Do Surname Differences Mirror Dialect Variation?

FRANZ MANNI,[1] WILBERT HEERINGA,[2] BRUNO TOUPANCE,[1] AND JOHN NERBONNE[2]

Abstract Our focus in this paper is the analysis of surnames, which have been proven to be reliable genetic markers because in patrilineal systems they are transmitted along generations virtually unchanged, similarly to a genetic locus on the Y chromosome. 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, an area of only 40,000 km^2, 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.

The major aim of this study is to evaluate to what extent the patterns of geographic variation of surnames overlap with those of linguistic diversity, because family names can be seen as a specific part of language.

This research was inspired by a similar paper in which we addressed the similarities between the geographic structures of genetic, surname, and linguistic variables in a small area of Italy (Manni and Barrai 2001). In that study we used a more rudimentary methodology that, nonetheless, was already fully computational, even concerning the linguistic treatment; thus statistical comparisons with diverse markers were possible. Using similar methods, our aim is now to compare the geographic patterns of surname-inferred genetic variation with corresponding linguistic data in the Netherlands (see Figure 1 for a map of the different Dutch

[1]UMR 5145, CNRS, Musée de l'Homme, National Museum of Natural History, 17 Place du Trocadéro, 75016 Paris, France.
[2]Alfa-Informatica, Faculty of Arts, University of Groningen, Groningen, The Netherlands.

Human Biology, February 2008, v. 80, no. 1, pp. 41–64.

KEY WORDS: GENETICS, LINGUISTICS, SURNAMES, DIALECTS, THE NETHERLANDS, BARRIERS, ISOLATION BY DISTANCE (IBD).

Figure 1. Location of the Dutch provinces in the Netherlands.

provinces), similarly to what we did in a paper addressing a linguistic audience (Manni et al. 2006).

Male-transmitted family names can be regarded as genetic markers because they simulate neutral alleles of a gene transmitted only through the Y chromosome. Therefore they satisfy the expectations of the neutral theory of molecular evolution (Cavalli-Sforza and Bodmer 1971; Crow 1980), which is entirely described by random genetic drift, mutation, and migration (Kimura 1983). Nevertheless, before a strict rule of transmission was established, family names were also words, and so they remain today (even if frozen to meet the needs of administration).

If surnames were words, we might expect them to pattern with other linguistic material; for instance, it is often possible to guess someone's geographic origin by the sound and spelling of his or her surname, that is, on a purely linguistic basis.

We remind readers that in most European countries surnames were not adopted until the Middle Ages, and in the Netherlands they were not obligatory until the Napoleonic period. As a consequence, surname-inferred demographic phenomena such as migrations, drift, and isolation can be dated at best only within the last six centuries for most of Europe and only within the last two centuries in the Dutch case.

We investigated the dual nature of patronymic markers by comparing the geographic patterns of variability of 19,910 Dutch surnames, accounting for 1,303,369 individuals, with the linguistic differences of the Netherlands measured by Heeringa (2004). In the present study we computed a general regression model between pairwise Levenshtein dialect distances (see Heeringa 2004, pp. 121–144) and geographic distances between dialect locations. Then, expected distances were subtracted from the observed distances, leading to the computation of residuals. The residuals reflect variance that is unrelated to geographic distance in general, as though corresponding distances corresponded to the ideal case of equidistant locations, meaning that geographic proximity (or distance) plays virtually no role in residual distances. Finally, we obtained a representation of boundaries (i.e., borders between the most differentiated areas) based on the residuals. We have applied this procedure to surname data as well, but, even though it can be considered a standard method in genetics, we note that the technique is pioneering in dialectology. When we apply the methodology in both areas, we see that surnames are not distributed in the same way as dialect differences are.

Although maps have long been used in dialectology, the idea that divergence increases with geographic distance has not been analyzed to the same degree as in genetics (see Cavalli-Sforza et al. 1994). Nonetheless, in linguistics a related idea can be traced back to the wave theory of Johannes Schmidt (1872) regarding Indo-European languages. According to Schmidt, a linguistic innovation spreads geographically so that its effects can be seen in continuously weakening concentric circles, like the ripples made when a stone is thrown into water. This should lead to convergence among nearby varieties.

From a great distance all languages consist of chains of pairs of mutually intelligible speakers (or speech-types) where different varieties gradually shade into one another and where the extremes of the chain are between the most differentiated areas (I. Dyen, personal communication, 2004). The role played by geographic distance in the steady increase of linguistic divergence is also the point of Chambers and Trudgill's traveler's distance (1998, p. 5). They tell of a traveler going across a linguistic area who repeatedly encounters dialects whose features overlap to a large extent with those of the last dialect he heard and those of the next. The traveler experiences the continuum that is frequently appealed to in dialectology: Neighboring dialects are usually quite similar. Heeringa and Nerbonne (2001) analyzed Chambers and Trudgill's traveler by comparing the step-by-step

linguistic changes along a path with a general regression in which linguistic distance was explained by geographic linguistic distance, following Séguy (1971). The rate of linguistic change was not steady, and some adjacent pairs of dialect variants differed by higher values, suggesting the significance of some borders and therefore of some areas as organizing forces. Further, Nerbonne et al. (1999) computed regression models for all pairwise distances in the matrix of sampling sites, making the computation more stable than that of the one-dimensional traveler's distance along a single line.

If we eliminate the variance explained by geography from a dialectometric distance matrix, we focus on the residual variance that is probably not due to the distance between speakers. From a historical point of view large residuals may signal a linguistic difference that is more profound—for example, one that has arisen through migrations. As a consequence of sparser population density, less contact between speakers, and less reliable transportation, we can (1) imagine that in ancient times linguistic (dialect) differences were stronger than they are today and (2) consider that present-day differences may be the relicts of patterns so remote that analysis of residuals is needed to highlight them effectively.

Materials and Methods

Data

Surnames. From the original database of Dutch surnames that is composed of 51,578 surnames, corresponding to 2,294,154 individual telephone users (1997 data) in 226 locations (Manni 2001), we have selected a new database composed of 19,910 surnames accounting for 1,303,369 individuals (8.1% of the entire Dutch population) and corresponding to those surnames recorded in at least 10 locations and no more than 100 locations, thus excluding rare and polyphyletic surnames, respectively. The upper limit of 100 locations, which was empirically determined, allows us to avoid almost exactly the same polyphyletic surnames identified by Manni et al. (2005). We describe how this new database was obtained in a later section.

Dialects. In the current study linguistic distances were computed over all pronunciations, using the same list of 125 words that Heeringa (2004, pp. 214–215 and pp. 292–294) did, although some technical constraints, related to software limitations, forced us to reduce the number of Dutch samples to 252. In contrast to the surname survey, the dialects of the province of Flevoland (see Figure 1) were not studied because Flevoland has a recent origin, as it was not reclaimed from the sea until 1968.

Computation of Diversity

Surnames. Several studies have focused on the variability of surnames on account of their ready availability, for example, from voters' lists or phone books.

They are useful in the investigation of genetic structures (i.e., meaningful differences in the geographic space) of populations. Although the use of patronymic markers is easy and provides large sample sizes, it also suffers from limitations such as nonpaternity, surname change, polyphyleticism, and a limited temporal depth. Thus the use of surnames is expected to give estimates for demographic inference that differ from real genetic studies that mirror more remote demographic phenomena and are usually based on several loci, unlike surnames. Nonpaternity and surname change are not at all major problems, affecting no more than 10% of the data, but polyphyleticism can decrease the reliability of surname studies. We have recently shown (Manni et al. 2005) how it is possible to reduce this source of error by using a neural network analysis (Kohonen 1995) of the geographic distribution of the surnames. In this way some clearly polyphyletic surnames can be identified because they share crucial properties, such as the absence of a coherent geographic heart of diffusion, a high average number of people sharing the surname, and a peculiar clustering in specific cells of the Kohonen map. The analysis of surnames provided in this paper has been implemented by the exclusion of polyphyletic surnames identified according to such criteria.

