# Dialectometry++

**John Nerbonne**
University of Groningen and Freiburg Institute of Advanced Studies, Germany

**William A. Kretzschmar, Jr**
University of Georgia, USA

## Abstract

Dialectology is one of the sub-disciplines in the humanities that embraced digital techniques early on. The use of computational and quantitative techniques in dialectology is known as 'dialectometry'. The present collection of articles contain several which proudly continue working within dialectometry's usual assumptions and toward its established goals, honing existing techniques, and experimenting with novel ones, but also, significantly, several articles that depart deliberately from earlier modes, returning to individual phenomena (as opposed to aggregates), examining new sources of data (not taken from atlases), applying dialectometric techniques to sociolinguistic and diachronic research questions, seeking explanations for geographic distributions in semantics and in complexity theory, and experimenting with techniques from spatial statistics, geographic information systems, and image analysis.

**Correspondence:**
John Nerbonne, Center for Language and Cognition, University of Groningen, P.O. Box 716, NL 9700 AS Groningen, The Netherlands
**E-mail:**
j.nerbonne@rug.nl

## 1 Introduction and Background in Dialectometry

Two leitmotifs have characterized dialectometry's approach to the study dialects, one explicit and the other largely implicit. First, dialectometry deliberately proceeds from the scientific conviction that dialectal data are too complex to be studied one phenomenon at a time, e.g. what word is used to refer to a particular body part, farm instrument, or domestic animal, or what sound is pronounced at the end of words such as 'floor', 'walking', or 'house'. The emphasis has therefore been on the analysis of large aggregates of differences, e.g. all the lexical differences in a sample of several hundreds. Second, dialectometry, unlike many other schools of thought struggling to understand language variation, has regarded the analysis of dialect atlas data as a duty incumbent on us, the heirs to the large collections of data for which previous generations of scholars have painstakingly compiled.

Dialectometry's effective birth can be precisely dated as the publication of Séguy's (1973) landmark paper 'La dialectométrie dans *l'Atlas linguistique de la Gascogne*'. We write 'effective birth' as there clearly were precedents in Haag's (1898, 1901) work (Streck and Auer, 2012). But while Haag's work received only little follow-up, Séguy's was taken up enthusiastically by Goebl (1982, 1984), who essentially codified the principles of dialectometric analysis for a decade and a half. Séguy and Goebl were concerned that the older studies on dialectology had focused so heavily on details that general relations were obscured. The forest of varietal relations could not be seen for the trees of details on lexical choice, pronunciation details, and occasionally, grammatical differences. A particularly bothersome aspect of the older work was that the more

doi:10.1093/llc/fqs062

precise the data analysis, the less attainable general principles seemed to become. They therefore broke with scholarly tradition and began to count the differences in their data rather than view them separately. They then focused on understanding the sums of differences between pairs of sites. Their 'aggregating' steps have largely been followed with the advantage that the relative differences among entire varieties (e.g. the speech of Gascony versus that of the Provence) emerge clearly and reliably. Worries about the relative importance of different data items or different data sorts (e.g. vocabulary versus pronunciation) could be more easily dismissed when it turned out that the aggregate relations among the varieties were not particularly sensitive to how one weighted the evidence. This was the major advance that Séguy and especially Goebl developed and applied over many years.

Naturally more was involved than mere counting differences. Goebl championed the analysis of the distributions of differences as seen from individual sites, not merely observing the center of the distribution (the mean difference to other sites), but also calling attention to the meaning of substantial skew in distributions. While most dialectometry has focused on the analysis of linguistic data seen categorically, Kessler (1995) and Nerbonne *et al.* (1999 and elsewhere) contributed a numerical measure of pronunciation distance that has been widely applied and validated (Gooskens and Heeringa, 2004; Heeringa *et al.*, 2006). See earlier collections of papers for many applications of dialectometry to questions in dialectology (Nerbonne and Kretzschmar, 2003, 2006; Nerbonne *et al.*, 2008).

A number of novel techniques for computational dialectology have arisen from computational linguistics. Kondrak's (2003) work on alignment appeared in a special issue on dialectometry (Nerbonne and Kretzschmar, 2003), and he later extended this to the more difficult task of detecting cognates in corpora of putatively related languages (Kondrak and Sherif, 2006). Eisenstein *et al.* (2010) examined postings (tweets) in Twitter in the USA and demonstrated that one could isolate a geographical signal, capitalizing especially on the fact that different topics dominate in different regions. Scherrer and Rambow (2010) experimented with knowledge-based, geo-referenced variation rules in order to identify Swiss German dialects based on their 'shibboleths', or typical pronunciations.

