein method in the dissertation (see also CHAPTER 1), reviewing the discussion it has stimulated in the literature seems appropriate here. Although some of the research included in the dissertation was conducted before certain criticisms were levelled, they cannot be ignored now.

First it should be made clear that the Levenshtein algorithm performed very well concerning the comparison of dialect varieties consisting of quite large wordlists. But for the purpose of identifying cognates, or for the purpose of recognizing loan words, or for the purpose of identifying sound correspondences, more sensitive measures were designed. They rely on how well the measure works per word rather than on sets of words and find application in computational phonology, including dialectometry (Covington 1996; Somers 1998; Gildea and Jurafsky 1996, Kessler 1995, Oakes 2000 -- See Kondrak 2003 and Kondrak and Sherif 2006 for a review). Among them, Kondrak's algorithms (2002) (COGIT, ALINE) gained renown in historical linguistics. While COGIT is meant to automatically identifying cognate words, ALINE is similar to the Levenshtein distance but it incorporates the notion of phonetic coalescence and break-up, and it uses a non-binary feature system. If the author does not compare his algorithm to the Levenshtein method (and to the other existing ones, besides Covington's (1996)), this is probably because the application is different, mainly concerning language reconstruction, whereas the Levenshtein approach was initially used to identify *language areas*. It would be interesting to see the degree of improvement that Kondrak's alterations would add to the performance of the Levenshtein method.

Explicit criticisms of the use of the Levenshtein approach were expressed by McMahon and McMahon (2005)[4] and Greenhill (2011). The latter author described the approach as blind, unable to distinguish chance similarities from real cognates and lying at the linguistic "surface", suggesting that better methods should be adopted to classify language varieties: a criticism that, from his standpoint, could have been le-veled at the use of a majority of the phonological distances cited. While Greenhill (2011) conceded that Levenshtein distance worked well at low time depths, he suggested to use, instead, adaptive algorithms that learn the transition weights through the application of naïve Bayesian classifiers or stochastic transducers. The first ones are generally used for natural language processing tasks and have encountered great success in the subfield of authorship analyses, the process of attributing the author of an anonymous text according to its writing characteristics (Juola *et al.* 2006). Stochastic transducers, popular in automatic translation, have been applied to dialectology by Wieling *et al.* (2007a) and Scherrer (2007). Scherrer's work is about comparing stochastic transducers with the Levenshtein distance, which was used as a baseline for experiments of bilingual lexical induction between the German-Swiss dialect of Bern and standard German. Although less efficient, the Levenshtein distance performed well, as it had been shown by Mann and Yarowsky (2001) that induced over 90% of the Spanish-Portuguese cognate vocabulary with a 11% F-measure improvement over the Levenshtein algorithm; however in several cases stochastic transducers offered no improvement over the Levenshtein distance. This result is in agreement with the findings of Wieling *et al.* (2007a) showing that the segment distances induced correlate well with distances in formant space.

Besides the efforts to improve the sensitivity of Levenshtein distance using features taken from phonetics and phonology, data-driven techniques have also been applied for this same purpose. Wieling *et al.* (2012) used an iterative technique to induce segment distances. They first applied the Levenshtein algorithm to a large da-taset of dialect pronunciations, collecting first all the alignments. From these, they extracted the segment correspondences and their frequencies. They then collected all these in a large contingency table from which they could calculate an information-theoretic measure of

[4] See the reply of John Nerbonne (2005) where he reviews a large number of methodological misunderstandings, by the McMahons, about the way the Levenshtein distance is computed. McMahon and McMahon (2005) criticise that Nerbonne and Heeringa's (1997, p. 11) *"earlier work calculated edit distance in the simplest possible way meaning that the pair [a,t] count as different to the same degree as* [a, *]"*. But in fact the paper they cite focuses on how to differentiate such sounds more subtly, exploiting phonetic and phonological features for this purpose.

the affinity of pairs of correspondences called *pointwise mutual information* (PMI).[5] Finally, they used (an inversion of) the PMI values as substitution costs in the following iteration of the procedure. Experiments with several datasets showed that the procedure stabilized within ten or fewer iterations. An evaluation based on alignment accuracy confirmed that the PMI-based version of Levenshtein algorithm reduced error only slightly (from 3% error to about 2.5%) with respect to the classical method. Jäger (2015) used a simplified version of this algorithm to avoid the need for expert judgments on cognacy in historical linguistics, showing that the results were confirmed by Glottlog classifications (Hammarström *et al.* 2016).

It has been said that the Levenshtein method was not originally introduced to linguistics for the task of identifying cognates, meaning that the distances it yields are not to be taken as an estimation of the divergence time between language varieties, and suggesting that the criticisms of Greenhill (2011) and McMahon and McMahon (2005) discussing its use for this purpose were hasty. While it would be interesting to compare some of the more sophisticated versions of the Levenshtein distance with other algorithms, that must be taken as a note for future work.

Actually, dialectology and historical linguistics have divergent aims. In the first the focus is on the overall similarity of entire varieties, whereas in the latter the attention goes to the identification of cognates and to the measure of the similarity of individual words. The primary use of the Levenshtein approach has been to seek and identify the signal of geographic provenance in dialect speech, while historical linguistics addresses a signal of historical "relatedness" at the level of variety and at the level of individual words. Another important difference between the two is their relation to geography, which influences the distribution of dialectal varieties massively, but not necessarily discretely. Heeringa and Nerbonne (2001) have shown how the dialectal analysis provides an analytical foundation for the notion "dialect *continuum*", where the classification into discrete groups, the very heart of phylogenetic analysis and historical classifications, plays no role. Because they were obtained outside an explicit historical linguistics frame, McMahon and McMahon (2005) judged Levenshtein classifications "to a great extent uncorroborated" (p.213) basing their judgment on expected methodological drawbacks that, in reality, turned far from true because, when the technique has been applied to Dutch, German, American English, Sardinian, Norwegian, Bulgarian and Catalan, the groupings provided enjoyed the recognition of specialists in these dialects and languages. We review formal validation efforts below.

[5] Point-wise mutual information (PMI), or point mutual information (Fano 1961), is a measure of association used in information theory, statistics and computational linguistics (see for example Church and Hanks 1990).

Concerning dialectology and Catalan varieties, the classificatory effectiveness of the Levenshtein algorithm has been tested in comparison to another computational method, the *m*COD (*Méthode* COD) (Clua *et al.* 2008; Clua and Lloret 2015), that embraces the study of linguistic variation in the areas of phonetics, phonology and inflection. The *m*COD approach differs from other dialectometric analyses in the fact that these are quantitatively surface-oriented, while the *m*COD was designed to capture the differences among varieties not only quantitatively but also qualitatively, in order to increase the accuracy of the groupings. The distances obtained with two methods (*m*COD *vs.* Levenshtein) correlate very highly ($r = 0.868$) and converge in identifying the same borders between dialect areas, the differences between the two methods concerning specific details (Valls *et al.* 2012).

To remind the extensive work conducted in Groningen (see Heeringa 2004, Nerbonne and Heeringa 2010) prior to an extensive application of the Levenshtein method, the classifications have been tested quantitatively by *i)* addressing the sensitivity of the measure to segment order and to phonological context, *ii)* taking into account the (non-) use of length normalisation, *iii)* testing the linguistic constraint that all alignments respect the consonant/vowel distinction but also by *iv)* measuring their (good) match with the overall similarity-judgments of dialect speakers (Gooskens and Heeringa 2004,[6] Gooskens and Heeringa 2006, Heeringa *et al.* 2006) and in comparison to native speakers' judgments of accent strength (Wieling *et al.* 2014). Inspired by the latter research direction Fontan *et al.* (2015) have successfully applied the Levenshtein method to measure intelligibility in a project concerning the tuning of hearing-aids, therefore setting up automatic measures of speech intelligibility for the recognition of isolated words and sentences, similarly to Sanders and Chin (2009) that found the Levenshtein method to correlate extremely well (r = 0.925**) with naïve human transcriptions of the speech of pediatric cochlear implant users. This is why, the Levenshtein distance can be seen as a good measure of intelligibility, that is the perception a speaker has of the linguistic difference of someone else's speech (Beijering *et al.* 2008).

[6] About Norwegian dialects, Gooskens and Heeringa (2004) report that perceived linguistic differences correlate at ~0.8 with measured Levenshtein distances, and the correlation would probably be higher if all the (naïve) speakers had a same (high) level of linguistic competence in assessing the varieties located more distantly that their neighbourhood. In fact, the perception of linguistic differences is finer within the radius of human interaction and decreases outside it, meaning that the speakers are less familiar with distant varieties that they tend to classify as "very different", whatever the real geographic or linguistic distance is.

8.1.3 The Levenshtein distance measures contact

To summarize, the Levenshtein algorithm has been criticized for not distinguishing cognates from chance similarities at great time depths, and for measuring the linguistic "surface", which is perhaps why it correlates well with the individual perception.[7] This latter aspect is essential in the frame of the research presented in this dissertation, which is work addressing the cultural proximity of human populations with respect to their ancestry, inferred through genetic markers or family names. By comparing population genetics data to cultural differences measured through linguistic diversity, we are adding detail to the same research question: *describing and explaining how people interact and mate*. Human mating is influenced by several factors, the first one being the chance to meet (which depends on geographical proximity and social stratification) but also, to a large extent, the perceived attractiveness of the partner. This is where cultural differences play a significant role, together with economic considerations, traditional rules of descent, taboos, *etc*. The speech conveys information about the geographical origin of the partner, about the social status of the family, about education, *etc*. Our speech plays a significant role, conscious and unconscious, in the feeling we have about the possible mutual understanding *lato sensu* with a new partner. When we speak to someone we are not mentally counting the number of shared cognates or borrowed words we both employ, instead we perceive, intuitively and very rapidly, the extent to which her/his speech is close to ours; and a sentiment of closeness or distance can arise, leading to a stronger or weaker desire to interact. The speech is also evocative of many preconceived opinions we have about the others, they arise from beliefs, traditions and history. A measure able to capture *perceived* linguistic proximity, like the Levenshtein distance does, is useful in the context of cross-disciplinary research involving population genetic and demographic inquiry. However, phonological differences do not concern only the linguistic surface. In an example about Dutch dialects, it has been shown that they correlate with syntactic differences (Spruit *et al.* 2009), meaning that they reflect a deeper level of the languages, showing how linguistic levels arise from a similar pattern of historical geographic contact.[8] This finding has

[7] But the book is not closed on the suitability of modified Levenshtein algorithms for cognate recogntion. T. Rama *et al.* (in preparation.) reject Greenhill's criticism forcefully, concluding that "PMI [-based Levenshtein] systems yield […] better accuracies than current state-of-the-art systems." (personal communication to J. Nerbonne, 2017)

[8] Spruit *et al.* (2009) find that pronunciation is associated with syntax and lexis, while syntax and lexis are only weakly associated. The main cause is that pronunciation and syntax are both strongly associated with geography, while lexis is not. When geographic distance is controlled for, as the underlying factor, the association between pronunciation and syntax remains but weakens considerably, while the association between syntax and lexis disappears.

special relevance as it contradicts the notion that (morpho)syntax is hostile to geographic diffusion and suggests that similar mechanisms of diffusion apply to grammar, syntax and pronunciation (Szmrecsanyi 2013, p. 159).