Because there are two widespread manners of calculating a measure of surname differentiation, we have computed both Nei and Lasker distances according to the following formulas. Nei's distance is computed as

$$\frac{\sum n_{si} n_{sj}}{\left(\sum n_{si}^2 \sum N_{sj}^2\right)^{1/2}}, \tag{1}$$

and Lasker's distance is computed as

$$-\log\left(2\frac{\sum n_{si} n_{sj}}{\sum n_{si} \sum N_{sj}}\right), \tag{2}$$

where n_{si} denotes the frequency of a given surname s in locations i and n_{sj} denotes the frequency of the same surname in location j. The sums are done for all surnames. By applying these expressions to the surname distributions (list of surnames and their relative frequency) in two locations, we can obtain Nei or Lasker pairwise distances, and therefore we can compile a pairwise distance matrix accounting for all surnames. If the two locations have completely different surnames, then their distance will be maximum, whereas if they share the same set of surnames with identical relative frequencies, their distance will be null. For further details on the computation of such distances, please refer to Nei (1973) and Rodriguez-Larralde et al. (1998).

Dialects. We applied sequence comparison to pronunciations belonging to different dialect varieties to measure the distance between them, similar to what is done by molecular biologists when two or more DNA sequences are aligned by dynamic programming techniques. To this end we adopt the Levenshtein (1966)

distance, an algorithm able to align pronunciations of variable lengths that was first applied to dialects by Kessler (1995) for the comparison of Irish Gaelic varieties. Alignment algorithms have been used to study language since at least the 1970s. Kruskal and Liberman (1983) provide an overview of dynamic time warping, a technique that gained popularity in speech recognition in the 1970s. The difference between two pronunciations is explained by the insertion, deletion, or substitution of sounds, and weights are assigned to these three operations. Assume that *afternoon* is pronounced as [æəftənʉn] (with a phonetic notation that ignores tones that are represented by accent marks) in the dialect of Savannah, Georgia, and as [æftərnʉn] in the dialect of Lancaster, Pennsylvania. Changing one pronunciation into the other can be done as follows:

æəftənʉn	delete ə	1
æftərnʉn	insert r	1
æftərnʉn	substitute ʉ/u	1
æftərnun		

Because the distance between longer pronunciations will generally be greater than the distance between shorter pronunciations, the sum of alignment operation costs is normalized by the length of the alignment:

Savannah	æ	ə	f	t	ə		n	ʉ	n
Lancaster	æ		f	t	ə	r	n	ʉ	n
Cost		1				1		1	

After the normalization the distance between the pronunciations of *afternoon* in Savannah and Lancaster is 0.33 [(1 + 1 + 1)/9]. We obtain a characterization of the distance between two dialects by computing the mean of their word distances (in this study we used 125 words). This procedure has been followed for all varieties to obtain an overall matrix of pairwise distances between them. Calculations have been performed with L04, a freely distributed dialectometric software package (available at http://www.let.rug.nl/kleiweg/L04/)

The simplest versions of such a method are based on a notion of phonetic distance in which phonetic comparison is binary: Nonidentical phones contribute to phonetic distance, and identical ones do not. Thus the pair [a, i] counts as different to the same degree as [e, ɛ]. In more sensitive versions phones are compared on the basis of their feature values, so the pair [a, i] counts as more different than [e, ɛ] because the second pair of vowels is more similar. The method that we use in this paper is more refined because the cost system is computed on the comparison of spectrograms of the sounds. A spectrogram is a representation of the acoustic signal in which sound intensities are tracked for a range of frequencies and times. With spectrograms it is not necessary to make decisions about the weight of different features.

We should note that such an approach is perceptually oriented, meaning that it emphasizes the differences that a speaker can hear in someone else's speech. We add that small differences can play a relatively strong role compared to larger differences in dialect perception, suggesting the use of natural logarithmic (ln) segment distances. Final distances were computed as

$$\frac{\ln(\text{distance} + 1)}{\ln(\text{maximum distance} + 1)} \times 100. \tag{3}$$

To reckon with syllabification in words, we adapted the Levenshtein algorithm so that a vowel can be aligned only with another vowel and a consonant only with a consonant. The only exceptions are [j] and [w], which can also be aligned with vowels; [i] and [u], which can also be aligned with consonants; and central vowels, which can match sonorants. All other alignments are ignored.

The measurements of phonetic differences are consistent for large samples of words [Cronbach (1951) $\alpha > 0.96$ for 100-word sample sets] and have been validated with respect to scholarly tradition (Heeringa et al. 2002) and, again, with respect to lay dialect speakers' judgment of dialectal distance (Gooskens and Heeringa 2004). Gooskens and Heeringa (2004) showed that the measurement correlated highly with lay speakers' judgments ($r = 0.80$). In addition, the technique has now been applied to Norwegian, American English, German, Sardinian, Bulgarian, and Bantu languages of Gabon. Interestingly, the same Levenshtein algorithm has been applied extensively to measurement differences in long genetic strings (Sankoff and Kruskal 1999).

Visualization of Diversity

Multidimensional Space: Principal Components Analysis. The topology of the 226 surname samples in the principal components analysis (PCA) was applied to the data to graphically detect possible patterns of similarity between the 226 surname samples and the 252 dialect samples. Principal components analysis involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability (variance) in the data as possible, and each subsequent component accounts for as much of the remaining variability as possible.

Multidimensional relationships between items can be seen in a bivariate or trivariate plot (if two or three axes are plotted against each other). The analysis was performed with the Excel applet GenAlEx (Peakall and Smouse 2001; available at http://www.anu.edu.au/BoZo/genAlEx).

Geographic Analysis: The Monmonier Algorithm. When sampling locations are known, it is possible to identify the areas where a given variable shows an abrupt rate of change. We use a computational geometry approach, which uses computed distances (surname, linguistic) to identify the locations of barriers and

which can therefore also show where the geographic patterns of two or more variables are similar. Inspired by this idea, we have implemented anew Monmonier's (1973) maximum difference algorithm (Manni 2004; Manni et al. 2004), and to identify genetic and linguistic barriers, we have used the program Barrier, version 2.2 (available at http: // www.mnhn.fr / mnhn / ecoanthropologie / software / barrier .html). To avoid ambiguity, we stress that we use the term *boundary* synonymously with *barrier*. Basically, the Monmonier method is based on a triangulation connecting all the samples; then the distance measures (genetic, linguistic) are associated with each edge of the triangulation, and a barrier is traced perpendicularly to those edges that have the highest distances. Because of space limitations, we do not describe the method in detail here because the provided link gives access to exhaustive documentation.

To test the confidence with which we can view the barriers in a genetic or linguistic landscape, we implemented a significance test in the software by means of bootstrap analysis. Using the Monmonier algorithm, we repeated the procedure of finding boundaries using matrices computed on data sets from which random elements had been removed and in which other elements, randomly selected as well, appear more than once. As with phylogenetic trees, a score is associated with all the different edges that constitute barriers, thus indicating how many times each edge is included in one of the boundaries computed from the N matrices (typically $N \geq 100$). In other words, if we have 100 matrices and we want to compute the first barrier, 100 separate barriers will be obtained. These 100 different barriers (different in the sense that they have been computed on matrices obtained from modified data sets) are displayed in a single picture by increasing the thickness of their edges in proportion to the number of times they belong to one of the 100 barriers. If a pattern exists, whatever the modification of the original data set, barriers should repeatedly emerge in certain areas of the plot. If barriers emerge everywhere in the plot, then the results may not be robust (in terms of geographic differentiation). The issue is similar to the use of bootstrap in phylogenetic trees (Felsenstein 1985), and similar advantages accrue to this way of computing barriers—notably, the confidence of the postulation of the barrier is reflected in the visualization. The method has been applied to original distances or to matrices of residuals.