Meanwhile, dialectometric techniques have been successfully applied outside of dialectology proper, surely a further sign of success. Gooskens and colleagues have applied essentially dialectometric techniques to determine the intelligibility of closely related varieties, including closely related languages (Gooskens *et al.*, 2008, Kürschner *et al.*, 2008; Gooskens, 2012). Although the use in sociolinguistics has been only modest, there are encouraging attempts (Wieling *et al.*, 2011; Nerbonne *et al.*, 2013). Heeringa and colleagues have assayed language contact effects using dialectometric techniques (Heeringa *et al.*, 2010), and Sanders and Chin (2009) have measured the deviance in the speech of cochlear implant users employing the edit-distance measure developed for pronunciation difference measures in dialectometry. Others have applied dialectometric techniques to obtain the sound correspondences that form the basis of diachronic research in linguistics (Prokić *et al.*, 2009; List, 2012). Wichmann and Holman (2009) use a version of the pronunciation distance metric made popular in dialectology in their 'Automated Judgment of Similarity Program' (AJSP), which they and colleagues have applied to a number of problems in diachronic linguistics and typology. They cleverly restrict the phonetic alphabet from which transcriptions may be made.

Some papers in this volume hone existing techniques and others strike out in new directions. In the first category, we find papers on MDS, on the new software package DiaTech, and on comparing principal component analysis (PCA) to bipartite spectral graph partitioning (BiSGP), and in the second, more innovative group several articles that depart deliberately from the earlier leitmotifs, returning to individual phenomena (as opposed to aggregates) and exploring corpora as an alternative source of data (to atlases). Further we find papers applying dialectometric techniques to sociolinguistic and diachronic research questions, seeking explanations for geographical distributions in semantics and in complexity theory, and experimenting with techniques from spatial statistics, geographic information systems and image analysis.

These studies indicate variously and cumulatively that dialectometry continues to flourish!

## 2 The Dialectometry++ Papers

### 2.1 Technical improvements in dialectometry

Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler present their further work on using multi-dimensional scaling (MDS) in dialectometry in 'Defining Dialect Regions with Interpretations: Advancing the MDS Approach'. MDS takes as input a distance matrix, normally the aggregate distances among all the sites in a sample, and attempts to place them in a low-dimensional space (i.e. assign *x*, *y*, *z*, and perhaps more coordinates to them) in such a way that the input distances between sites correlate as well as possible with the distances implicit in the assignment of coordinates. Embleton (1987) first introduced MDS to dialectometry, and it has since become a standard technique (Nerbonne, 2010). It has been a remarkable empirical finding that three and sometimes even only two dimensions suffice to represent large tables of aggregate linguistic distances, often containing several hundred sites. This article will be of interest to the historian of dialectometry for its extensive bibliography on the use of MDS, including many papers from the collaborators in Canada that regrettably are not cited as often as they might be.

One can visualize two dimensions of an MDS solution easily enough by plotting the values as *x* and *y* coordinates in a scatter plot. This article describes software for visualizing MDS results in three dimensions. The software supports rotation in a three-dimensional visualization so that the representation may be examined from a variety of perspectives. As the authors point out, an advantage of the dynamic rotation is that it clarifies the contribution of the different dimensions: whenever one can see that the points lie in a plane—in some rotation or other—then at least one dimension is superfluous. As the authors note, the dynamic three-dimensional view is easier to interpret exactly than the colors that have often been used to project the three MDS dimensions on to two-dimensional geographic maps (Nerbonne *et al.*, 1999). Finally, the authors report on using their system in analyzing data from the *Romanian Online Dialect Atlas*. Not only were they able to detect heterogeneity in an unexpected area in the north of the country, but they also noted a surprising lack of correlation in the analyses based on two sorts of linguistic date, namely on lexical data on the one hand versus those based on morphophonemic data on the other. The latter will undoubtedly spark speculation as to what underlies the rift in the diffusion patterns of the two linguistic levels.

Gotzon Aurrokoetxea, Karmele Fernández-Aguirre, Jesus Angel Rubio, and Jon Sanchez have recently developed *DiaTech*,[1] a package to support dialectometry. Naturally they are aware of the *Visual DialectoMetry* (VDM) package,[2] constructed by Haimerl (1998, 2006) in close collaboration with Goebl (see above) and also of the web application *GabMap*,[3] which incorporates many of VDM's functions and also supports edit distance analyses of phonetic transcriptions (Nerbonne *et al.*, 2011). *DiaTech* is still under development, so the article is still somewhat programmatic, but Aurrokoetxea *et al.* clarify two of their most important goals in undertaking the development of *DiaTech*. One is different and more flexible ways of treating multiple responses and the second is the attempt to provide language-independent characterizations of dialect properties. Independent characterizations of dialect areas would allow precise comparisons, e.g. of the dialect diversity in different areas. (Grieve's and Pröll's papers (described later) also suggest characterizations of diversity.)