8.1.4 The Levenshtein distance measures historical divergence

Besides chance similarities, two cognate words generally result in the alignment of strings that are similar, meaning that there are fewer differences between related words (*Kaninchen/konijn; konijn/coniglio*) than between words that are unrelated (*Kaninchen/lapin; konijn/*'rabbit'). Given that the less variable part of a set of related words generally concerns the left part of the alignment (*Kaninchen, konijn, coniglio*), Kondrak (2003) suggested that the Levenshtein distance overestimates the differences between historically related words because all the segments in a word, located on the left or on the right, equally contribute to the measured distance. When the words under scrutiny have significantly diverged, the corresponding pairwise distances account less for their common origin and more for their divergence through time and space, as correctly noted by Greenhill (2011). A strict cognate-based method would not classify *Kaninchen*, *konijn*, *coniglio* as "the same word = zero difference", even though they all come from the Latin word *cuniculus*, because the Dutch and German words are borrowing French. A Levenshtein method would note their similarity, which exceeds that of chance words. This is to say that the Levenshtein distance captures, at the same time, a part of the historical signal that words convey by being cognates (or not), perhaps via borrowing, *and* the signal related to the phonological change that occurred after the separation, like linguistic divergence or contact; it captures everything related to similarity, agnostically.

Linguistic contact is one of the main reasons explaining the generally observed high degree of correlation between Levenshtein distances and corresponding geographic distances. Two varieties can become increasingly similar through extensive borrowing, to the point that their original difference (historical) is obscured. Since the degree of borrowing and the intensity of communication is proportional to geographic proximity, I have experimented with residual Levenshtein distances in order to see if the historical signal would be emphasized after correcting Levenshtein measurements by controlling for the language contact related to geographic proximity (see section 8.2.2).

While the Levenshtein distance measures the signal of historical relatedness *and* the contact between the languages, its ability to match classifications based on shared cognates identified by the comparative method is much higher than the criticisms mentioned above would suggest. In CHAPTER 6 the very good match between

the clusters identified by Grollemund *et al.* 2015 and the corresponding Levenshtein classifications has been reported (Fig. 6.25). This result might be explained by the fact that Bantu languages are linguistically quite close, often forming dialect-chains: this is a scenario closer to the initial application of the Levenshtein method to dialectology. Nevertheless, when the Levenshtein classification of six Indo-Iranian and Turkic Central Asian languages (Tajik, Yaghnobi and Kazakh, Karakalpak, Kyrgyz, Uzbek, respectively), described in CHAPTER 7, is compared to a tree only accounting for shared cognates, the two representations overlap well again (Fig. 8.1), without discrepancies, suggesting that the Levenshtein distance, after all, normally captures the *same* historical signal that a cognate-based approach does, additionally delivering sub-clusters that are less capricious.

Jäger (2015) presents a very promising application of a modified Levenshtein algorithm (Needleman-Wunsch)[9] to the problem of detecting historical relatedness. Jäger notes *inter alia* that the application of the Levenshtein method does not require that language family experts first annotate all of the data to indicate which words are cognates, as the classificatory experiments of the Bantu languages of CHAPTER 6 also indicate. It is therefore much more widely applicable.

In addition to the efforts to adapt the Levenshtein method to the task of identifying cognates carried on by other groups, future research practices might be based on the comparative use of *both* methods, appropriate versions of the Levenshtein distance *versus* standard cognate-based classifications. The degree of their divergence, when there is one, is likely to be proportional to a wide panel of effects, linguistic contact and convergence being the first candidates. The discrepancies between Levenshtein and cognate-based classifications, instead of being reported as flaws, could be investigated as clues able to shed light over demographic, sociolinguistic and geographic phenomena that pair off with the verticality of the linguistic transmission.

After 20 years of hectic research focused on the application of computational methods on understanding linguistic variability, the moment has come to recognize that more research has been devoted to methods better able to establish a correct phylogeny of the languages than to approaches able to explain how the contact of the speakers took place and with which consequences: the Levenshtein distance is certainly one of those.

[9] The Needleman-Wunsch (1970) method is a dynamic programming algorithm for global sequence alignment, a technique particularly appropriate when sequences are of a same length; it finds application in many aspects of computer science.

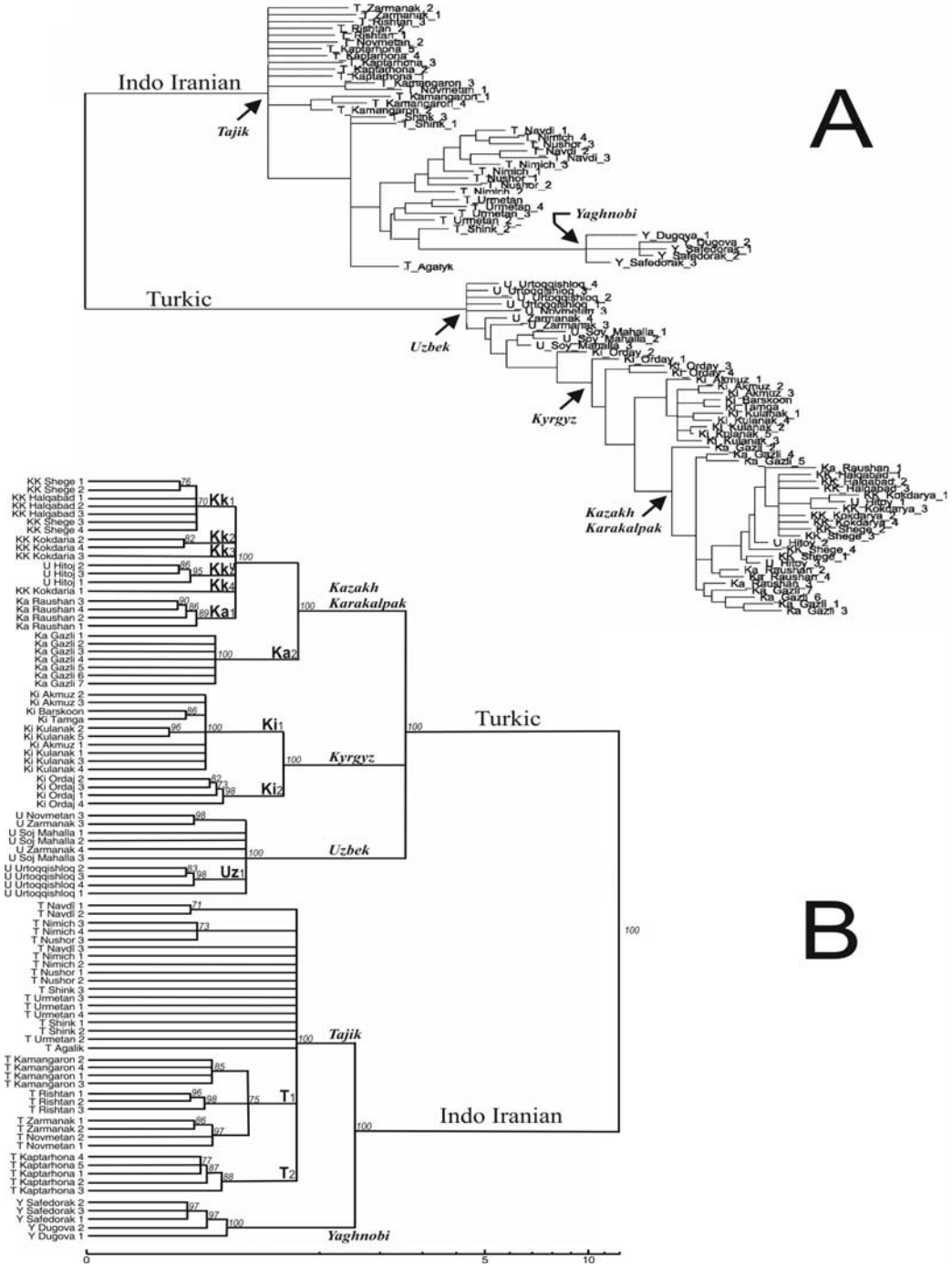

Figure 8.1 ▸ **Linguistic similarities and differences between 88 informants interviewed in 23 test sites as in CHAPTER 7**. The two trees show a very similar topology. *A) Expert cognacy judgments*. Cladistic majority-rule consensus tree obtained from 100 bootstrap trees. Branch lengths are posterior estimations. The nodes supported by less than 70% of the trees have been collapsed. Tree length = 1596; Consistency index = 0.4518; Homoplasy Index = 0.5482; Retention index = 0.8699; Rescaled consistency index = 0.3930. Courtesy of Pierre Darlu (National Museum of Natural History, Paris). *B) Levenshtein distance***.** Consensus tree obtained from 100 UPGMA bootstrap trees (Fig. 7.2 in CHAPTER 7). The nodes supported by less than 70% of the trees have been collapsed.

8.2 CURRENT CHALLENGE: GOING BEYOND GEOGRAPHY

In this and the following sections we turn to challenges we are now better prepared to face on the basis of work in this dissertation. Given the result that Levenshtein distance is a valid indicator of aggregate similarity, and that it reflects the influence of separation among linguistic varieties, we may ask what other factors influence. For this purpose we explore an examination of residual Levenshtein distances here.

8.2.1 The spread of linguistic innovations

It has long been admitted in linguistics that the geographic distance between varieties has an effect on their evolution, namely that closer varieties are generally more similar than distant ones. The first model about the spread of linguistic innovations probably was the WAVE THEORY of Johannes Schmidt (1872). However, a mathematical modelling of this empirical evidence came only a century later, when Séguy (1971) started to develop computational dialectology. In a similar vein, a theoretical formalization of the phenomenon states that the similarity of dialects is a function of both the geographic proximity and the population size of speech communities. Like the WAVE THEORY, Trudgill's (1974) GRAVITY MODEL explains the spread of linguistic innovations as a radiation from a centre but, and this is the novel aspect, one which has an effect on larger centres at first, and then spreading to the smaller ones in a cascade of effects (Labov 2001, p. 285) depending on the population size, that is the frequency of linguistic contact.[10] When contact occurs, the speakers are influenced by one another's speech and modify their own, sometimes adopting innovations (Lewis 1979). This model does not take into account geographic features that are likely to increase or decrease the contact (rivers, deserts, mountains, etc.) or social strata, different levels of economic attractiveness or other factors such as the perceived prestige of a given variety, probably because the model was primarily meant to formalize theoretical thoughts, rather than providing explicit clues about the way to empirically test it.

The spread of innovations is a central aspect in historical- and socio-linguistics and it is gaining increasing attention (see Eisenstein *et al*. 2014), because today's social networks enable the measure of the spread of innovations in real time and make it possible to follow their geographical directions. However interesting, such investigation deals with technologies enabling very easy contact between the speakers (that can remain virtual to each other) with the consequence that the present-day spread of linguistic innovations might actually deviate from the neighbourhood dynamics that

[10] The gravity model assumes that populations are sedentary.