We carried out the bootstrap test for both surname and linguistic data, but in addition, we are preparing a more exhaustive manuscript focusing on the bootstrap approach in dialectologic classifications.

Results

Surnames

Lasker Distance Versus Nei Distance: A Comparative Test Including or Excluding Polyphyletic Surnames and Rare Surnames. We computed two distance matrices, one from the original data set of all Dutch surnames and another from the remaining surnames after the withdrawal of polyphyletic ones. This step was

repeated twice, once using Nei measures and once using Lasker measures. The correlation between matrices, computed with or without polyphyletic surnames, was higher when analyzing Nei distances (MT $r = 0.913$, $p < 0.001$) and lower when correlating matrices of Lasker distances (MT $r = 0.791$, $p < 0.001$) [MT stands for Mantel test (Mantel 1967), the significance of which has been computed according to Manly (1997)]. The calculations of the statistical significance for all the correlations mentioned in this article are based on this test. Furthermore, average distances were higher, as expected, when excluding polyphyletic surnames from the data set, whatever the estimator. Similarly, the matrices computed after the elimination of polyphyletic surnames from the database correlate better with geographic distances; the correlation with Lasker distances improves from MT $r = 0.438$ to MT $r = 0.691$, $p < 0.001$; and the correlation with Nei distances improves from MT $r = 0.497$ to MT $r = 0.563$, $p < 0.001$.

Nei distances seem less sensitive, with respect to the change in correlations, to the presence of polyphyletic surnames than Lasker distances, and we noticed that the principal components and barrier analyses are more satisfactory when polyphyletic surnames are excluded from the database because a higher percentage of variance is explained. Finally, we also tested the influence that rare surnames had on the computation of matrices. We found that their contribution to the final computation of matrices was irrelevant. Whatever the measure, the correlation between a distance matrix computed by keeping rare surnames and another distance matrix obtained by eliminating rare surnames from the whole data set approached 1. We note that the polyphyletic and rare surnames correspond to a similar fraction of individuals in each of the 226 samples; therefore their exclusion does not bias the data set, because the sample size for each sample after the withdrawal is proportional to the initial size. Overall, the described results prompted us to adopt Nei distances and to base our study on them.

Multidimensional Analyses of Surname Variability (Nei Distance). The topology of the 226 surname samples in the principal components plot of Figure 2 is reminiscent of the geography of the Netherlands (see Figure 1 for a geographic map of the Dutch provinces): A well-defined Limburg cluster and a cluster composed of North Brabant samples are apparent. The remaining eastern and western samples are close to each other in a continuous swarm of points, whereas Zeeland samples are intermediate. A more detailed analysis of the topology of the plot reveals that within the swarm there is no overlap between the topological area occupied by the northeastern and northwestern samples. Furthermore, the topology of samples suggests more heterogeneity in the south of the Netherlands than in the north, where samples are visualized closer to each other. The two axes account for approximately 30% of the total variance, and further axes (even through the tenth) still account for significant fractions of the total variance. The low fraction of variance explained by the first two axes—a frequent phenomenon when large numbers of samples are analyzed—means that Figure 2 is less than optimally representative and suggests a rather complex topology of samples in the multidimensional

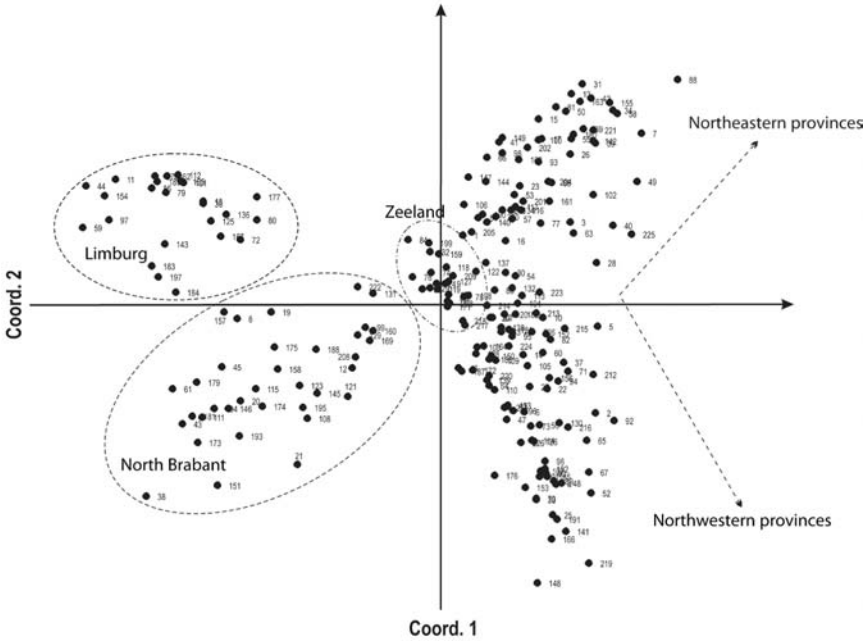

Figure 2. Principal components analysis of the surname differences in the Netherlands (Nei distances). Two almost distinct clusters corresponding to the North Brabant and Limburg samples can be identified. The remaining samples, belonging to the other provinces, cluster in a single swarm of points. Further details can be found in the text. The first axis explains 17.6% of the total variance, and the second axis accounts for 11.7% of it. The third and fourth axes (not shown) explain 5.8% and 4.3% of the total variance and highlight the diversity of the Limburg and Zeeland provinces, respectively. Further axes (fifth, sixth, seventh) point, in different ways, to the differences between the north and the south of the Netherlands. Samples are numbered as in Barrai et al. (2002).

space that cannot be straightforwardly projected to two dimensions. This suggests that geography is not the only factor influencing surname variability. Nevertheless, Figure 2 still provides a reasonable first approximation of overall variability, and in fact further axes point to the specificities of both Limburg and Zeeland and, more generally, to the differences existing between the northern and the southern parts of the country.

To understand the geographic variability of surnames, given that general correlations are not informative about local variation, we analyzed the surname distance matrix with Monmonier's algorithm (not shown). The barriers computed highlight some differentiation zones in the northeastern provinces and along the northern border of the Limburg and Dutch Brabant provinces. Moreover, the Zeeland province appears fragmented, indicating that surnames are heterogeneous in

the area with important differences from one location to another. These conclusions are reinforced by the analysis of 100 bootstrap matrices computed after a resampling procedure of original surname data (thick black lines in Figure 3). Bootstrap analysis leads to a clearer picture because some minor barriers, previously computed from the whole data set, in the northern part of the Netherlands and in Zeeland, disappear. In Figure 3 we note the presence of a major barrier across the former Zuiderzee (the internal sea in the northern Netherlands, visible in Figure 1, now called IJsselmeer after its polder and the construction of a dam that separates it from the Northern Sea), a result that will be discussed in what follows.