Multiple responses arise when field workers hear two different lexicalizations of the same concept at a single data collection site, or encounter two different morphological realizations or two different pronunciations. They occur frequently in dialect atlases, and therefore must be analyzed somehow by the dialectometrist. As corpus data are increasingly analyzed, the need for reliable and probative procedures for handling multiple responses, or indeed, frequency data, will only increase. Goebl (1997 and elsewhere) has hypothesized that multiple responses first arose in dialectal records as sociolinguistic interests began to play a role, but be that as it may, the analytical problem remains. Nerbonne and Kleiweg (2003) had noted that the simple procedure of averaging the distance of all pairs of items where multiple responses were involved would overrate distances between sites and argued that one ought to seek a 'covering set' of pairs from the two multiple response items with minimal cost. Aurrokoetxea *et al.* are concerned that this might underestimate the true linguistic distance between such sites and suggest an alternative. The differences are subtle, but it will be beneficial to stimulate more discussion.

Martijn Wieling, Robert Shackleton, Jr and John Nerbonne examine two alternatives for identifying dialect markers in 'Analyzing Phonetic Variation in the Traditional English Dialects: Simultaneously Clustering Dialect and Phonetic Features', namely cluster analysis and PCA on the one hand and BiSGP (Wieling and Nerbonne, 2011) on the other hand. While dialectometry originally focused on aggregates of linguistic differences, traditional dialectology has always attended to the details of dialectal differences. Naturally one can extend dialectometrical analyses to do just that, which has the advantage of enabling a closer comparison to results in traditional dialectology, namely by measuring the degree to which traits are representative and distinctive of a dialect region with respect to a language area. The latter, quantitative perspective is a further contribution of dialectometry to the study of language variation. Given that aggregates are composed of individual items, it was probably not to be expected that PCA and BiSGP, which is designed to cluster sites and features (items) simultaneously, would yield radically different results, and indeed they do not. From a typological view, it remains puzzling that the PCA process of identifying linguistic items that tend to co-occur (anywhere!) yields not only coherent groups of linguistic features but, indirectly, when one examines the sites in which the PCs are strong, geographically coherent areas. One might have expected to find some confounding due to typological influence. This might take the form of apparent relatedness due to typological similarity. One might imagine that areas that develop asymmetric vowel inventories for initially completely different reasons would then be fated to share further similarities for that reason. The authors do not examine their PCA results in search of such potential influences, but the clean geographical interpretation of the PCA results suggest that typological effects were not substantial. The authors conclude that the main benefit of BiSGP lies in its ability to identify geographical clusters of sites together with their corresponding linguistic features.

## 2.2 Innovative sources of data

Kristel Uiboaed, Cornelius Hasselblatt, Liina Lindström, Kadri Muischnek, and John Nerbonne present a dialectometric analysis of some syntactic aspects of Estonian in 'Variation of Verbal Constructions in Estonian Dialects'. Aside from adding to the still meagre dialectometric scholarship focusing on syntax (Spruit, 2008; Szmrecsanyi and Kortmann, 2009; Wiersma *et al.*, 2011), this study enhances the prospects for a marriage between corpus linguistics and dialectometry (Keune *et al.*, 2005; Scherrer, 2012; Szmrecsanyi, 2012), with its accompanying benefits of casting a broader net and of potentially including more natural data. While in some data collection projects, 'direct questioning was greatly disfavored, if not downright prohibited' (Prokić *et al.*, 2009), in general, data from dialect atlases might have been collected in a variety of less natural ways. Uiboaed *et al.* use data from spontaneous spoken corpora. Corpus data, particularly if it is taken from conversational speech, complement traditional sources with data better suited to grammatical analysis and improve researchers' chances of detecting unexpected phenomena. Finally, the article shows that syntactic variation, at least of the sort investigated, unexpectedly appears to be distributed differently than phonological and lexical variation.

The article in this volume concentrates on 'collostructions', grammatical constructions that are marked by specific lexical items, a paradigmatic example of which is the English phrase 'waiting to happen', which tends to take as subject words denoting events that are (nearly) inevitable and often negative (Stefanowitsch and Gries, 2003, p. 220). The wish to use spoken corpus data are no sinecure, however, as the relevant data normally must be extracted via programs, and corpus linguistics is a thriving field with a great deal of methodological discussion. Wiechman (2008) examines forty-seven different statistical measures of association strength to determine which seems to function best in the detection of collostructions. Uiboaed *et al.* use Fisher's exact test to identify collostructions, following Stefanowitsch and Gries (2003). They then use correspondence analysis (Cichocki, 2006) to simultaneously detect sites that are similar in their use of collostructions, and collostructions that tend to appear to the same degree at different sites.