Labov and Trudgill assume in their modelling. We note, however, that Eisenstein *et al.* (2014) definitely find locality effects in Twitter data even though it is a medium allowing world-wide contact. While Seguy (1971) presented dialect distances as function of the square root of geographic distance, Trudgill (1974) suggested that the spread of innovations declines quadratically. Nerbonne and Heeringa (2007) and Nerbonne (2010) found a logarithmic model to better function, similarly to the models of population genetics concerning the biological differences of neighbouring populations that are function of migration processes.

The mathematical relations between genetic and geographic distances have long been addressed and Wright (1943) postulated ISOLATION BY DISTANCE (IBD), a model explaining that the genetic similarity of human groups decreases with their geographic distance with reference to spatially limited gene flow, a frequent phenomenon in natural populations. Gustave Malécot (1948) pushed this analysis onwards by establishing that the increase is not linear but logarithmic, and this is what is currently found with surname studies addressing the diversity of local populations (see Scapoli *et al.* 2007 for European case-studies). The agreement existing in linguistics and population genetics about the exponential decay of human interaction as a function of the geographic distance is probably *not* due to a chance similarity. The easiest and most logical explanation is to admit that the speakers interact in a neighbourhood that leads, at the same time, to the dissemination of linguistic innovations and offspring.

8.2.2 Levenshtein residual distances

At this point it is interesting to speculate on future directions about the research direction illustrated in the previous sections. When a matrix of aggregate linguistic distances, such as those produced by the Levenshtein algorithm, is found to be significantly correlated with the corresponding matrix of geographic distances, it is possible to compute a regression (*linguistic distance* vs. log [*geographic distance*]) in order to compute, from it and for each paiwise comparison, the linguistic distance that is *expected* according to the geographic distance. This procedure leads to a matrix of *expected* pairwise linguistic distances that can be *subtracted* from the linguistic distances obtained from the original data. The matrix that results after the subtraction consists in residual distances that can be positive, negative or null. They will be positive when the linguistic distance computed on original data is higher than the one expected from the regression; they will be negative when two localities exhibit a linguistic distance that is lower than what is expected according to the regression. The idea is that residual distances account for the fraction of the linguistic variability that is *not* explained

by normal linguistic contact between neighbours, in fact residuals correspond to the virtual case in which all the populations would be located at the same geographic distance one from each other.[11]

The matrix of original distances can be compared to the matrix of residual distances, once they are both represented as multidimensional projections or as trees. The differences between the two representations are likely to correspond to the relations among the varieties before they drifted apart due to geographic remoteness or to convergence/divergence phenomena related to contact with other varieties, operationalized as geographic distance. In the two representations, the differences of single linguistic varieties with respect to the others should be considered with caution, because their different positions rely on a geographic correction (the regression model) that arises by taking into account the *whole* set of pairwise distances, while single varieties might be affected by specific phenomena that the general model of regression does not take into account. In fact, some subsets of the entire dataset, when they are taken separately, can lead to a regression having a different slope (Simpson's paradox), meaning that the computation of residuals according to the full dataset is not optimal for every sample, because it does not take into account local geographic phenomena. The regression is correct on the whole, but not necessarily in its details. For this reason a matrix of residuals provides tendencies that should be interpreted generally, that is in terms of clusters, to answer questions like the following ones: *Are the clusters appearing in the projection based on residual distances the same ones as those that the matrix of original linguistic distances delivers? Does the relative distance between the clusters change from one plot to the other?* To explain the usefulness of residual distances, three empirical examples taken from datasets analyzed in this dissertation will be analyzed in this way.

8.2.2.1 The Netherlands

This first example concerns the Dutch dialect areas presented in CHAPTER 4. The matrix of residuals reveals two aspects that the "regular" Levenshtein matrix does not show: *i)* hidden structures in the phonology of the province of Groningen that still testify to proximity with the Frisian varieties that used to be spoken in the province of

[11] Controlling for geographic distance is very easy with distance-based methods, like the Levenshtein, but is not readily applicable to multi-character-based methods, like cladistics ones for example. To do so, it is necessary to obtain a distance matrix from the phylogenetic tree by computing, for all the *taxa*, patristic distances on tree branches and, then, to establish regressions between linguistic and kilometric distances. This routine does not seem to have been applied frequently in the literature.

Groningen some centuries ago and which have gradually been replaced by the language of the city of Groningen (Lower Saxon),[12] and *ii)* a transcribers' barrier in the southern part of the country (see Fig. 4.7 in CHAPTER 4).[13]

8.2.2.2 Tanzania

When the regression method is applied to the dataset of Bantu languages from Tanzania (see Fig. 6.11 and section 6.3.1.1 in CHAPTER 6) and residual linguistic distances are compared to the original ones, the topologies of corresponding projections match but the clusters occupy different surfaces (see Fig. 8.2).

With residuals the cluster {E50, E60} becomes more compact and closer to other varieties, while the group {F20, M10, M20, M30} less so. The cluster {N10, P10, P20, G50} remains stable. By definition, the topology of the samples portrayed by residual distances is not linked to the linguistic contact between neighbouring varieties (geography is controlled for), meaning that the expansion/contraction of clusters is explained by other factors, probably historical.[14]

A working hypothesis to be tested could be that the first Bantu speakers that settled in Tanzania spoke varieties that, while divergent, did not belong to clearly identified separate groups. Once the speakers became sedentary, differential linguistic contact between the Bantu immigrants led to phenomena of linguistic convergence in some areas (F20, M10, M20, M30) but not in other ones (E50, E60). Methodologically it is interesting to see that the stress values of multidimensional representations of residuals' matrices are generally considerably higher than those of original distances. This is a clear indication that residuals' variability is not easily represented in few dimensions, differently from data that have been shaped by linguistic contact happening in the two dimensions that geography allows.

8.2.2.3 Gabon

Concerning the ALGAB dataset (see section 6.3.1.2 in CHAPTER 6), the residuals distances have been computed after a linear regression (R^2=0.216; the logarithmic transformation of geographical distances makes almost no difference: R^2=0.222). They form clusters that correspond well to those obtained by plotting original distances, but give rise to groups that are more evenly dispersed and better distinguished (Fig. 8.3, Table 8.2).

[12] Note that the distinction between Friesland and Groningen is quite strong in pronunciation and in lexis, and much weak in syntax. Compare Srpuit et al.'s (2009) Figures 6 and 7 (pronunciation and lexis, respectively) on the one hand and Figure 8 (syntax) on the other.
[13] See Mathussek (2016) for examples about computationally inferred transcribers' borders.
[14] No transcribers' borders here as each variety has been transcribed by a different person.

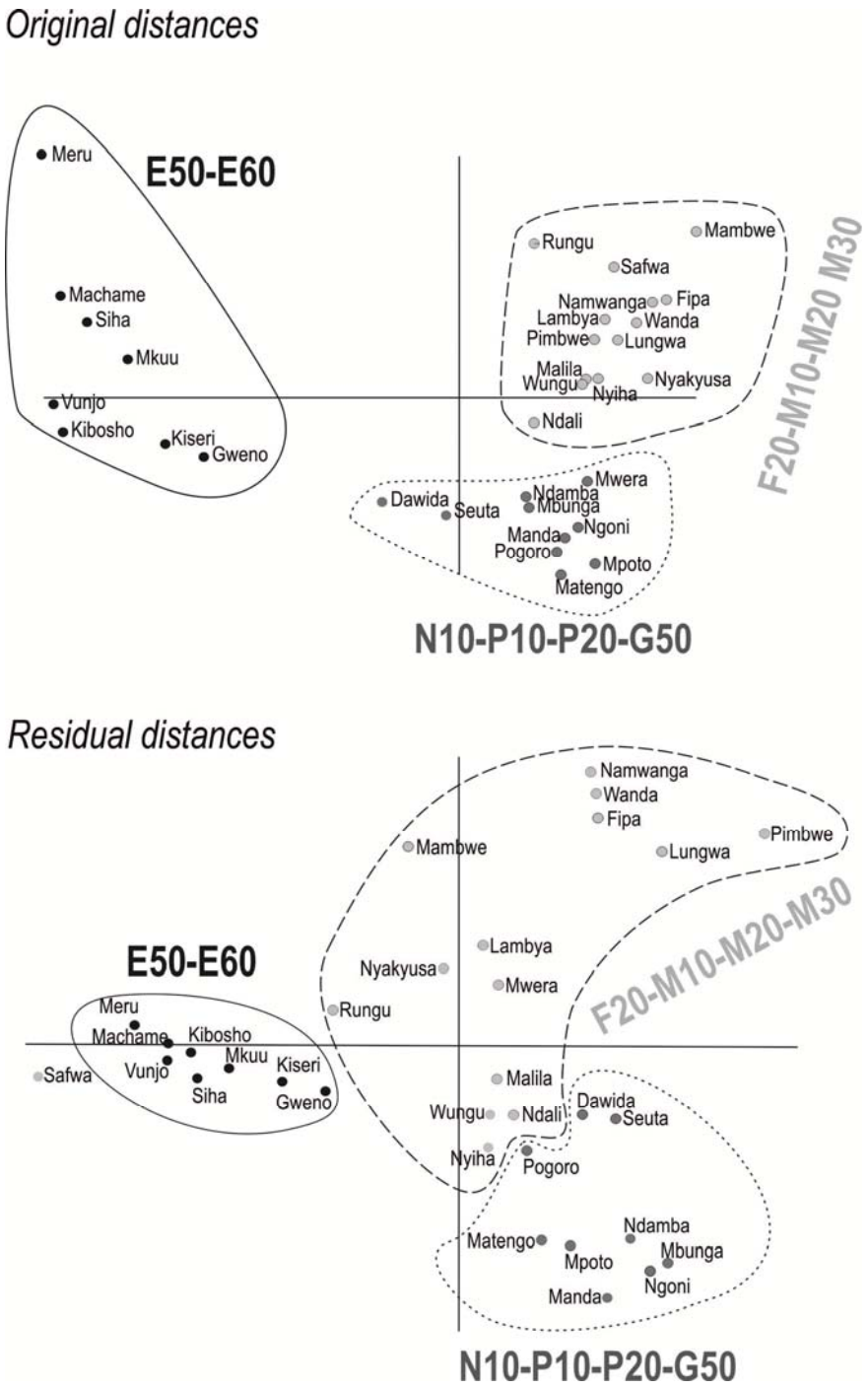

Figure 8.2 *(see also Fig. 6.11 CHAPTER 6)* ▸ **Multidimensional scaling plots concerning 32 Tanzanian languages and 1052 concepts**. *Left:* Projection based on original gradual segmental Levenshtein distance. Stress values: in 1 dimension = 0.1588, in 2 dim. = 0.0966 (plot reported), in 3 dim. = 0.0688. Correlation between geographic and linguistic distance = 0.7. *Right:* Projection based on residual distances after computing the regression (R^2 = 0.4983) between the logarithm of kilometric distances and the corresponding Levenshtein distances. Stress values: in 1 dimension = 0.3779, in 2 dim. = 0.2471 (plot reported), in 3 dim. = 0.1885. Residuals are normally distributed.