To focus on the variance that is not explained by geographic distance, we computed a general regression between geographic and surname distances after a double logarithmic transformation, $\ln y = 0.155 \ln x + k$, which is equivalent to $y = \exp(\ln 0.155x)\exp(k) = 0.155cx$, where $c = \exp(k)$. Thus we computed the expected surname distance, according to geography, between two sampled localities. The residual distance between observed and expected values, either positive or negative, reflects the influence of phenomena other than geography (e.g., history, systematic errors in data recording when several scientists are involved in fieldwork). In Figure 3 (solid gray lines) we show the Monmonier analysis of the matrix of such residuals. Besides some local barriers (barriers 5, 7, 9, 12–14, and 16), previously observed patterns are confirmed, with the exception of the IJsselmeer boundary, which disappears. Methodologically, it is interesting to note that the IJsselmeer boundary was traced across some of the longest edges of the Monmonier triangulation [a similar issue was discussed by Manni et al. (2004, p. 16)]. We can conclude that the IJsselmeer boundary mirrors a surname differentiation that, given the longer length of the Delaunay triangulation edges (triangles visible in the background of Figure 3) connecting the shores of the IJsselmeer, was expected when comparing the samples on the opposite sides of this inland sea. Seen from this perspective, the IJsselmeer does not seem to have been a substantial geographic barrier to internal migrations.

Dialects. Using an identical methodology, we analyzed the dialect data of the Netherlands. The matrix of Levenshtein dialectometric distances is visualized in the two-dimensional principal components plot of Figure 4, which suggests good geographic structure in the dialect data. Low Saxon and Low Franconian dialects, the two major groups, are displayed in separate clusters, and Frisian samples are represented by three different clusters that describe the (rural) Frisian, archaic Frisian, and Friso-Franconian varieties. We recall that Frisian is one of the two official languages of the Netherlands (the other is Dutch). Intermediate between the Friso-Franconian and Low Saxon dialect varieties we find a small Friso-Saxon group. Gray circles represent varieties spoken in central Gelderland, and open circles correspond to varieties of the Dutch province of Zeeland (Figure 4). The first and second axes account for 40.8% and 36.7% of the total variance, respectively, thus indicating that the two-dimensional plot reflects the multidimensional shape

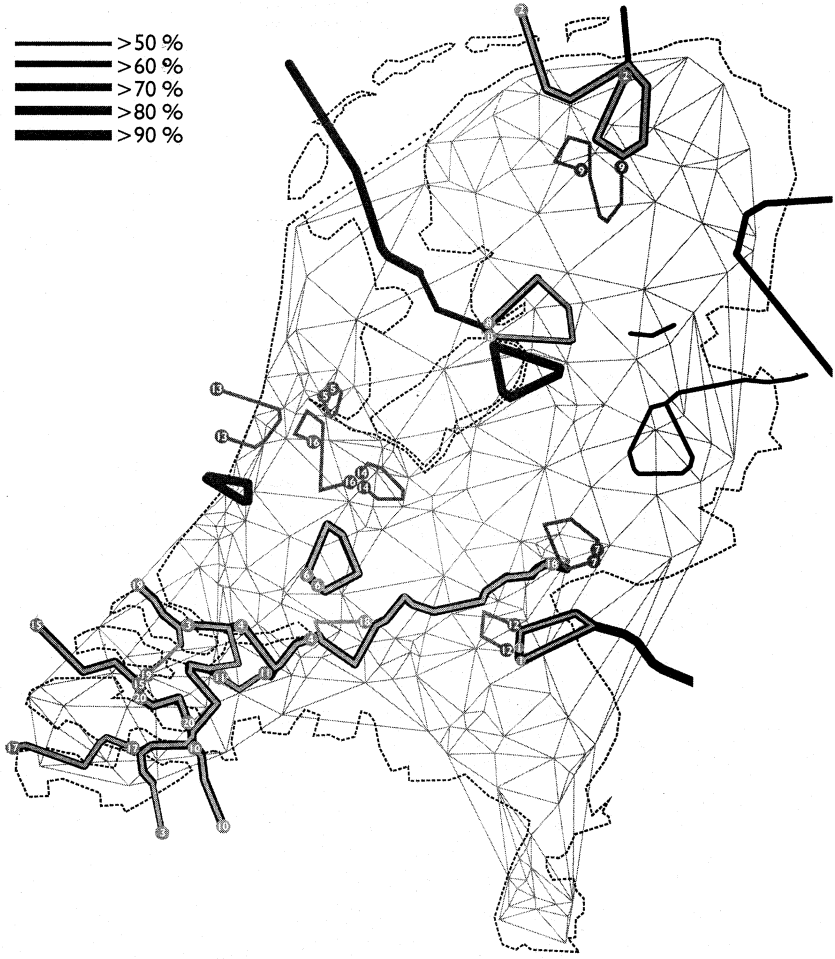

Figure 3. Comparison of barriers detected on the basis of surname distances: Thick black lines of variable width correspond to barriers obtained with the Monmonier algorithm on 100 matrices of surname distances computed according to Nei's method. The different matrices were computed using a bootstrap resampling of original surnames. Only the first 20 barriers of each matrix are shown (2,000 barriers in total). The thickness of barriers is proportional to their bootstrap score, and barriers whose score is lower than 50% are not shown (see scale). Gray lines correspond to barriers obtained from a matrix of residual surname distances. After a linear regression between the logarithms of geographic and Nei distances, we computed the expected surname distance according to the regression. Such values were subtracted from observed distances, thus leading to the residuals. The first 20 barriers are shown (numbered at both extremes from 1 to 20). The Delaunay triangulation is visualized by a light gray network.

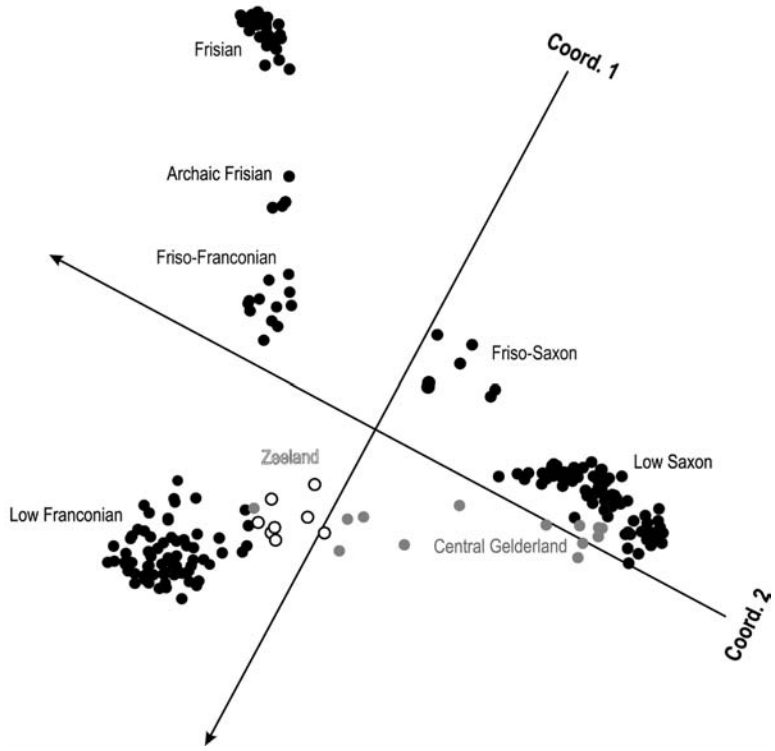

Figure 4. Principal components plot on the basis of 252 Dutch dialects. Low Saxon and Low Franconian dialects are grouped into separate clusters, whereas Frisian samples are represented by three different clusters that describe the (rural) Frisian, archaic Frisian (Hindeloopen, Schiermonnikoog, and Terschelling Island), and Friso-Franconian varieties (Frisian cities, Midsland, Ameland Island, and Het Bildt). Intermediate between Friso-Franconian and Low Saxon we find a small Friso-Saxon group (Westerkwartier and Stellingwerf). For a rough mapping of Franconian and Saxon varieties, see Figure 5. Gray circles represent varieties spoken in central Gelderland, and open circles correspond to varieties of the Dutch province of Zeeland. The first and second axes account for 40.8% and 36.7% of the total variance, respectively. Recall that the Dutch area is traditionally divided into Low Saxon and Low Franconian dialect varieties. A further group is represented by Frisian varieties; Frisian is a language distinct from Dutch and has its own dialects.

of the data accurately. We do not describe the linguistic classification of dialect varieties in more detail here because they have been fully discussed by Heeringa (2004).