## 2.3 Geographical perspectives

Simon Pickl's paper 'Lexical Meaning and Spatial Distribution. Evidence from Geostatistical Dialectometry' is explicit in breaking away from earlier dialectometry in eschewing the standard dialectometrical step of aggregating the differences of many variables, which step enables a more robust characterization of the relations among sites. There is an aggregation step in his analysis, but this consists in examining, for each concept he investigates, an associated geographical density map. The 'density' refers to how frequently some lexical variants are preferred to others, and the density of a variant at a given site depends not only on how often it was elicited there (if it indeed is a data collection site), but also on how often it was elicited at neighboring sites. It represents therefore a different response to the complexity and variability of linguistic data, and moreover, one that privileges rather than punishes the status of the individual variable.

Pickl's paper continues the innovative approach of the Augsburg-Ulm project that has emphasized image analysis techniques in dialectology (see the references in Pickl's paper). These techniques extract properties from maps depicting the distribution of individual variables, including the 'complexity' of a variable's distribution (how many and how jagged are the isoglosses needed) and its 'homogeneity', or how little the variable varies on a local scale in the area under study. Pickl notes that the two are normally strongly correlated, but prefers to examine both.

From this vantage point, Pickl is in an excellent position to ask the onomasiological question of whether word meanings influence their geographical distributions, in particular their complexity and homogeneity. Pickl resumes the earlier philological tradition of *Wörter und Sachen* (Schuchardt, 1912; Anttila, 1989) in his focus on semantics as influential in geographical diffusion. And as the author himself notes, his work represents a continuation of a dialectometric research line begun by Speelman and Geeraerts (2008), who also identify the quantitative techniques they employ as dialectometric.

Jack Grieve's paper 'A Statistical Comparison of Regional Phonetic and Lexical Variation in American English' breaks new dialectometrical ground in a way that may be surprising. Given the fact that Geographical Information Systems (GIS) have certainly matured to the point where they are straightforward to use and also extremely powerful, and likewise the circumstance that dialectometry has concentrated almost exclusively on geographical influences on variation, it may be surprising that GIS systems are not in constant use in dialectometry. But they are only sporadically deployed in dialectometry, probably because dialectometry focuses on data that normally represent relations between sites, always implicitly involving a comparison, while GIS focuses on measurements (e.g. concentrations of pollutants in ground water, per capita income, or political preferences) at single sites.

Grieve compares lexical and phonetic variation in the continental USA asking about the degree to which the differences coincide (Spruit *et al.*, 2009). He uses on the one hand the (mean) vowel formant data (of nineteen vowels) from The *Atlas of North American English* (Labov *et al.*, 2006) and on the other hand, frequency data on lexical variation taken from earlier work (Grieve *et al.*, 2011). Both sets of data were subjected to spatial autocorrelation analysis to identify variables that display geographical conditioning, and a spatially normalized value (the Getis *Gi* *z*-score) is used in subsequent factor analyses of both the phonetic and the lexical data. This represents a first use of GIS techniques. Now Grieve wishes to compare the two linguistic levels for the degree to which they are similarly distributed across North America, and it would be straightforward if his measurements came from the same sample of cities, but unfortunately, they do not. So Grieve turns a second time to GIS techniques, this time to 'kriging', a technique for estimating values at undocumented sites that lie between the sites in the original sample.

With interpolated (kriged) values in hand, Grieve is able to assay the strong correlation between lexical and phonetic variation ($r = 0.73$). On the way, he also introduces to dialectometry 'theoretical variograms', a measure of the variance in a variable as measured at various distances, an analytical tool which might beneficially see further use. Grieve's work is also one of the technically most sophisticated papers in this collection.

Simon Pröll has collaborated with Simon Pickl in the Ulm-Augsburg project and contributed to this volume the paper 'Detecting Structures in Linguistic Maps—Fuzzy Clustering for Pattern Recognition in Geostatistical Dialectometry'. Pröll shares with Pickl (see the earlier discussion) the point of departure examining the geographical distributions of individual lexical items and analyzing these distributions using image-analysis techniques. He also shares an interest in the potential influence of concepts on geographical diffusion, but he approaches the subject by clustering lexical items based on characteristics of their geographical distributions. In this he makes use of an aggregating step (his fuzzy clustering), but since his goal is to understand the influence of concept characteristics on geographical distributions, he aggregates concepts, not sites, and he suggests that a different sort of clustering be preferred.

As an intermediate step, Pröll introduces 'empirical covariance functions', which plot the covariance of pairs of sites as function of their geographical distance from each other, which he proposes as devices to characterize the degree to which geographically close sites share linguistic traits. They might be compared with the theoretical variograms used in GIS systems (see the earlier discussion on Grieve's paper). Both of the constructs might be used to characterize the influence of geography on linguistic variation in a way that would support cross-linguistic comparison.