Residual distances convey very interesting clues about the possible historical scenario of linguistic diversification of the Bantu varieties in Gabon, a setting that is quite hard to interpret (see section 6.4.5.3 in CHAPTER 6). They point to a certain amount of linguistic diversity between different languages that long-lasting linguistic contact and convergence has progressively defaced (see Table 8.2 for a summary). With residual distances the varieties B50, B60, B70, spoken in a more densely inhabited area, remain close to each other, but we see changes in their relative distances from the groups corresponding to the languages classified as B40 and B20. Further, with residual distances the group B20 is resolved in two separate clusters (corresponding to the two subclusters reported in the bootstrap tree of Fig. 6.17 in CHAPTER 6: {B20-I, B20-II} and {B20-III, B20-IV}), suggesting that varieties B20 are not genealogically related.

Interestingly, the relative distance between the clusters B10 and B30 increases from one plot to the other (Fig. 8.3), recalling the debate about their possible convergence after a separate phylogenetic origin (Nurse and Philippson 2003). The fact that residual distances put the Fang languages (A75) much closer to the group B40 than measured Levenshtein distances do is also intriguing, and can be related to a similar geographic provenance.

Table 8.2 ▸ Summary of the possible effects on the Bantu varieties from Gabon that linguistic contact has determined (ALGAB data, see section 6.2.1.2 in CHAPTER 6)**.** This scenario is inferred by comparing the 2 plots of Fig. 8.3.

Varieties	*With reference to the initial stage (inferred by residuals) the* Varieties *later…*
B10	Converged with B30
B30	Converged with B10
B20	Two initially separate clusters converged together
B40	Converged with B50
B50	Diverged slightly from B60/B70 becoming closer to B40 and some varieties B20
B60/B70	Diverged slightly from B50
A75	A75 arrived in Gabon recently (~5 centuries ago), when B40 was already spoken. Their closeness (residuals' plot) might correspond to a similar geographic origin in Cameroon: but today they are very different.

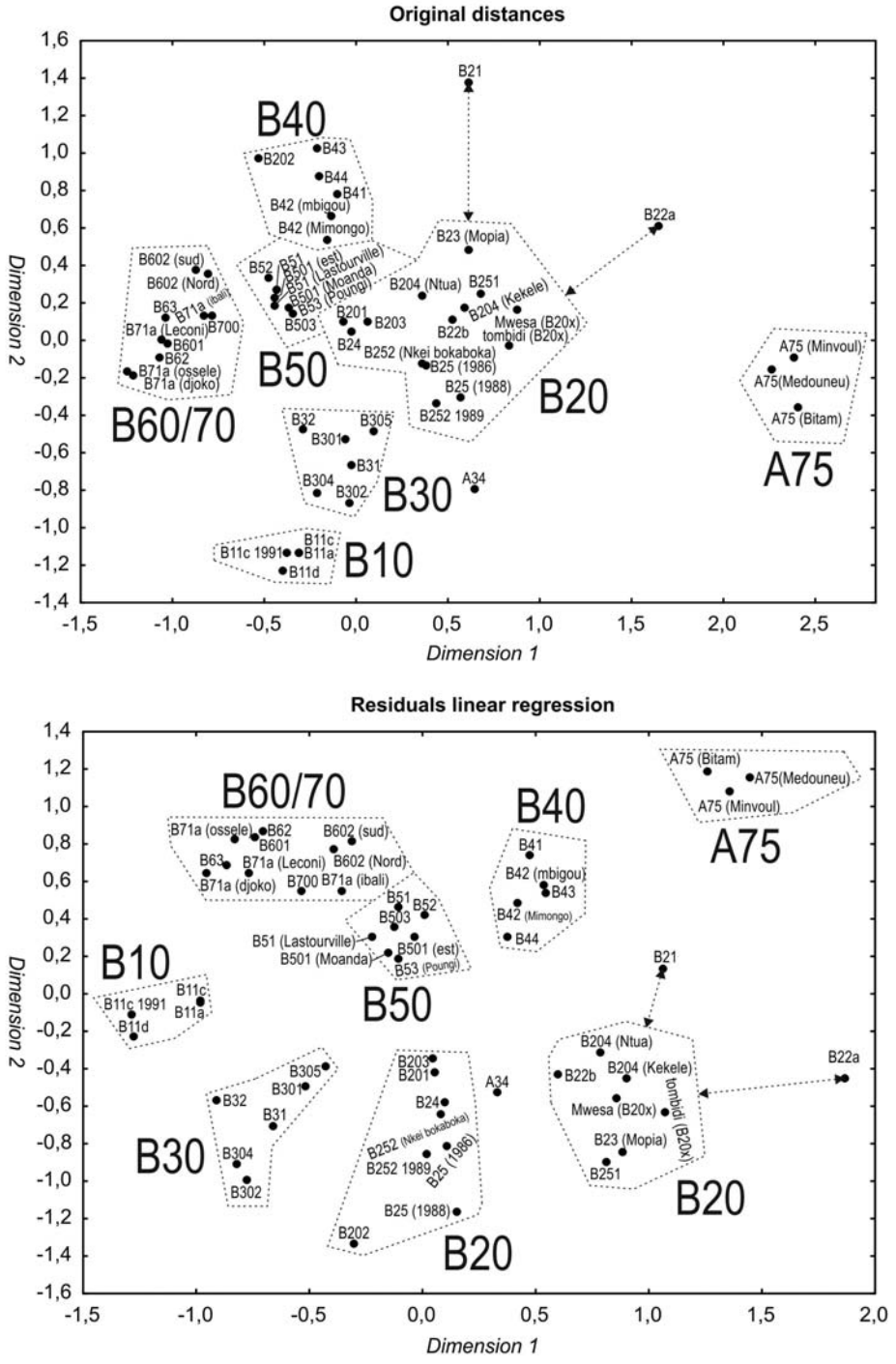

Figure 8.3 ▸ Multidimensional scaling projections concerning 53 languages from the Linguistic Atlas of Gabon (ALGAB). *Top:* Original Levenshtein distances. Stress values: in 1 dimension = 0.3247, in 2 dim. = 0.1641 (plot reported), in 3 dim. = 0.1215 (plot shown in Fig. 6.18 in CHAPTER 6). Correlation between geographic and linguistic distances = 0.478. *Bottom:* Residual distances after computing the regression ($R^2 = 0.216$) between the kilometric distances and the corresponding Levenshtein distances. Stress values: in 1 dimension = 0.399, in 2 dim. = 0.249 (plot reported), in 3 dim. = 0.171. Residuals are normally distributed.

8.3 THE INFLUENCE OF MIGRATION ON REGIONAL LANGUAGES [15]

We examine in this section a second promising arena for future work, namely the influence of migration on language. Population genetics and demography can provide evidence about the sources, destinations and sizes of population movements, providing a wealth of data on which to test hypotheses about the effect of migration on language. Our goal here is complementary to Falck et al. (2012), who showed that people prefer to move to areas in which the local dialect is more similar to their own.

8.3.1 The effect of linguistic contact

The GRAVITY MODEL (Trudgill 1974) explains the spread of linguistic innovations using as parameters geographic distance between speech communities and the number of speakers of the settlements/inhabited areas. While a demographic influence (population size, migrations) on linguistic diversity is obvious (linguistic barriers can orient migrations, however), there are few quantitative studies on this influence, probably due to the lack of detailed and readily-available historical demographic data concerning a full linguistic domain. Let's first focus on linguistic contact to address the quantification of the demographic phenomena that drive it. For 40 years a large body of research has been focused on investigating the effect of contact between mutually intelligible dialects.[16] Many case-studies have demonstrated that different varieties, in direct contact through face-to-face interaction (linguistic accommodation), become more alike with the time. Contact-induced linguistic accommodation involves several processes that can be briefly summarized as follows:

1. *Levelling*, the reduction in either the number of linguistic variants or in the degree of their variability. When two alternative forms exist, usually only one is preserved.
2. *Emergence of intermediate varieties*, where some linguistic forms may be new and may not have occurred in any of the dialects before the contact (*koine*).
3. *Reallocation*, in which alternative forms are retained but assigned different roles in the sociolinguistic use of the dialects, or in their grammatical use.
4. *Simplification* and increase in regularity.

[15] This section is largely based on the work about linguistic contact of J. Chambers (ex Univ. of Toronto), W. Labov (Univ. of Pennsylvania) and P. Trudgill (Univ. of East Anglia). Among others, some key references are Dodsworth (2017) and Kerswill (2006).

[16] Urban districts receiving immigrants from rural areas should be distinguished from sparsely populated areas that became the target of massive immigration, like new towns or in co-lonial settlements, because linguistic innovations spread faster in recent speech communities than in pre-existing ones.

In the phase that follows the initial dialect-contact, rudimentary levelling or extreme inter-speaker / intra-speaker variability can be noticed. Later a new focused variety gets established and "homogeneously" adopted by the whole speech community. This process lasts at least one generation because new variants get fixed at adolescence age and adults are less likely to modify their speech.

An attractive area, say an economically dynamic town, has probably been destination of migrants for centuries; initially they came from close areas but, with the time passing, immigrants from more distant areas (where more divergent dialects are spoken) moved in. This is to say that linguistic levelling and simplification are expected to be stronger in the dialect of an attractive area than in the dialect of an area that is not, because immigration from distant regions is less likely in the second case, and contact phenomena are stronger when the linguistic difference between varieties in contact is higher. On the other hand, an unappealing town has likely lost a large part of its population that migrated to find better living conditions. A linguistic consequences of this phenomenon is that the dialect of the latter has remained stable over the time, and has *not* undergone processes of linguistic simplification (see Fig. 8.4).[17]

8.3.2 Extensive linguistic contact and demography

8.3.2.1 Dialect change in the Netherlands

Some attempts to measure extensive dialect change were based on the comparison of linguistic atlases made in different epochs, or by comparing the pronunciation of speakers of different ages within a same family or community. Dutch dialectology offers interesting computational work addressing the linguistic change of dialect varieties with the time. Wieling (2007b/c) compared two pronunciation datasets collected, approximately, at two generations-interval (50 years)[18] and found that Friesland and Limburg are areas of dynamic convergence, while the south-eastern part of Low Saxony (Groningen, Drenthe, Overijssel, and the eastern part of Gelderland) is an area of divergence. His results do not align well with those of Heeringa and Hinskens (2015) that compared the pronunciation of present-day older male speakers and younger female speakers (two generation interval) obtaining capricious

[17] To explain the considerable linguistic effect of immigration it should be recalled that emigrants are generally younger than the average of the population, meaning that they are more likely to bring linguistic innovations.