Not unexpectedly, Monmonier boundaries (Figure 5) confirm the results of the principal components plot for the most part. We find a northwestern Frisian area (surrounded by barriers 1 and 2), a small northeastern area (part of the pro-

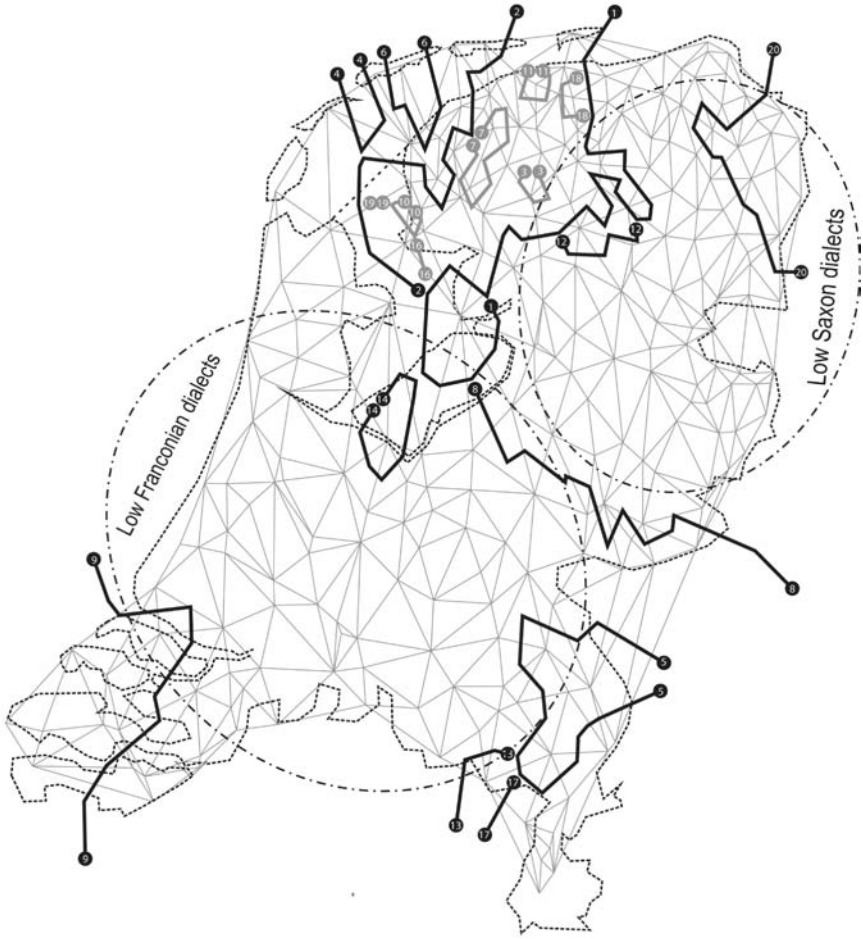

Figure 5. Barriers (bold black lines) obtained with the Monmonier algorithm on a matrix of dialect (Levenshtein) distances between 252 localities. The first 20 barriers are shown (numbered from 1 to 20). A light gray network visualizes the Delaunay triangulation. Boundaries identify areas corresponding to Friesland (local barriers corresponding to different Frisian varieties are displayed in gray to provide a clearer overall representation) and to parts of Zeeland and Limburg. On a wider scale it appears that some major barriers distinguish the geographic locations where Low Franconian and Low Saxon varieties are spoken (see labels).

vince of Groningen surrounded by barrier 20), a larger northeastern area (Low Saxon) adumbrated by barriers 1 and 8, a large more or less southwestern area (Low Franconian), a small southwestern part (province of Zeeland, barrier 9), and a small area in the southeast (part of the province of Limburg encircled by bound-

aries 5, 13, and 17). According to scholarly tradition, one of the major barriers was expected to lie along the border between the Low Saxon and Low Franconian varieties; we recognize it in Figure 5 as barrier 8 because barrier 1delimits just the Frisian area. We recall that the Monmonier algorithm finds barriers in a hierarchic order; therefore barrier 8 is the eighth computed among the 20 displayed. Now, the bootstrap test of barrier robustness (where original word transcriptions are resampled; not shown) does not give a special emphasis to this barrier, because its ranking decreases from the eighth to the sixteenth, a result that may derive from traditional views of dialect variability in the Netherlands. Otherwise, the bootstrap test confirms the patterns and the ranking of the remaining barriers plotted in Figure 5. Fragmentation is found in Friesland because of the well-known similarity among the urban Friso-Franconian varieties that are closer to Dutch.

As with the surname data, we continued the analysis by computing, after a double logarithmic transformation, a regression model ($\ln y = 0.287 \ln x + k$) between geographic and Levenshtein distances to obtain the matrix of residuals that is plotted in the principal components analysis (not shown). This is a novel treatment of the linguistics data because previous attempts to describe the relation between geography and linguistic variables ended with models that are different from the power law relation that is standard with genetic and surname data. Séguy (1971) described the relation between linguistics and geography as $\text{ling} = 0.5(\text{geo})$ and Heeringa and Nerbonne (2001) as $\text{ling} = \ln(\text{geo})$.

In proceeding this way, we are applying to linguistics the concept of isolation by distance that was first introduced in genetics by Wright (1943) and later developed by Malécot (1955). Malécot developed a model in which the number of migrants from one location to another is a function of the geographic distance between two areas, thus explaining the increase of genetic distances by increased geographic distances. Interestingly, the similarity between genetic and linguistic data can be pushed further because in both cases the correlation with geographic distances is not linear and the same logarithmic transformation is applied to both data sets to obtain an improved sublinear model. Anyhow, in linguistics it is difficult to distinguish the analyses based on the logarithms of geography from those postulating a power law with a fractional coefficient. The empirical predictions of the two make accord with one another in the domain under study.

We find that the remaining structure in the multidimensional principal components plot, computed on residuals (not shown because of space limitations), is still striking and appears at some points to reflect geography after all, perhaps suggesting that the influence of geography is not constant. Further research and some creativity may be necessary to address such new issues, which might a priori be expected to shed light on the mechanisms of linguistic differentiation through space.