An empirical covariance function is then constructed for each lexical item, and the set of variograms is input to a fuzzy version of the *k*-means clustering algorithm (Mackay, 2003, Chap. 20) to produce clusters of concepts (fuzzy clustering allows that a given item belongs to a cluster to a degree, rather than categorically). Pröll does not evaluate his results with respect to an independent classification of semantic concepts, but rather suggests how to interpret the results using ideas from 'prototype theory' (Rosch, 1973; Saeed, 2003).

## 2.4 Social and historical perspectives

Esteve Valls, Martijn Wieling, and John Nerbonne examine the effect of a change in school policy with respect to Catalan in 'Linguistic advergence and divergence in northwestern Catalan: a dialectometric investigation of dialect leveling and border effects'. They combine an analytic perspective taken from sociolinguistics, namely apparent time, in which linguistic change is gauged by comparing speakers of different ages, with the aggregate perspective common in dialectometry. Since in general sociolinguists prefer to study small numbers of changes in isolation (Labov, 2001), the aggregate perspective in combination with a sociolinguistic research question is novel.

The study is prompted in part by the fact that northwestern Catalan is split by a border separating Aragon, in the west, where Catalan is normally spoken but has no official status, and Catalonia and Andorra, in the east, where Catalan is an official language whose standard variety had been taught in schools when the youngest respondents were enrolled in schools (the older speakers had enjoyed compulsory education in Spanish). So the effect of schooling using a standard is reflected in the data.

An even mix of different ages, men and women, urban and rural speakers formed the base of the study, which indeed confirms Auer and Hinskens' (2005) note that, while a standard increases convergence within the borders in which it takes places, it also inevitably increases divergence with respect to areas outside those borders. These effects are seen in the comparison of younger and older speakers. Detecting the influence of standardization is confounded in northwestern Catalan, however, by a process known in the scholarly literature as 'orientalization', a convergence within Catalan toward the more eastern varieties. Selected phenomena suggest that both processes are taking place. Finally, Valls and colleagues examine their data from both cascade models of diffusion, in which more populous settlements are more important, and 'contagion models', in which proximity is solely important.

Maria-Pilar Perea's paper 'Dynamic Cartography with Diachronic Data: Dialectal Stratigraphy' introduces 'dialectal stratigraphy', a mapping technique in which several historical layers, each depicting the dialect maps of a specific period, are joined. Each layer thus depicts the geographical distribution of the realizations of the concept, similarly to the way a single map in a dialect atlas might display the distribution of the different variants of a concept. As such the software aims to support research in historical dialectology and less in historical semantics, where the changes in meaning of a given word are the focus of analysis (including shifts in meaning, narrowings, etc.). The work is well poised to examine how constant (or how changeable) dialectal distributions are over time, and in particular borders between substantially different dialects.

Perea's focus is the historical lexicon of Catalan from the 12th through the 21st century, represented by about sixty concepts, also referred to as 'geo-synonyms'. Crucially these involve lexical alternatives that need not be etymologically related. British 'petrol' versus American 'gasoline, gas' might be members of the same set of geo-synonyms. The data are taken from a long-term project at the University of Barcelona. Perea closes her paper with a list of opportunities open to researchers with good methods in dialectal stratigraphy, e.g. the investigation of whether all varieties tend to undergo lexical divergence.

Simonetta Montemagni, Martijn Wieling, Bob de Jonge, and John Nerbonne employ dialectometric techniques to trace the history of a well-known consonant lenition in Tuscan varieties of Italian, so-called *Gorgia Toscana* 'Tuscan throat', in which e.g. $k > x > h > Ø$ (deletion). In their paper, entitled 'Synchronic Patterns of Tuscan Phonetic Variation and Diachronic Change: Evidence from a Dialectometric Study', the authors invoke Bàrtoli's (1925) neolinguistic principles on detecting diachronic change in synchronic patterns of diffusion. In a nutshell Bàrtoli predicted that one should find older linguistic forms in more isolated areas, in peripheral versus central areas (where these may be contrasted), and in larger rather than smaller areas.

The authors also make use of BiSGP (Wieling and Nerbonne, 2011) (see the earlier discussion), in this case using data restricted to sound correspondences (automatically extracted using a technique developed in dialectometry, see Wieling *et al.*, 2012) involving only those segments (voiceless stops) that are potential candidates for the process. By exploiting the 'two-way' BiSGP clustering, the authors sought a link between the areal distribution of correspondences and their historical age. The results indeed confirm that the older forms are found peripherally, in areas that the change, radiating from Florence, failed to reach. Several intermediate degrees of change may also be distinguished.

The authors also examine segment correspondences in context, e.g. VkV:VhV, in an attempt to trace the gradual generalization of the phonetic process to more and more environments, and this too, succeeds to some extent. This might also be noted as a partial answer to critics of the use of edit distance as a measure of pronunciation difference (Heggarty *et al.*, 2005) as it indicates how the measure may be applied in a contextually sensitive manner. This has been noted in principle earlier (Heeringa *et al.*, 2006), but the use of *n*-grams has not been popular, perhaps due to the need for larger amounts of data.