[18] The altlas of Blancquaert and Peé (1982), created during the period 1925-1982 (but data generally correspond to the first half of the period), was compared with the Goeman-Taeldeman-van-Reenen dataset (Goeman and Taeldeman, 1996; Van den Berg 2003) collected over the years 1980 – 1995.

patterns. While the latter study is based on more consistent generation samples than the first one, the results of both have not been related to the population-growth and -size of corresponding provinces. In fact the increase in the population size could have been a key to link dialect change, population growth and immigration. Actually, Wieling *et al.* (2007b, 2007c) and Heeringa and Hinskens (2015) addressed a linguistic change that took place in very recent times, when the use of dialects was already growing less frequent and when the speakers were extensively exposed to the linguistic norm (standard Dutch), two factors that make the interpretation of the observed change difficult, because several sociolinguistic effects overlap.

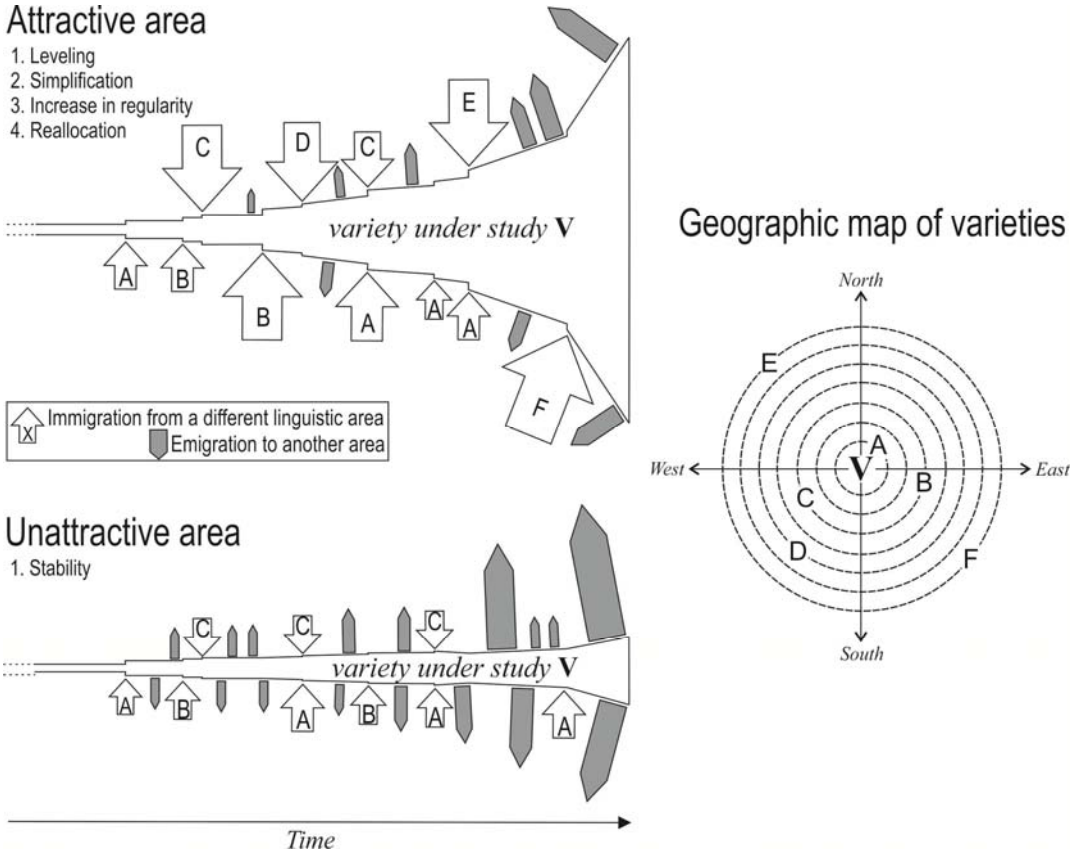

Figure 8.4 ▸ Two extreme scenarios of dialect contact according to migration. *Attractive Area:* The variety spoken here is frequently in contact with new varieties; earlier immigration comes from the neighbourhood, while later immigrants come from distant areas. Besides the normal population growth over time, the demographic balance is as positive as the migratory balance. The timeline reported can be assumed to cover some centuries. *Unattractive Area:* The spoken variety is less exposed to the contact with different varieties because the majority of immigrants comes from neighbouring areas. The population size may fall or remain somewhat stable over the time because there are few immigrants and many emigrants. The growth of the population counterbalances the loss of populations only partially. The timeline reported can be assumed to cover some centuries.

Ideally, to better assess the linguistic change that dialects experienced over the recent demographic transition from the rural to the urban way of life,[19] one would require a linguistic atlas of some centuries ago (of course *not* available) to be compared to a linguistic atlas concerning data collected in the first half of the last century, when dialects were still largely spoken and less influenced by the norm (such atlases *are* available for a majority of European countries). Along those lines, the study of Heeringa and Joseph (2007) was aimed at comparing the pronunciation data of the linguistic atlas of Blancquaert and Peé (1982) to the reconstructed pronunciation of proto-Germanic lexicon, and at estimating the degree of conservatism of contemporary Dutch dialects accordingly. I will not mention the special case of Frisian[20] and just focus on the findings of Heeringa and Joseph (2007) concerning the other Dutch varieties, that is Low Franconian dialects (central western part of the Netherlands) that tend to be more conservative (particularly Holland, the eastern part of North Brabant and the southern part of Gelderland) and Low Saxonian dialects, that are phonologically more innovative than the first group, according to reconstructed proto-forms.

Is this pattern the outcome of historical phenomena that took place during the linguistic differentiation from proto-Germanic over two millenia? While it is difficult to answer directly, a paradox should be noted: nearly all the most conservative areas identified by Heeringa and Joseph (2007) are located in the four provinces (South Holland, North Holland, North Brabant and Gelderland), which have had continued and very high population growth (and immigration) since about one century ago (Fig. 8.5). Actually, the phonological conservatism of many Franconian dialects might be a *recent effect* if we adopt the following sociolinguistic perspective: the high immigration and population expansion in some regions of the Franconian linguistic domain led to linguistic levelling, with an overall reduction of the number of different realizations of phonological variables but, also, with a general and increasing exposure to Standard Dutch, a variety that has a remarkably conservative sound system (Donaldson 1983, p. 161). In the opposite way, the Low Saxon dialects are more innovative as they experienced a inferior degree of levelling because of the lower population growth and immigration, meaning that standard Dutch was not needed as *lingua franca*, and, also, because of their different and more rural socioeconomic environment (but see Haartsen *et al.* 2003).[21]

[19] It must be emphasised that in middle of the seventeenth century a large proportion of the Dutch population was already living in towns and cities.

[20] That turned out to be conservative too.

[21] Some influence of phonologically more innovative varieties located across the German border, influenced by the Standard German norm, cannot be excluded. For example, the First Germanic Sound Shift (Campbell 2004) can be found reasonably intact in Dutch, and

If none of the studies cited above (Wieling 2007b, 2007c, Heeringa and Joseph 2007, Heeringa and Hinskens 2015) provides conclusive results and patterns about the evolution of Dutch dialects over time, it may be because *i)* they are exclusively based on phonology while dialect mixing, stability, hypervariability, reallocation and convergence *also* concern morphosyntax *and* rhythmic differences (see Dodsworth 2017 for a review of modern case-studies),[22] and because *ii)* the timeframe addressed by Wieling *et al.* (2007b, 2007c) and Heeringa and Hinskens (2015) is blurred by the progressive contemporary levelling of dialects and, finally, since *iii)* the sampling scheme was not designed to correspond to areas demographically comparable in terms of migration and population growth.

8.3.2.2 Migrations inferred from surname data

Aside from special cases where the time of the arrival and the provenance of immigrants are known (as in the Canadian province of Québec for example)[23] it is often difficult to have this information, even for historically recent times. This is why surnames can be of help when no alternative documentation is available. They make it possible to identify the direction of migrations that took place in, say, a European country over the last four or five centuries but, unlike historical registers, they do not show *when* such migrations took place. They could have happened anytime between the first introduction of family-names to the last generation, that is for a time span of about five centuries.[24]

We know that the a majority of these migratory movements took place after the industrial revolution, when new means of displacement became available and new jobs where massively created, therefore contributing to the establishment of coherent migration routes (like the northwards substantial immigration that took place from the rural southern part of Italy to its economically-dynamic northern side). The detailed comparison of these migration routes (Manni *et al.* 2005, Boattini *et al.* 2012, Rodriguez Diaz *et al.* 2015, 2017) reveals that neighbouring provinces can be quite different in the number and the provenance of the immigrants they attracted, but also concerning the directions of the emigrants that left them. This heterogeneity is valu-

all other Low German and Scandinavian languages for that matter, but a further shift of *p*/*t*/*k* occurred only in German which momentarily obscures the origin of some sounds (Donaldson 1983, p. 123).

[22] The examples concern the dialects of London (UK), Sao Paulo (Brazil), Xining (PRC), Amman (Jordan) and the Spanish varieties spoken in New York (USA).

[23] In Quebec migration registers have been kept since the beginning of the French rule.

[24] In the Netherlands surnames have a more recent origin.

able to test the cumulative effect[25] that migrations had on the dialects spoken in two neighbouring areas and that *initially* were linguistically very similar, meaning that they later diverged as a consequence of the dialect contact that migration processes drove.

A possible experimental set-up would be to compare couples of locations that initially had a comparable population size and where very close varieties were spoken before linguistically diverging because of a different migration history, similarly to the two cases shown in Fig. 8.4. The linguistic differences between couple of localities[26] (inferred using a linguistic atlas) selected in this way are expected to correspond to different types of migration-induced dialect contact.

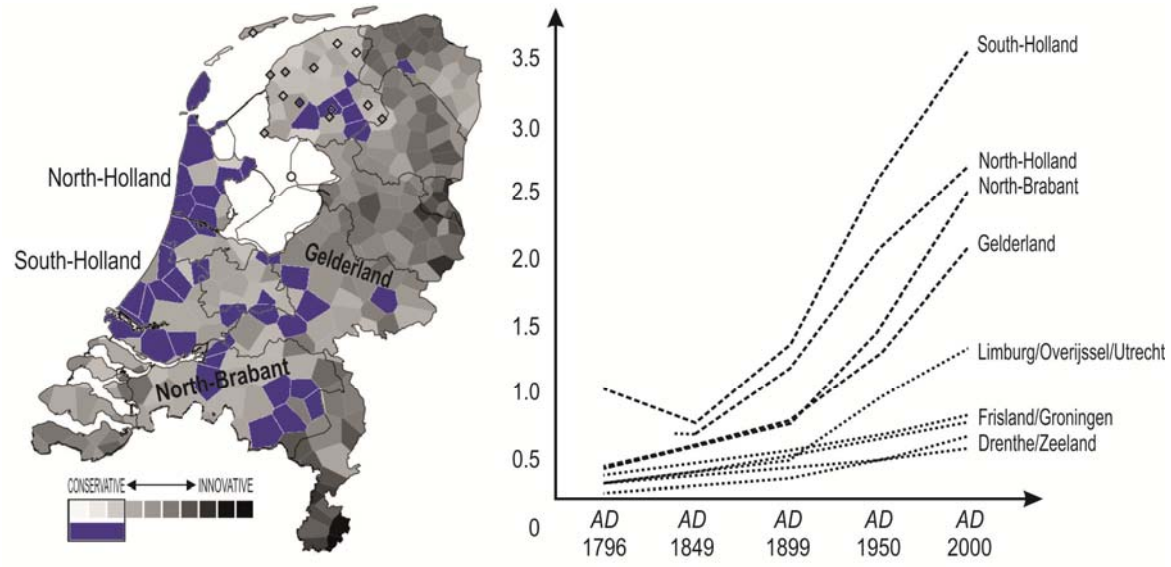

Figure 8.5 ▸ Conservativeness of Dutch dialects and demographic growth by province.
Right: Phonetically conservative and innovative dialects in the Netherlands according to Heeringa and Joseph (2007). Ten shadings cover the spectrum from conservative (lighter gray) to innovative areas (darker gray). The three most conservative classes of the spectrum are here coloured in blue, to show that they are mainly located in the provinces of North and South Holland, North Brabant and Gerderland. Friesland is not discussed in the text. Adapted from Heeringa and Joseph (2007). *Left:* Demographical growth per Dutch province according to several sources, including Blink (1897) and the Dutch Central Bureau of Statistics. The *x*-axis reports the years of the census, the *y*-axis indicates the population-size of each province expressed in millions. Sorce: http://www.populstat.info/Europe/netherlp.htm

[25] Cumulative means the aggregation of all migration movements over the time, because, as it was said, surnames do not allow distinguishing their timing.
[26] Only dialect contact of similar dialects spoken within a same country is taken into account here, not the linguistic contact with other languages.