The shape of the Monmonier barriers, based on the matrix of residual Levenshtein distances (Figure 6), confirms the pattern previously found in Zeeland (boundaries 11, 15, and 19) as well as the Saxon dialect area that is still surrounded by several barriers (10 and 14). The northern part of the Saxon area seems less

Figure 6. Barriers (bold black lines) obtained with the Monmonier algorithm on a matrix of residual dialect distances (to be compared with the identical analysis on the original matrix in Figure 5). The provinces of Friesland and Groningen appear as linguistically continuous, but see the text for further details. The first 20 barriers are shown (numbered 1 to 20). As in Figure 5, barriers corresponding to different Frisian varieties are displayed in gray. The Delaunay triangulation is visualized as a light gray network. As in Figure 5, the lines in light gray are simply not emphasized in the discussion; they correspond to some distinct Frisian varieties.

contoured compared to Figure 5, because its northern part (corresponding to the province of Groningen) is now geographically continuous with Friesland but is separated from the province of Drenthe by boundaries 1 and 3. As in the original matrix of Levenshtein distances (Figure 5), Friesland is still fragmented (as shown by the shape of barriers 2, 5, 7, and 9) because of the dialect islands of the urban Frisian mixed varieties (Friso-Franconian) in the Frisian dialect continuum. A completely new feature of Figure 6 is boundary 16, which begins on the left (west), follows the border between North and South Holland, and then veers south to pass vertically through the provinces of Utrecht and North Brabant. Even though this border has not been discussed extensively in previous studies (so we cannot easily compare alternative explanations of its meaning), it is nonetheless interesting because it can be attributed to heterogeneous transcriptions (Heeringa 2004).

Cross-Comparison of Surname and Dialect Variability. To be sure, distributions of linguistic and surname variation correlate significantly ($r = 0.417$, $p < 0.001$, based on the 74 locations common to the surname and dialect samples). But such correlation just reflects the link existing between linguistic and geographic distances ($N = 252$; $r = 0.407$, $p < 0.001$) on the one hand and between surname and geographic distances ($N = 226$; $r = 0.507$, $p < 0.001$) on the other. Because both pronunciation and surnames correlate strongly with geography, they seem to be correlated with each other. But there is no correlation between matrices of residual Nei and Levenshtein distances; that is, there is no correlation between surname and linguistic differences once their common dependence on geographic distance has been included in a statistical model.

Discussion

The major aim of this study was to determine to what extent the motifs of geographic variation of surnames overlap with those of linguistic diversity because family names are a specific part of language. Following the geographic approach used here and thus focusing on the barriers where geographic influence is insufficient as an explanation of genetic or linguistic difference, we note no striking correspondences between the two markers, for example, in comparing the areas of differentiation in Figures 2 and 4 and in Figures 3 and 6. We can then conclude that the pressure of the linguistic environment on surname distributions is absent or negligible and that surname variability differs from linguistic variability, which shows that the social contacts reflected by dialect varieties do not seem to be related to the demographic history of the populations speaking those dialects. Because the two processes, social and demographic, took place over a similar time frame, we can reasonably extend our claim, as a null hypothesis, to any future work addressing the comparison of surnames and linguistic markers.

If we describe the situation from the point of view of a multiple regression model in which we test geography and surname differences as independent

predictors of linguistic distance, then the two predictors are collinear, leading a hasty analysis to attribute influence to both predictors, whereas a careful analysis in fact displays none. The correlation between linguistic and surname markers is entirely explained by their common collinearity with geography. We can strengthen our own conclusion that in the Netherlands there has been no demonstration of a relation between linguistics and surnames by noting the differences between the model used here and the models used in our earlier dialectometric work. Nerbonne et al. (1999) calculated a correlation coefficient between linguistic and corresponding geographic distances of MT $r = 0.656$, ($p < 0.01$) using the full data set (which includes the Flemish part of Belgium) with a sample that is similar to the one used in this paper, but they used a logarithmic regression model rather than the power law (doubly logarithmic) model used here. The logarithmic model ($r^2 \approx 0.40$) clearly explains a great deal more variance than the power law model ($r^2 \approx 0.16$). We conclude from this that the optimal linguistic model takes a logarithmic form, in distinction to the power law relations favored in genetics. This reinforces our main conclusion, namely, that the linguistic and genetic patterns of variation are different, even if they are both conditioned strongly by geography.

Our results differ strikingly from those of a similar study comparing surnames and dialects in France by Scapoli et al. (2005). But we suspect that Scapoli and colleagues failed to control their matrices of genetic and linguistic distances for common geographic conditioning, leading them to the incorrect conclusion that language similarity is an indicator of genetic kinship even at local levels. This may occasionally be true but needs to be systematically verified by analyzing residuals.

Concerning the Netherlands, the only close match between the variation of surnames and dialects is found in the province of Zeeland, which is also geographically separate from surrounding areas (see Figures 3, 5, and 6). This special status of Zeeland may be due to its geography, because it was constituted by several islands that, starting in the 14th century (*Atlas van Nederland* 1986), increased in size and, thanks to land reclamation efforts, eventually turned into peninsulas at the beginning of the 20th century. Relative social and geographic isolation, together with an economy based on fishing and trade, may have maintained and reinforced a closed social structure still visible in surname and dialect variability—a diversity that is also mirrored by the different agriculture practices between insular Zeeland and Zeeland Flanders (see Figure 1). Finally, an additional and complementary explanation is represented by more intense contacts with the adjacent western Flemish area (Belgium).

The computation of a regression model leading to matrices of residuals is expected to better illuminate demography (surnames) and social patterns (dialects), both of which are related to history (in a broad sense) rather than to geography. As a consequence, we can interpret the surname barriers found along the northern borders of Zeeland, North Brabant, and Limburg as the results of historical phenomena. The significance of such major separations is confirmed by bootstrap ma-

trices visualized through the Monmonier algorithm and by the analyses of residual distances (barriers 4 and 18 in Figure 3)—which brings up new issues.

As we said earlier, the geographic variation of surnames only mirrors demographic phenomena, without showing any effect of linguistic environment. Therefore, when we seek explanations for such barriers, which linguistic culture does not support, we must turn to other factors. In this case we are struck by the correspondence between the border induced by common surnames and the border of the Protestant/Roman Catholic area of the Netherlands (Figure 7). The strength of the surname border suggests that the frequency of intermarriage between Catholics and Protestants was low. This religious distinction may have acted as a social boundary, thus increasing surname differences between populations on the border's sides. The fact that there is no linguistic evidence (Figure 5 and 6) of such separation means that more casual social contacts and interchange were not diminished between Catholic and Protestant populations. Communication proceeded despite a profound social cleft.

Intuitively, the observed incoherence between markers of genetic relatedness as surnames and linguistic space distributions can be regarded as misleading, once culture (language) is assumed to be matrilineally transmitted and surnames are paternally transmitted. This was a concern expressed by Bob Shackleton, one of the scholars who assisted us in our research. In other words, his question was, Would our findings have been the same if surnames were maternally transmitted? Some readers may remember a popular study pointing to the greater dispersion of females compared to males (Seielstad et al. 1998). Such results, based on the comparison between specifically paternal (Y-chromosome) and specifically maternal (mitochondrial DNA) genetic markers, were explained in terms of patrilocality. Even if alternative explanations (Dupanloup et al. 2003) and different conclusions (Wilder et al. 2004) have been provided since, we note that such debate mainly concerns the frame of prehistorical times instead of the recent time frame of surname data.

Even keeping in mind that matrimonial migrations, at least in rural Europe, generally consist of only a few kilometers and that we are dealing with differences that can be traced back for only eight generations, it is likely that patrilocality plays a role in our surname data set, meaning that females move more than males. Nevertheless, patrilocality is counteracted by the observation that sons inherit their propensity to migrate from their fathers, whereas such transmission is largely absent among women (Gagnon et al. 2006). Gagnon and colleagues concluded that the pool of migrants is not a random sample of the whole population because there is an intergenerational dependency in the probability of migration that can be explained by social factors (Heyer 1993). Once settled somewhere, the newcomers seldom become the owners of the land (or of other means of production), so their sons are more likely to move out. In this process their new Y-chromosome variants, together with their surnames, tend to disappear in the next generation. Differently, daughters of immigrants can become part of the new community by marriage and therefore have a higher chance of enriching the local pool of genetic diversity.