## 2.5 An excursion into complexity theory

William A. Kretzschmar, Jr, Brendan A. Kretzschmar, and Irene M. Brockman present ideas from complexity theory in their contribution 'Scaled Measurement of Geographic and Social Speech Data'. Proceeding from the observation that the Zipf-like curves of frequency that we know from lexical frequencies (Baayen, 2001) is not only found in the distribution of dialectal variants but also at smaller scales, e.g. when one restricts one's attention to a subset of respondents. The authors, following Kretzschmar (2009), refer to the Zipf-like curves as 'A-curves', and brashly adduce evidence suggesting that even the acoustic variants of single vowels are distributed this way (note that such distributions are often sketched as Gaussian ellipses).

In a far-ranging essay highlighting links to economics, physics, and other areas of inquiry, the authors introduce the 'Gini coefficient' to linguistics, a measure of the inequality of a distribution. Perfectly uniform distributions have Gini coefficients of zero, and very uneven distributions have coefficients approaching one (1.0). Normal distributions have coefficients that may rise with their standard deviations but are much smaller than the coefficients associated with the distribution of dialectal variants, where typically one variant is extremely dominant. They then try categorizing data to various degrees of discrimination, showing that the Gini coefficient is affected: at very rough degrees of discrimination, the unevenness in the underlying distribution is thoroughly hidden. Similarly, using too few respondents fail to reflect how unevenly the population data are distributed. Samples become unrepresentative.

In conclusion, the authors suggest that their explorations have consequences for how finely one should categorize linguistic data and for the sample size of respondents one should try to use.

## 3 Conclusion

Dialectometry continues to focus on the measure of dialectal differences, but it now includes a large variety of techniques in addition to the fairly simple aggregations of differences that served it well for so long. Regrettably, we are not able to introduce papers in this collection that explore whether dialect distributions resemble other cultural distributions (Goebl, 2005; Manni *et al.*, 2006; Falck *et al.*, 2012) for we find in such studies yet another major virtue of the dialectometric approach, namely that dialectometry provides numerical characterizations of dialectal affinity that may be compared with biological and economic measures, e.g. how families are distributed geographically (Goebl, 2005; Manni *et al.*, 2006), or how likely contemporary migration is (Falck *et al.*, 2012). We hope that dialectometry may in this way contribute to a more general understanding of culture.

## References

**Anttila, R.** (1989). *Historical and Comparative Linguistics*. Amsterdam: Benjamins.

**Auer, P. and Hinskens, F.** (1996). The Convergence and Divergence of Dialects in Europe. New and not so New Developments in an Old Area. In Ammon, U., Mattheier, K. L., and Nelde, P. (eds), *Konvergenz und Divergenz von Dialekten in Europa. Convergence and divergence of dialects in Europe. Convergence et divergence des dialectes en Europe*. Tübingen: Niemeyer, pp. 1–30.

**Baayen, R. H.** (2001). *Word Frequency Distributions*. Berlin: Springer.

**Bàrtoli, M.** (1925). *Introduzione alla Neolinguistica*. Genève: Olschki.

**Cichocki, W.** (2006). Geographic variation in Acadian French/r/: what can correspondence analysis contribute toward explanation?. *Literary and Linguistic Computing*, **21(4)**: 529–41.

**Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P.** (2010).A Latent Variable Model for Geographic Lexical Variation. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Shroudsburg, PA Association for Computational Linguistics, pp. 1277–87.

**Embleton, S.** (1987). Multidimensional Scaling as a Dialectometrical Technique. In Babitch, R. M. (ed.), *Papers from the Eleventh Annual Meeting of the Atlantic Provinces Linguistic Association*. New Brunswick, Canada: Atlantic Provinces Linguistic Association, pp. 33–49.

**Falck, O., Heblich, S., Lameli, A., and Südekum, J.** (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, **72**(2–3): 225–39.

**Goebl, H.** (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften (Denkschriften, Bd. 157).

**Goebl, H.** (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer (Beihefte zur Zeitschrift für romanische Philologie, Bd. 191).

**Goebl, H.** (1997). Some Dendrographic Classifications of the Data of CLAE 1 and CLAE 2. In Viereck, W. and Ramisch, H. (eds), *The Computer Developed Linguistic Atlas of England 1*. Tübingen: Max Niemeyer, pp. 23–32.

**Gooskens, C.** (2012). Methods for Measuring Intelligibility of Closely Related Language Varieties. In Bayley, R., Cameron, R., and Lucas, C. (eds), *Handbook of Sociolinguistics*. Oxford: Oxford University Press [in press].

**Gooskens, C. and Heeringa, W.** (2004). Perceptual evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, **16**(3): 189–207.