This is a kind of contact that surname studies can help to describe in detail, meaning as it is possible to say how many speakers of each variety came into contact with the original dialect spoken in the two locations under study. A working hypothesis is that immigration of very different varieties has greater impact on the receiving speech communities than does the immigration of similar varieties, where we would expect that the receiving community's speech should remain more "stable". This stability can imply a higher level of areal heterogeneity and a lower number of innovations. According to Trudgill (1992, p. 199) innovation and simplification can be synonymic because the growth of new forms, that were not present in the initial mixture but that developed out of the interaction between varieties, gives rise to interdialects that are more regular. Speech communities having frequent contacts with other groups tend to have simplified (innovative) languages or dialects. An exemplification comes from a forthcoming article concerning Spanish data (Rodriguez-Diaz 2017)

8.3.2.3 Spanish surnames and internal migrations

The way to account for the intensity and the directions of the internal migrations that took place in Spain after the introduction of surnames is quite simple. It consists in coding the surnames listed in the database of current Spanish residents (*Padrón municipal*)[27] as vectors whose components correspond to the relative frequency of each surname in, say, all the 47 continental Spanish provinces.[28] Then all the vectors (surnames) are classified in a discrete number of clusters by using Kohonen maps (Kohonen 1982, 1984, Kaski 1997) or other similar methods, so that each cluster corresponds to a group of surnames having a comparable geographic distribution over the country: frequent in some provinces and not in others. Finally, such groups are plotted over a geographic map to see if there are visible peaks of frequency corresponding to a single province.[29] If one assumes that the province, where the relative frequency of each surname is the highest, corresponds to the geo-historical origin of corresponding surnames, it is possible to measure migrations because the diffusion centre of these family names is known as well as their present-day distribution. In this way all migration patterns can be summarized in two migration matrices, one for the aggregate

[27] Only surnames occurring at least 20 times have been processed.

[28] Example: *Rodriguez; (Province 1 =) 0.0047; (Province 2 =) 0.0030; …; (Province 47 =) 0.0018.*

[29] In some cases, the peak of frequency is geographically ambiguous because it corresponds to two (or more) provinces. Such ambiguities are related to the fact that many surnames, spelled in a same way, *independently* became the name of *unrelated* families located in different areas (as it is the case for *de Boer*, *van Dijk*, *de Jong*, *Visser* in the Netherlands). Only the surnames with a *clear* origin in *one* province have been kept.

immigration- and one for the aggregate emigration-processes that took place over the last five centuries (Fig. 8.6).[30]

As was anticipated in the introduction of the dissertation (see CHAPTER 1, section 1.2), emigration and immigration phenomena are not symmetrical, this is why Spanish provinces can be classified into four groups: *1) Isolated provinces* (low emigration, low immigration); *2) Corridor provinces* (high emigration, high immigration); *3) Unattractive provinces* (high emigration, low immigration)*; 4) Attractive provinces* (low emigration, high immigration) as in Fig. 8.7.

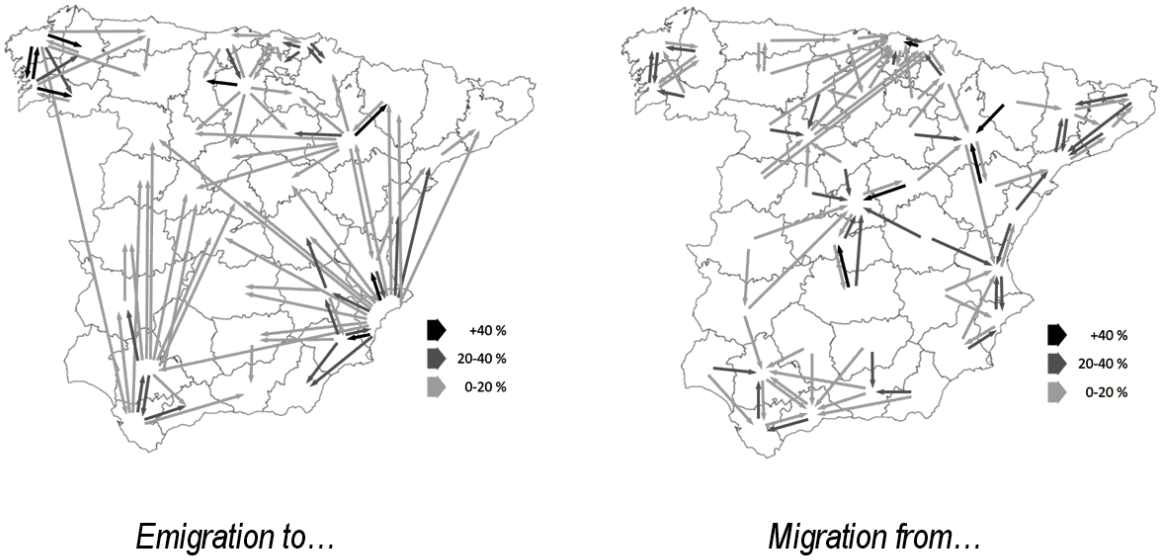

Figure 8.6 ▸ Immigration and emigration in Spain. *(Left):* Major emigration routes from each continental Spanish province. *(Right):* Major immigration routes from all continental Spanish provinces to each one of them. Analysis based on the distribution of 25,714 single surnames (like 'Diaz'; 'Rodriguez'; *etc*.) corresponding to the 12,348,109 Spanish residents processed by Rodriguez-Diaz *et al.* (2017).

Concerning migration distances, they can be classified as short-, medium- and long-range (Fig. 8.8). It is reasonable to think that the medium and long range movements took place in more recent times, when the mechanization of transportation and the industrialization of the country led to *massive* displacement of the population that progressively abandoned rural life. Differently, other provinces are characterized by very *local* emigration distances directed to neighbouring areas; they correspond to processes that took place within a more traditional frame of displacement, probably when people used to move by their own means, progressively diffusing as described in the WAVE THEORY of Schmidt (1872).

[30] Spanish surnames became fixed starting with the 16th century.

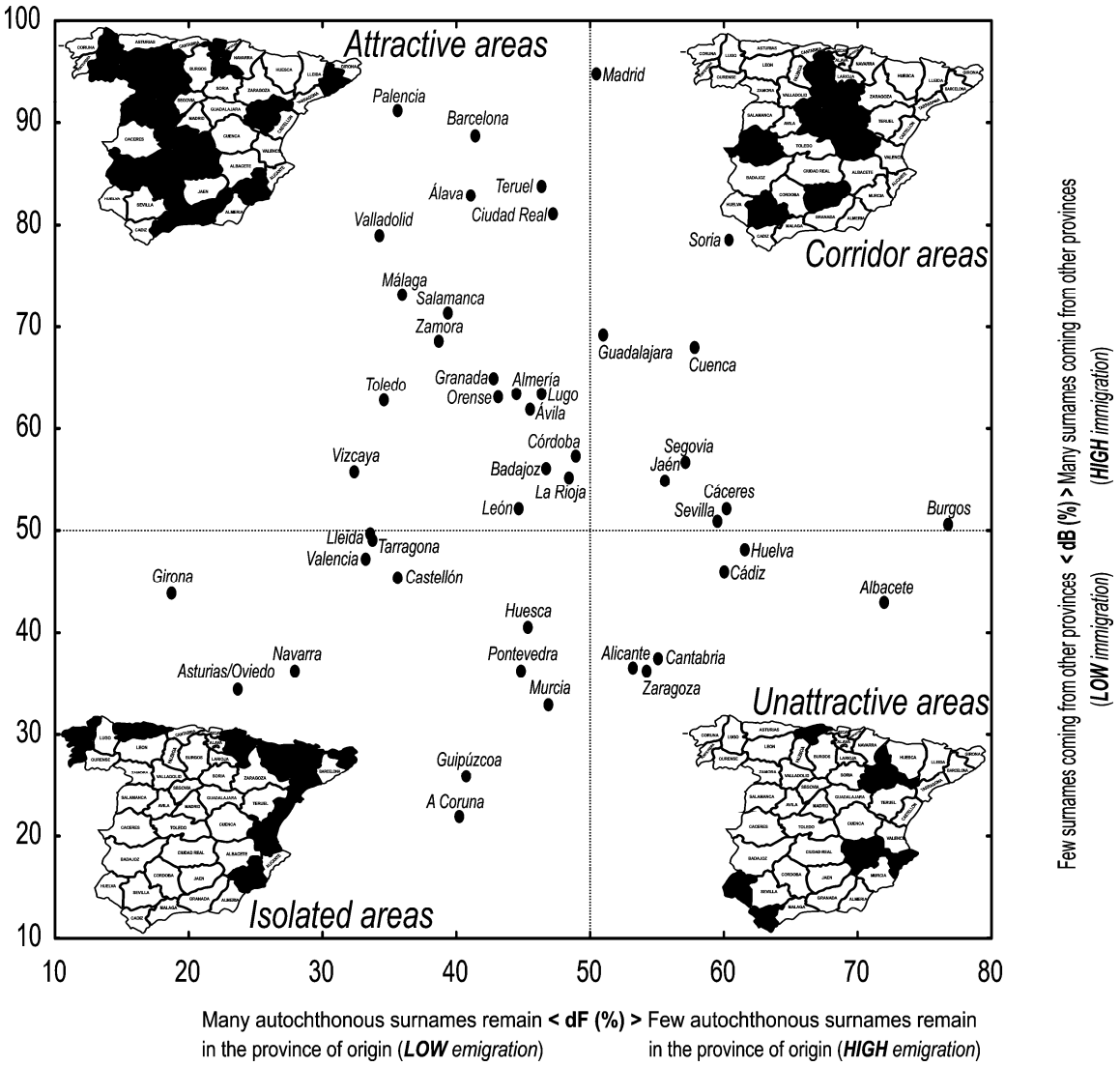

Figure 8.7 ▸ Bidimensional plot of immigration (*x*-axis: [%] of surnames of foreign origin in each Province) and emigration (*y*-axis: [%] of surnames located outside the province of origin) by province. In the plot it is possible to identify four different cases. Dataset as in Fig. 8.6.