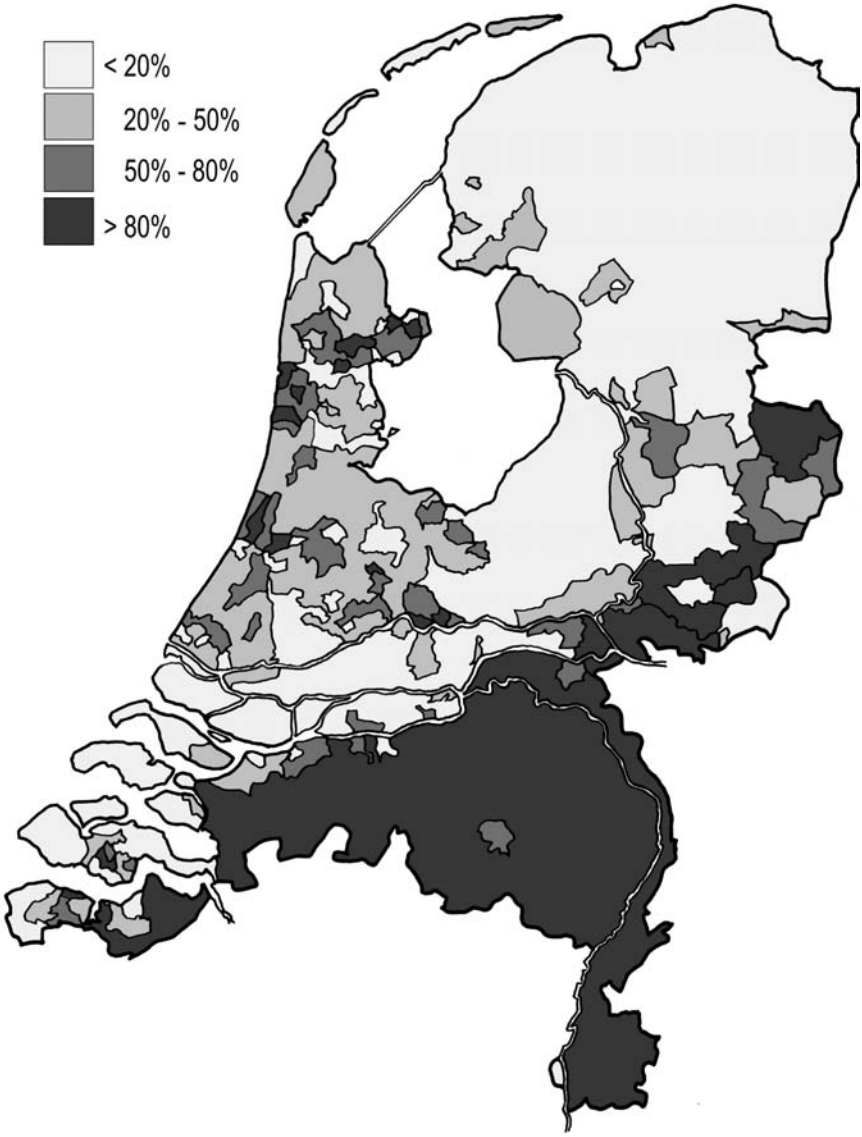

Figure 7. Frequency of Roman Catholics in the Netherlands in 1954. Redrawn from van Heek (1954).

To answer the thorny question of our colleague: If women transmitted Dutch surnames, we would have computed pairwise surname distances that were smaller than patronymic ones (more migration equals more surname homogeneity). The

general portrayal of such a hypothetical matrilineal-inferred surname structure would have been similar, albeit less detailed—more migrations imply lower local differences—unless we allow that females and males might migrate in different directions [a hypothesis that seems excluded by Kok (1997) concerning the Netherlands].

Linguistically, the role of the mother in language transmission is a valuable area for further research, because, in wider studies comparing genetic and linguistic differences, culture is often assumed to be maternally transmitted. Actually, most linguistic research emphasizes the importance of the peer group, outside the immediate family, in influencing adolescent patterns of speech, and the general suspicion is that these are normally then resistant to change in later life. This is a question that may be of special interest in genetic samplings performed according to ethnological and linguistic criteria.

Besides the major research question of this paper, we think that some methodological outcomes should be reviewed. First, the use of matrices of residual linguistic distances obtained after the computation of a regression between geographic and linguistic distances has been rewarding. The approach has enabled us to visualize the geographic affinity of the province of Groningen to the Frisian-speaking area (see Figure 6). We recall again that Frisian and Dutch are two distinct languages, with Frisian being spoken in the north of the Netherlands, almost exclusively in the province of Friesland (other Frisian speakers are found in Denmark and Germany). This closer relation probably mirrors the early linguistic history of the Groningen area, where some Frisian varieties were last spoken in the early part of the 16th century [see Hoekstra (2001, p. 139) and Niebaum (2001, p. 431)]. Besides some few contemporary phonetic features, presently there is no linguistic evidence that a different language was once spoken in this area, thus underscoring the great potential interest of the methodological approach we undertook. Spruit's (2006) analysis of syntactic variability showed that the north of the Netherlands appears much less heterogeneous than it does in lexical and phonetic analyses. We are aware that our approach to the computation of residuals, in both data sets, is mainly empirical and that a deeper insight into the relations between real data and theoretical models describing isolation by distance is needed. This area of investigation represents our next target because, concerning genetics, only two recent papers have addressed it properly (Rousset 1997, 2000) and, concerning linguistics, such approaches are still to come.

A second methodological outcome concerns the use of a bootstrap methodology to assess the robustness of data. We experienced that barrier analyses based on resampled data correspond, to a large extent, to the patterns obtained from matrices of residuals, thus suggesting that a bootstrap approach is a better approximation to the patterns of variability of both dialect and surname variation than the analysis of a single matrix obtained from the entire data set. If resampling procedures are of general use in phylogenetic or phenetic studies, they were surprisingly never applied to surnames, even though such studies belong to the same disciplinary area. Similarly, dialectologists do not normally use resampling proce-

dures to assess the robustness of linguistic patterns because, concerning language varieties, computational studies are far more recent.

Although resampling procedures are not common among dialectologists, McMahon and McMahon (2005) applied them to historical linguistics, and Kleiweg et al. (2004) experimented with techniques that add random degrees of noise to distance matrices in an effort to counter the instability of clustering classifications. Nerbonne et al. (2007) showed that this technique results in the same classifications as bootstrap analysis.

We hope that future directions of investigation will be focused on an interdisciplinary understanding of linguistics and biology, discussed at length by Goebl (1996), and in particular of the interrelations existing between surnames and dialects together with real genetic markers on wider scales. Computational linguistics and phylogenetic methods applied to languages (Forster and Renfrew 2006) are setting new standards in multidisciplinary studies. They offer a key to understanding the geographic dissemination of both cultural features and hereditary markers through generations, with surnames standing at the interface between them.

Acknowledgments We are indebted to Pierre Darlu for compiling the program DISNEI to compute different surname distances, to Bob Shackleton for commenting on the manuscript, to Isabelle Dupanloup for inspiring us, to Alain Gagnon for providing insightful references, and to Raphael Leblois for discussing the regression procedures. We also thank George Welling for his critical remarks concerning the history of the Netherlands and Meindert Schroor for improving our geographic insight. A special thanks to Evelyne Heyer and Serge Bahuchet, who provided the best working environment to carry out such cross-disciplinary collaboration. John Nerbonne was supported by an Invited Scholar grant from the National Museum of Natural History of France. This study is an outcome of a larger research program, financed by the Wenner Gren Foundation (grant 7247 awarded to F. Manni), concerning the Netherlands.