**Gooskens, C., Beijering, K., and Heeringa, W.** (2008). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, **2**(1–2): 63–81.

**Grieve, J., Speelman, D., and Geeraerts, D.** (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, **23**: 193–221.

**Haag, K.** (1898). *Die Mundarten des oberen Neckar- und Donaulandes* (Schulprogramm). Reutlingen.

**Haag, K.** (1901). Über mundartgeographie. *Zeitschrift des Freiburger Geschichtsvereins*, **17**: 228–47.

**Haimerl, E.** (1998). A Database Application for the Generation of Phonetic Atlas Maps. In Nerbonne, J. (ed.), *Linguistic Databases*. Stanford: CSLI Press, pp. 103–16.

**Haimerl, E.** (2006). Database design and technical solutions for the management, calculation and visualization of dialect mass data. *Literary and Linguistic Computing*, **21**(4): 436–44. (Special issue, *Progress in Dialectometry: Toward Explanation*, edited by Nerbonne, J. and Kretzschmar, W., Jr).

**Heeringa, W., Kleiweg, P., Gooskens, C., and Nerbonne, J.** (2006). Evaluation of String Distance Algorithms for Dialectology. In Nerbonne, J. and Hinrichs, E. (eds), *Linguistic Distances*. Sydney: Workshop at COLING-ACL, pp. 51–62.

**Heeringa, W., Nerbonne, J., and Osenova, P.** (2010). Detecting Contact Effects in Pronunciation. In Norde, M., de Jonge, B., and Hasselblatt, C. (eds), *Language Contact. New Perspectives*. Amsterdam: Benjamins, pp. 131–53.

**Heggarty, P., McMahon, A., and McMahon, R.** (2005). From Phonetic Similarity to Dialect Classification: A Principled Approach. In Delbecque, N., van der Auwera, J., and Geeraerts, D. (eds), *Perspectives on Variation*. Amsterdam: Mouton de Gruyter, pp. 43–91.

**Kessler, B.** (1995). *Computational Dialectology in Irish Gaelic. In Seventh Conference of the European Chapter of the Association for Computational Linguistics*. Dublin: Association for Computational Linguistics, pp. 60–67.

**Keune, K., Ernestus, M., van Hout, R., and Baayen, R. H.** (2005). Social, geographical, and register variation in Dutch: from written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory*, **1**: 183–223.

**Kondrak, G.** (2003). Phonetic alignment and similarity. *Computers and the Humanities*, **37**(3): 273–91. (Special issue, *Computational Methods in Dialectometry*, edited by Nerbonne, J. and Kretzschmar, W.A., Jr).

**Kondrak, G. and Sherif, T.** (2006). Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In Nerbonne, J. and Hinrichs, E. (eds), *Linguistic Distances*. Sydney: Workshop at COLING-ACL, pp. 43–50.

**Kretzschmar, W. A., Jr.** (2009). *The Linguistics of Speech. Cambridge*: Cambridge University Press.

**Kürschner, S., Gooskens, C., and van Bezooijen, R.** (2008). Linguistic Determinants of the intelligibility of Swedish words among danes. *International Journal of Humanities and Arts Computing*, **2**(1–2): 83–100.

**Labov, W.** (2001). *Principles of Linguistic Change (II): Social Factors*. Oxford: Blackwell.

**Labov, W., Ash, S., and Boberg, C.** (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

**List, J. M.** (2012). *Multiple Sequence Alignment in Historical Linguistics. A Sound Class Based Approach. Proceedings of ConSOLE XIX*, pp. 241–260. http://www.sole.leidenuniv.nl (26 November 2012, date last accessed).

**MacKay, D.** (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

**Manni, F., Heeringa, W., and Nerbonne, J.** (2006). To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands. *Literary and Linguistic Computing*, **21**(4): 507–27.

**Nerbonne, J.** (2009). Data-driven dialectology. *Language and Linguistics Compass*, **3**(1): 175–98.

**Nerbonne, J.** (2010). Mapping Aggregate Variation. In Lameli, A., Kehrein, R., and Rabanus, S. (eds),.

*Language and Space. International Handbook of Linguistic Variation*. Vol. 2, *Langauge Mapping*. Berlin: Mouton De Gruyter, Chap. 24, pp. 476-495, maps 2401-2406. (Series *Handbooks of Linguistics and Communication Science* 30.2.).

**Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., and Leinonen, T.** (2011). Gabmap—a web application for dialectology. *Dialectologia.* Special Issue II, 65–89. (Special issue *Production, Perception and Attitude*, edited by Nerbonne, J., Grondelaers, S., Speelman, D., and Perea, M.-P.).

**Nerbonne, J., Gooskens, C., Kürschner, S., and van Bezooijen, R.** (eds), (2008). Language variation. *International Journal of Humanities and Arts Computing* (special issue), 2(1–2).