8.3.2.4 Spanish migrations and regional languages

Concerning Spain, it is interesting to note the significant overlap between:

i) The areas that remained rather isolated in terms of internal Spanish migrations (see bottom part of Fig. 8.7);

ii) The provinces in which emigration was mainly directed to neighbouring areas following a short-range migration discipline of isolation by distance (Wright 1943, Malécot 1948) (see the bottom left part of Fig. 8.8);

iii) The regions where languages other than Castilian have resisted the political will to set Castilian as *The* national language [31] (see Fig. 8.9), therefore showing the positive effect that reduced immigration had on the persistence of language areas.

To conclude on this challenging result, I note that, in the past, linguistics has driven a considerable amount of hypotheses about the anthropological diversity of human populations. Today, demographers and geneticists can also help by providing an accurate and large-scale quantification of the demographic processes that led to linguistic contact, setting a new framework to understand linguistic differentiation.

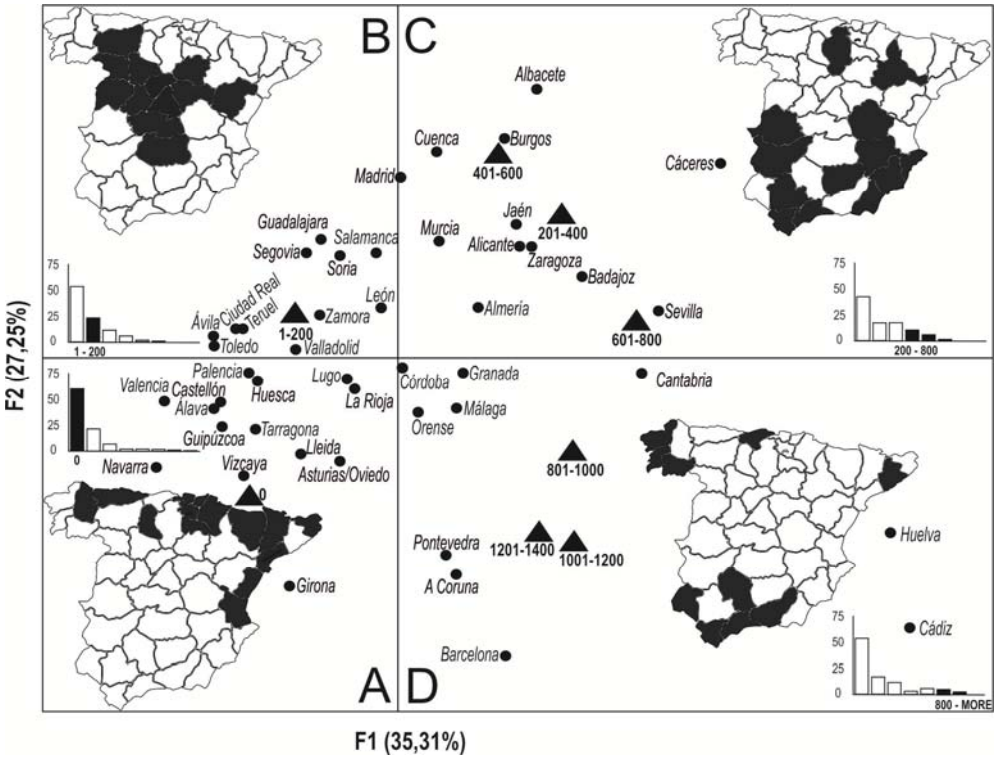

Figure 8.8 ▸ Emigration by distance classes from each Spanish province. By comparing the present-day distribution of Spanish surnames to their inferred geographical origin (where they were first adopted about five centuries ago), it is possible to dissect emigration distances (emigration distances have been ranked in 8 distance classes; see triangles). The 47 vectors (one per province) accounting for the distance classes have been the input of the Principal Component Analysis shown above. Provinces from which emigration was directed to the closest neighbouring areas (bottom left) can be distinguished from the others. Dataset as in Fig. 8.6.

[31] This policy lasted from the beginning of the 16th century, with the unification of the crowns of Aragon and Castilla, to the times of the regime of General Franco ended in 1975.

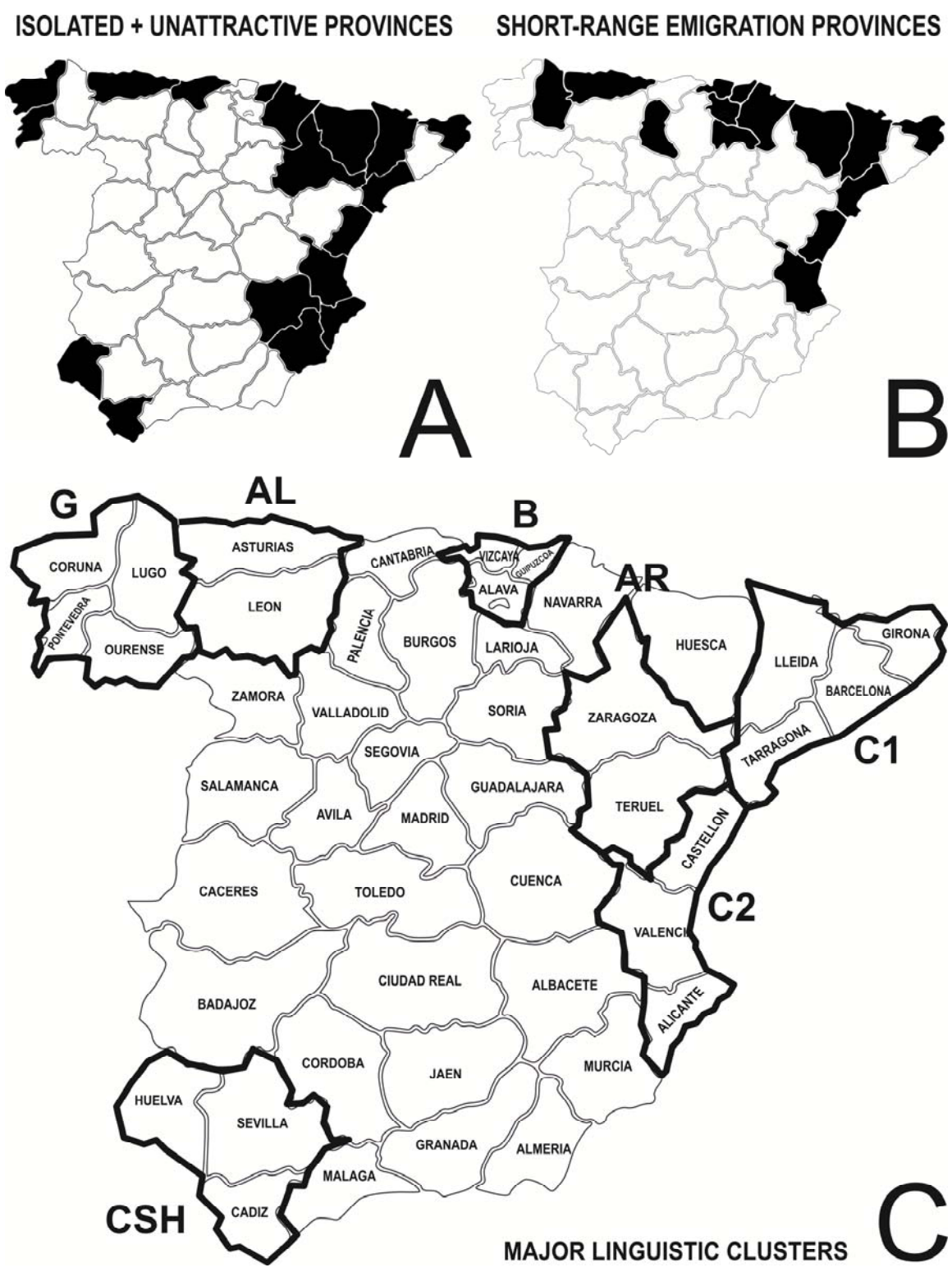

Figure 8.9 ▸ Migration and linguistic diversity in Spain. *(A):* Provinces that have attracted a low number of immigrants (see Fig. 8.7). *(B):* Spanish provinces from which migration has been local and directed to neighbouring areas (Fig. 8.8). *(C):* Major linguistic areas according to a computational linguistics analyses reported in CHAPTER 5.

References:

Almeida A., Braun A. 1986. 'Richtig' und 'Falsch' in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, **53**:158-172.

Almeida, A., Braun, A. 1985. What is Transcription? In: W. Kurschner, R. Vogt (eds.) *Grammatik, Semantik, Textlinguistik. Akten des 19 Linguistischen Kolloquiums Vechta 1984*. Vol. 1, Tűbingen, pp. 37-48.

Beijering K., Gooskens C., Heeringa W. 2008. Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. *Linguistics in the Netherlands*, **25**: 13-24.

Blancquaert E., Peé, W. (eds.). 1925–1982. Reeks Nederlans(ch)e Dialectatlassen. Antwerpen: De Sikkel.

Blink H. 1897. Tegenwoordige staat van Nederland. Amsterdam: S.L. van Looy.

Boattini A., Lisa A., Fiorani O., Zei G., Pettener D., Manni F. 2012. General Method to Unravel Ancient Population Structures through Surnames. Final Validation on Italian Data. *Human Biology*, **84**: 235-270.

Bolognesi R., Heeringa W. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap* **9**: 45-84.

Campbell L. 2004. Historical linguistics (2nd ed.). Cambridge: MIT Press.

Church K.W., Hanks P. 1990. Word association norms, mutual information, and lexicography. *Computatational Linguistics*, **16**: 22–29.

Clua E., Valls E., Viaplana J. 2008. Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà. In: G. Blaikner Hohenwart *et al.* (eds.) *Ladinometria Miscellanea per Hans Goebl per il 65º compleanno* Edizione multiligue , vol. 2. Vigo di Fassa: Istituto Culturale Ladino, pp. 27-42.

Clua E., Lloret M-R. 2015. COD2: An Oral Dialectal Corpus for the Analysis of Spatial and Temporal Variations in Catalan. In: *Proceedings of the 7th International Conference on Corpus Linguistics. Current Work in Corpus Linguistics. Working with Traditionally-conceived Corpora and Beyond* (CILC 2015). Valladolid, 5-7 March, pp. 89-94.

Covington M. A. 1996. An Algorithm to Align Words for Historical Comparison. *Computational Linguistics*, **22**: 481-496.

Cucchiarini C. 1993. Phonetic Transcription: a Methodological and EmpericalStudy. PhD dissertation. Nijmegen: Katholieke Universiteit Nijmegen.

Dodsworth Rn. 2017. Migration and Dialect Contact. *Annual Review of Linguistics*, **3**: 331-346.