Received 7 December 2006; revision received 19 September 2007.

Literature Cited

Atlas van Nederland. 1986. 's-Gravenhage, the Netherlands: Stichting Wetenschappelijke.

Barrai, I., A. Rodriguez-Larralde, F. Manni et al. 2002. Isonymy and isolation by distance in the Netherlands. *Hum. Biol.* 74:263–283.

Cavalli-Sforza, L. L., and W. Bodmer. 1971. *Human Population Genetics.* San Francisco: Freeman.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press.

Chambers, J. K., and P. Trudgill. 1998. *Dialectology*, 2nd ed. Cambridge, U.K.: Cambridge University Press.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334.

Crow, J. F. 1980. The estimation of inbreeding from isonymy. *Hum. Biol.* 52:1–4.

Dupanloup, I., L. Pereira, G. Bertorelle et al. 2003. A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *Mol. Ecol.* 13:853–864.

Dyen, I. 2004. Personal communication based on the discussion of the paper "Johannes Schmidt's 'wave theory' and the Homomeric method," presented at the workshop Phylogenetic Methods and the Prehistory of Languages, McDonald Institute for Archaeological Research, Cambridge, U.K., July 9–12, 2004.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.

Forster, P., and C. Renfrew, eds. 2006. *Phylogenetic Methods and the Prehistory of Languages.* Cambridge, U.K.: McDonald Institute for Archaeological Research.

Gagnon, A., B. Toupance, M. Tremblay et al. 2006. Transmission of migration propensity increases genetic divergence between populations. *Am. J. Phys. Anthropol.* 129:630–636.

Goebl, H. 1996. La convergence entre fragmentations géo-linguistique et géo-génétique de l'Italie du Nord. *Rev. Linguist. Romane* 60:25–49.

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

Heeringa, W. 2004. *Measuring Dialect Pronunciation Differences.* Groningen, the Netherlands: Groningen University Press.

Heeringa, W. J., and J. Nerbonne. 2001. Dialect areas and dialect continua. In *Language Variation and Change*, D. Sankoff, W. Labov, and A. Kroch, eds. New York: Cambridge University Press, 375–400.

Heeringa, W., J. Nerbonne, and P. Kleiweg. 2002. Validating dialect comparison methods. In *Classification, Automation, and New Media*, W. Gaul and G. Ritter, eds. Heidelberg, Germany: Springer, 445–452.

Heyer, E. 1993. Population structure and immigration: A study of Valserine valley (French Jura) from the 17th century until present. *Ann. Hum. Biol.* 20:565–573.

Hoekstra, E. 2001. Frisian relics in the Dutch dialects. In *Handbuch des Friesischen* [Handbook of Frisian Studies], H. H. Munske, ed. Tübingen, Germany: Niemeyer, 138–142.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics.* Dublin, Ireland: European Association for Computational Linguistics, 60–67.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge, U.K.: Cambridge University Press.

Kleiweg, P., J. Nerbonne, and L. Bosveld. 2004. Geographic projection of cluster composites. In *Diagrammatic Representation and Inference*, A. Blackwell, K. Marriott, and A. Shimojima, eds. Berlin: Springer, 392–394.

Kohonen, T. 1995. *Self-Organizing Maps.* Berlin: Springer.

Kok, J. 1997. Youth labour and its family setting, the Netherlands 1850–1930. *Hist. Family* 2:507–526.

Kruskal, J., and M. Liberman. 1983. The symmetric time-warping problem: From continuous to discrete. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence*, D. Sankoff and J. Kruskal, eds. Reading, MA: Addison-Wesley, 125–161.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10:707–710.

Malécot, G. 1955. The decrease of relationship with distance. *Cold Spring Harbor Symp. Quant. Biol.* 20:52–53.

Manly, B. F. J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, 2nd ed. London: Chapman and Hall.

Manni, F. 2001. *Strutture genetiche e differenze linguistiche: Un approccio comparato a livello micro e macro regionale.* Ph.D. thesis, University of Ferrara, Ferrara, Italy.

Manni, F. 2004. *Barrier, Version 2.2: Manual of the User.* Available at http://www.mnhn.fr/mnhn/ecoanthropologie/software/manual.pdf

Manni, F., and I. Barrai. 2001. Genetic structures and linguistic boundaries in Italy: A microregional approach. *Hum. Biol.* 73:335–347.

Manni, F., E. Guérard, and E. Heyer. 2004. Geographic patterns of (genetic, morphologic, linguistic)

variation: How barriers can be detected by using Monmonier's algorithm. *Ann. Hum. Biol.* 76:173–190.

Manni, F., W. J. Heeringa, and J. Nerbonne. 2006. To what extent are surnames words? Comparing the geographic patterns of surname and dialect variation in the Netherlands. *Literary and Linguistic Computing* 21:507–527.

Manni, F., B. Toupance, A. Sabbagh et al. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. *Am. J. Phys. Anthropol.* 126:214–228.

Mantel, N. A. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27:209–220.

McMahon, A., and R. McMahon. 2005. *Language Classification by Numbers.* Oxford, U.K.: Oxford University Press.

Monmonier, M. 1973. Maximum-difference barriers: An alternative numerical regionalization method. *Geogr. Anal.* 3:245–261.

Nei, M. 1973. The theory and estimation of genetic distance. In *Genetic Structure of Populations,* N. E. Morton, ed. Honolulu: University Press of Hawaii, 45–54.

Nerbonne, J., W. Heeringa, and P. Kleiweg. 1999. Edit distance and dialect proximity. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and J. Kruskal, eds. Stanford, CA: CSLI Press, v–xv.

Nerbonne, J., P. Kleiweg, W. Heeringa et al. 2007. Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In *Data Analysis, Machine Learning, and Applications*, C. Preisach, L. Schmidt-Thieme, H. Burkhardt et al., eds. Berlin: Springer (in press).

Niebaum, H. 2001. Der Niedergang des Friesischen zwischen Lauwers und Weser. In *Handbuch des Friesischen* [Handbook of Frisian Studies], H. H. Munske, ed. Tübingen, Germany: Niemeyer, 430–442.

Peakall, R., and P. E. Smouse. 2001. *GenAlEx, vs. 5: Genetic Analysis in Excel—Population Genetic Software for Teaching and Research.* Canberra: Australian National University.

Rodriguez-Larralde, A., C. Scapoli, M. Beretta et al. 1998. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Ann. Hum. Biol.* 25:533–540.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* 145:1219–1228.

Rousset, F. 2000. Genetic differentiation between individuals. *J. Evol. Biol.* 13:58–62.

Sankoff, D., and J. Kruskal, eds. 1999. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Stanford, CA: CSLI Press.

Scapoli, C., H. Goebl, S. Sobota et al. 2005. Surnames and dialects in France: Population structure and cultural evolution. *J. Theor. Biol.* 237:75–86.

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

Séguy, J. 1971. La relation entre la distance spatiale et la distance lexicale. *Rev. Linguist. Romane* 35:335–357.

Seielstad, M. T., E. Minch, and L. L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. *Nat. Genet.* 20:278–280.

Spruit, M. 2006. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing* 21:493–506.

van Heek, F. 1954. *Het geboorteniveau der Nederlandse Rooms-Katholieken.* Leiden, the Netherlands: Stenfert Kroese.

Wilder, J. A., S. B. Kingan, Z. Mobasher et al. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat. Genet.* 36:1122–1125.

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