**Nerbonne, J., Heeringa, W., and Kleiweg, P.** (1999). Edit Distance and Dialect Proximity. In Sankoff, D. and Kruskal, J. (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI Press, pp. v–xv.

**Nerbonne, J. and Kleiweg, P.** (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, **37**: 339–57.

**Nerbonne, J. and Kretzschmar, W. A., Jr** (eds), (2003). Computational methods in dialectometry. *Computers and the Humanities* (special issue), 37(3).

**Nerbonne, J. and Kretzschmar, W. A., Jr** (eds), (2006). Progress in Dialectometry: Toward Explanation. *Literary and Linguistic Computing.* (special issue), 21(4).

**Nerbonne, J., van Ommen, S., Wieling, M., and Gooskens, C.** (2013). Measuring Socially Motivated Pronunciation Differences. In Borin, L. and Saxena, A. (eds), *Approaches to Measuring Linguistic Differences*. Berlin: Mouton De Gruyter.

**Prokić, J., Nerbonne, J., Zhobov, V. *et al.*** (2009a). The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, **3**(3): 269–98.

**Prokić, J., Wieling, M., and Nerbonne, J.** (2009b). Multiple Sequence Alignments in Linguistics. In Borin, L. and Lendvai, P. (eds), *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. (LaTeCH—SHELT&R 2009). Stroudsburg, PA: Association for Computational Linguistics, pp. 18–25.

**Rosch, E.** (1973). Natural cateogies. *Cognitive Psychology*, **4**(3): 328–50.

**Saeed, J. I.** (2003). *Semantics*. 2nd edn. Oxford: Blackwell.

**Sanders, N. C. and Chin, S. B.** (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, **43**: 96–114.

**Scapoly, C., Goebl, H., Sobota, S., Mamolini, E., Rodriguez-Larralde, A., and Barrai, I.** (2005). Surnames and dialects in France. Population structure and cultural evolution. *Journal of Theoretical Biology*, **237**: 75–86.

**Scherrer, Y.** (2012). *Recovering Dialect Geography from an Unaligned Comparable Corpus. Proceedings of the EACL Workshop on Visualization of Language Patterns and Uncovering Language History from Multilingual Resources*. Avignon: Association for Computational Linguistics, pp. 65–71.

**Scherrer, Y. and Rambow, O.** (2010). *Word-Based Dialect Identification with Georeferenced Rules. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Shroudsburg, PA: Association for Computational Linguistics, pp. 1151–61.

**Schuchardt, H.** (1912). Wörter und Sachen. *Anthropos*, **7**(4): 827–39.

**Séguy, J.** (1973). La dialectométrie dans *l'Atlas linguistique de la Gascogne*. *Revue de linguistique romane*, **37**: 1–24.

**Speelman, D. and Geeraerts, D.** (2008). The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing*, **2**(1–2): 221–42.

**Spruit, M. R.** (2008). *Quantitative Perspectives on Syntactic Variation in Dutch Dialects.* Ph.D. thesis, University of Amsterdam.

**Spruit, M. S., Heeringa, W., and Nerbonne, J.** (2009). Associations among linguistic levels. *Lingua*, **119**: 1624–42.

**Stefanowitsch, A. and Gries, S. T.** (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, **8**(2): 209–43.

**Streck, T. and Auer, P.** (2012). Das Raumbildende Signal in der Spontansprache: Dialektometrische Untersuchungen zum Alemannischen in Deutschland. *Zeitschrift für Dialektologie und Linguistik*.

**Szmrecsanyi, B.** (2012). *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. (Series: Studies in English Language). Cambridge: Cambridge University Press.

**Szmrecsanyi, B. and Kortmann, B.** (2009). The morphosyntax of varieties of english worldwide: a quantitative perspective. *Lingua*, **119**(11): 1643–63. (Special issue 'The Forests behind the Trees', edited by Nerbonne, J. and Manni, F.).

**Wichmann, S. and Holman, E. W.** (2009). Population size and rates of language change. *Human Biology*, **81**(3): 259–74.

**Wiechmann, D.** (2008). On the computation of collostruction strength: testing measures of association as expressions of lexical bias. *Corpus Linguistics & Linguistic Theory*, **4**(2): 253–90.

**Wieling, M., Margaretha, E., and Nerbonne, J.** (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, **40**(2): 307–14.

**Wieling, M. and Nerbonne, J.** (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, **25**: 700–15.

**Wieling, M., Baayen, R. H., and Nerbonne, J.** (2011). Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS One*, **6**(9): e23613.

**Wiersma, W., Nerbonne, J., and Lauttamus, T.** (2011). Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, **26**(1): 107–24.

## Notes

[1] See http://eudia.ehu.es/diatech.

[2] See http://www.dialectometry.com/.

[3] See http://www.gabmap.nl.