Donaldson B.C. 1983. Dutch: A Linguistic History of Holland and Belgium. Leiden: Martinus Nijhoff.

Eisenstein J., O'Connor B., Smith N.A., Xing E.P. 2014. Diffusion of lexical change in social media. ***PLoS ONE***, **9**. http://dx.doi.org/10.1371/journal.pone.0113114

Falck O., Heblich S., Lameli A., Südekum, J. 2012. Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, **72**: 225-239.

Fano R. M. 1961. Transmission of Information: A Statistical Theory of Communications. Cambridge, (MA): MIT Press.

Fontan L., Farinas J., Ferrané I., Pinquier J., Aumont X. 2015. Automatic intelligibility measures applied to speech signals simulating age-related hearing loss. In: *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2015), Dresden, 6-10 September.

Gildea D., Jurafsky D. 1996. Learning Bias and Phonological-Rule Induction. *Computational Linguistics*, **22**: 497–-530.

Goeman T., Taeldeman J. 1996. Fonologie en morfologie van de Nederlandse dialecten: een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, **48**: 38–59.

Gooskens C., Heeringa W. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, **16**: 189-207.

Gooskens C., Heeringa W. 2006. The Relative Contribution of Pronunciation, Lexical and Prosodic Differences to the Perceived Distances between Norwegian dialects. *Literary and Linguistic Computing.* **21**: 477-492.

Greenhill S. 2011. Levenshtein distances fail to identify language relationship accurately. *Computational linguistics*, **37**: 689-698.

Grollemund R., Branford S., Bostoen K., Meade A., Venditti C., Pagel M. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences USA*, **112**: 13296-13301.

Haartsen T., Groote P., Huigen P.P.P. 2003. Rural areas in the Netherlands. *Tijdscrift voor Economische en Sociale Geografie*, 94: 129-136.

Hammarström, H., Forkel R., Haspelmath M., Bank S. 2016. Glottolog 2.7. Jena: Max Planck Institute for the Science of Human History. (Avail. http://glottolog.org, Accessed on 2017-01-25.)

Heeringa W. 2004. Measuring dialect pronunciation differences using Levenshtein distance. PhD Doctoral disserationthesis. Groningen: Rijksuniversiteit Groningen.

Heeringa W., Braun A. 2003. The Use of the Ameida-Braun System in the Measurement of Dutch Dialect Distances. *Computers and the Humanities*, **37**: 257-271.

Heeringa W., Hinskens F. 2015. Dialect change and its consequences for the Dutch dialect landscape. How much is due to the standard variety and how much is not? *Journal of Linguistic Geography*, **3**: 20-33.

Heeringa W., Joseph B. 2007. The Relative Divergence of Dutch Dialect Pronunciations from their Common Source: An Exploratory Study. In: J. Nerbonne, T. Mark Ellison, G. Kondrak G. (ed.), *SigMorPhon 07 ACL 2007, Computing and Historical Phonology, Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology.* Prague, (Czech Republic), Stroudsburg (PA): The Association for Computational Linguistics (ACL), June 28, pp. 31-39.

Heeringa W., Kleiweg P., Gooskens C., Nerbonne J. 2006. Evaluation of String Distance Algorithms for Dialectology. In: J. Nerbonne, E. Hinrichs (eds.) *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, July, pp. 51-62.

Heeringa W., Nerbonne J. 2001. Dialect areas and dialect continua. *Language Variation and Change*, **13**: 375-400.

Heeringa W., Nerbonne J., Kleiweg P. 2002. Validating Dialect Comparison Methods. In: W. Gaul, G. Ritter (eds.), *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*, pp. 445-452.

Jäger G. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences USA*, **112**: 12752-12757.

Juola P., Sofko J., Brennan P. 2006. A Prototype for authorship attribution studies. *Literary and Linguistic Computing*, **21**: 169-178.

Kaski S. 1997. Data exploration using self-organizing-maps. *Acta Polytechnica. Scandinavica*. **82**:1-57.

Kerswill P. 2006. Migration and language. In: K. Mattheier, U. Ammon, P. Trudgill (eds.), *Sociolinguistics/Soziolinguistik. An international handbook of the science of language and society*, 2nd edition, volume 3, Berlin: De Gruyter.

Kessler B. 1995. Computational Dialectology in Irish Gaelic. In: *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 60–67.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**: 59–69.

Kohonen T. 1984. Self-organization and Associative Memory. Berlin: Springer.

Kondrak G. 2002. Algorithms for language reconstruction. (Doctoral dissertation, University of Toronto). *Dissertation Abstracts International*, **63**: 5934.

Kondrak G. 2003. Phonetic alignment and similarity. *Computers and the humanities*, **37**: 273-291.

Kondrak G., Sherif T. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In: *Proceedings of the ACL Workshop on Linguistic Distances*. Sydney: Australia, pp. 43-50.

Labov W. 2001. Principles of linguistic change: Social factors. Vol. II. Malden: Blackwell.

Lewis D.K. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, **8**: 339-359.

Malécot G. 1948. Les mathématiques de l'hérédité. Paris: Masson.

Mann G.S., Yarowsky D. 2001. Multipath translation lexicon induction via bridge languages. In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, June 1-7, pp 1-8.

Manni F., Toupance B., Sabbagh A., Heyer E. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *American Journal of Physical Anthropology*, **126:** 214-28.

Mathussek A. 2016. On the problem of field worker isoglosses. In: M-H Côté, R. Knooihuizen, J. Nerbonne (eds.), *The future of dialects, dialects: Selected papers from Methods in Dialectology XV*. Berlin: Language Science Press, pp. 99-116.

McMahon A., McMahon R. 2005. Language classification by the numbers. Oxford: Oxford University Press.

Needleman S.B., Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**: 443–53.

Nerbonne J. 2005. Review of April McMahon and Robert McMahon *Language Classification by the Numbers.* Oxford: Oxford University Press. *Linguistic Typology* **11**: 425-436.

Nerbonne J. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B*, **365**: 3821-3828.

Nerbonne J., Heeringa W. 1997. Measuring Dialect Distance Phonetically In: J. Coleman (ed.) *Workshop on Computational Phonology. Special Interest Group of the Association for Computational Linguistics*. Madrid: ACL, pp. 11-18.

Nerbonne J., Heeringa W. 2007. Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: S. Featherston, W. Sternefeld (eds.) *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton De Gruyter, pp. 267-297.

Nerbonne J., Heeringa W. 2010. Measuring dialect differences. In : P. Auer, J.E. Schmidt (eds.) *Language and Space: Theories and Methods*. Berlin: Mouton De Gruyter, pp. 550-566.

Nerbonne J., Heeringa W., van den Hout E., van de Kooi P., Otten S., van de Vis W. 1996. Phonetic Distance between Dutch Dialects. In: G. Durieux, W. Daelemans, S. Gillis (eds.eds.) *CLIN VI: Proceedings. of the Sixth CLIN Meeting.* Antwerp: Centre for Dutch Language and Speech (UIA), pp.185-202.

Nurse D., Philippson G. 2003. Towards a historical classification of the Bantu. In: D. Nurse and G. Philippson (eds.), *The Bantu languages*. London: Routledge, pp. 164-179.

Oakes M.P. 2000. Computer estimation of vocabulary in protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, **7**: 233–243.

Rama T., Wahle J., Sofroniev P., Jäger G.. 2017 Unpublished. Fast and unsupervised methods for multilingual cognate clustering. *ArXiv preprint*, arXiv:1702.04938.

Rodríguez Díaz R., Manni F., Blanco Villegas M-J. 2015. Footprints of Middle Ages Kingdoms Are Still Visible in the Contemporary Surname Structure of Spain. *PLoS ONE*, **10**. doi:10.1371/journal.pone.0121472

Rodríguez Díaz R., Blanco Villegas M-J, Manni F. 2017. From surnames to linguistic and genetic diversity: Five centuries of internal migrations in Spain. *Journal of Anthropological Sciences*, **95**: 000-000, forthcoming.

Sanders N., Chin S. 2009. Phonological distance measures. *Journal of Quantitative Linguistics*, **16**: 96-114.

Scapoli C., Mamolini E., Carrieri A., Rodriguez-Larralde A., Barrai I. 2007. Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. *Theoretical Population Biology*, **71**: 37-48.

Scherrer Y. 2007. Adaptive string distance measures for bilingual dialect lexicon induction. *ACL '07 Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, Prague (Czech Republic), June 25-26, pp. 55-60.

Schmidt J. 1872. Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar: H. Böhlau.

Séguy J. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, **35**: 335-357.

Somers H. L. 1998. Similarity metrics for aligning children's articulation data. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 1227–1231.

Spruit M.R., Heeringa W., Nerbonne J. 2009. Associations among linguistic levels. *Lingua*, **119**: 1624-1642.

Szmrecsanyi B. 2013. Grammatical variation in British English dialects. A study in corpus based dialectometry. Cambridge (UK): Cambridge University Press.

Trudgill P. 1974. Linguistic Change and Diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, **2**: 215-246. pp.

Trudgill P. 1992. Dialect tipology and social structure. In: E.H. Jahr (ed.) *Language contact*. Berlin, New York (NY): Mouton de Gruyter, pp. 195-212.

Valls E., Nerbonne J., Prokic J., Wieling M., Clua E., Lloret M-R. 2012. Applying the Levenshtein distance to Catalan dialects: A brief comparison of two dialectometry approaches. *Verba*, **39**: 35-61

Van den Berg B.L. 2003. Phonology and Morphology of Dutch and Frisian Dialects in 1.1 million transcriptions. Goeman-Taeldeman-Van Reenen project 1980-1995, *Meertens Instituut Electronic Publications in Linguistics* 3. Amsterdam: Meertens Instituut (CD-ROM).

Vieregge W.H., Rietveld A.C.M., Jansen C.I.E. 1984. A distinctive feature based system for the evaluation of segmental transcription in Dutch. In: M.P.R. Van den Broecke, and A. Cohen (eds.), *Proceedings of the 10th International Congress of Phonetic Sciences*, Dordrecht and Cinnaminson: Foris Publications, pp. 654-659.

Wieling M., Leinonen T., Nerbonne J. 2007a. Inducing Sound Segment Differences using Pair Hidden Markov Models. In: John J. Nerbonne, M. Ellison, G. Kondrak (eds.) *Computing and Historical Phonology: 9th Meeting of ACL Special Interest Group for Computational Morphology and Phonology* Workshop at ACL. Prague, pp. 48-56.

Wieling M. 2007b. Comparison of Dutch Dialects. Master thesis, University of Groningen.

Wieling M., Heeringa W., Nerbonne J. 2007c. An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen Project data. *Taal en Tongval*, **59**: 84-116.

Wieling, M., Bloem J., Mignella K., Timmermeister M., Nerbonne J. 2014. Automatically measuring foreign accent strength in English. Validating Levenshtein Distance as a Measure. *Language Dynamics and Change*, **4**: 253-269.

Wieling, M., Margaretha E., Nerbonne J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, **40**: 307-314.

Wright S. 1943. Isolation by distance. *Genetics*, **28**: 114-138.