ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Advances in Dialectometry

Martijn Wieling[1,2] and John Nerbonne[1,3]

[1]Department of Humanities Computing, University of Groningen, 9700 AS Groningen, Netherlands; email: wieling@gmail.com, j.nerbonne@rug.nl

[2]Department of Quantitative Linguistics, University of Tübingen, 72074 Tübingen, Germany

[3]Freiburg Institute for Advanced Studies, University of Freiburg, 79104 Freiburg, Germany

Annu. Rev. Linguist. 2015. 1:243–64

First published online as a Review in Advance on July 28, 2014

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

This article's doi: 10.1146/annurev-linguist-030514-124930

Copyright © 2015 by Annual Reviews. All rights reserved

Keywords

language variation, aggregate (analysis of) variation, dialect geography, sociolinguistics, language change

Abstract

Dialectometry applies computational and statistical analyses within dialectology, making work more easily replicable and understandable. This survey article first reviews the field briefly in order to focus on developments in the past five years. Dialectometry no longer focuses exclusively on aggregate analyses, but rather deploys various techniques to identify representative and distinctive features with respect to areal classifications. Analyses proceeding explicitly from geostatistical techniques have just begun. The exclusive focus on geography as explanation for variation has given way to analyses combining geographical, linguistic, and social factors underlying language variation. Dialectometry has likewise ventured into diachronic studies and has also contributed theoretically to comparative dialectology and the study of dialect diffusion. Although the bulk of research involves lexis and phonology, morphosyntax is receiving increasing attention. Finally, new data sources and new (online) analytical software are expanding dialectometry's remit and its accessibility.

1. INTRODUCTION

Edit distance (or Levenshtein distance): minimum number of insertions, deletions, and substitutions to transform one (phonetic) string into the other

Dialectometry studies dialects using exact methods, especially computational and statistical approaches. In this article, we review exciting developments, roughly focusing on the past five years, when dialectometry has gained substantially both in the range of techniques under study and in the number of practitioners.

The great tradition of European dialect geography produced innumerable detailed maps depicting the geographic distribution of variation, especially in word choice, pronunciation, and morphology. Researchers naturally sought to identify the deeper geographic and social structures that might be assumed to underlie many details and that might be examined as potentially explanatory. But as Bloomfield's (1933, p. 340ff) classic discussion of this work noted, the maps of individual features often simply did not coincide, leading him to conclude that "in this respect [. . .] dialect geography proved to be disappointing." The problem usually revolved around how one should distinguish dialect areas, but modern dialectology recognizes that geographic distributions may involve continua or even scattered settlements.

Jean Séguy (1971, 1973) is credited with taking the liberating step of examining not individual features, but rather large aggregates of features, effectively asking how often two sites differ with respect to a given set of features (such as lexicalizations, but also the pronunciation of selected sounds, or the realization of a given morpheme). It is historically noteworthy that Haag (1898) had suggested something very similar, namely counting the isoglosses that separated sites to assay the strength of a putative border separating them, as noted by Bloomfield (1933) in the chapter cited above. Séguy not only took this step but presciently applied it to one of the foundational questions in dialect geography, the relation between aggregate linguistic differences and geographic distance (Séguy 1971).

In a programmatic article, Nerbonne (2009, p. 179) summed up the motivation for dialectometry's attention to aggregates rather than individual features, arguing that the common practice of abstracting away from many details of phonetic variation is an implicit sort of aggregation that all variationists have accepted, and further noting that individual features are inevitably noisy (interpreting Bloomfield's point above in this way). He also observed that the sheer number of available features makes it likely that a researcher focused on individual features can find some feature or other that coincides with a putative social or geographical influence, exposing the researcher to the danger of "cherry picking"—working with features that are selected (perhaps innocently) to confirm his or her hypotheses. Nerbonne (2009, pp. 190–91) finally notes that moving the analysis from the (categorical) level of individual features to the (numerical) level of aggregates enables language variationists to study general relations such as the law-like relation between linguistic differences and geographic distance demonstrated by Séguy (1971).

Séguy was sadly lost to dialectometry through an accident, but his work was continued and deepened enormously by Hans Goebl, who may be regarded with Séguy as the prime contributor of this direction in dialectology (Goebl 1982a,b). Goebl has focused mainly on Romance languages, and especially notable among his many contributions to dialectometry are his use of Thiessen polygons as a means of dividing maps into areas around data-collection sites, his experimentation with (inverse) frequency weightings in calculating aggregate differences, and his introduction of a range of descriptive statistics as well as cluster analysis to dialectometry. Goebl (1982b) also pleaded for computational procedures in dialectology early on.

An important innovation for dialectometry was the introduction of edit distance [or Levenshtein distance (Levenshtein 1965)], first used in dialectometry by Kessler (1995). Heeringa's (2004) influential dissertation revolved around the use of edit distance in dialectometry and has enabled the analysis of large databases of material collected during the heyday of dialect atlases,

obviating the usual need to manually characterize and classify data with respect to specific phonetic linguistic variables (Goebl refers to this step as *Taxierung* 'appraisal'). Heeringa also included a great deal more data in his analyses than had been used before. This point can be appreciated if one recalls that the phonetic transcription of words had normally been used to study the pronunciation of a single segment, or sometimes a single segment with respect to a specific context. Heeringa (2004) used edit distance to analyze the phonetic transcriptions of entire words and phrases, typically consisting of four to five segments rather than a single segment. This step shelters dialectometry further from the bias introduced by focusing only on a few preselected features.

Multidimensional scaling (MDS): a dimension-reduction technique for distance matrices

Whereas Black (1976) introduced multidimensional scaling (MDS) to linguistics, Embleton (1993) applied the technique specifically to dialectometry (see Embleton et al. 2013 for more recent work on alternative MDS visualizations). MDS takes a site × site distance table as input and tries to assign the sites in the table to coordinates in a small-dimensional space, typically consisting of two or three dimensions. Nerbonne et al. (1999) mapped MDS coordinates to color values for the first time, providing visual correlates in response to the frequent critique found in dialect atlases and treatises that the division of the language area into different dialect areas did little justice to the gradual nature of dialect boundaries. **Figure 1** shows an example of one of these MDS maps, visualizing Dutch phonetic variation, together with a legend providing examples of words and how they are pronounced in their "fuzzy" areas. Heeringa's (2004) dissertation used this form of presentation as well. Heeringa identified "typical" word pronunciations by selecting words whose distances correlated highly with the (distances on the basis of the) dimensions proposed in MDS, effectively the intensity of the colors shown in **Figure 1**.

2. RELATED WORK

An exciting aspect of recent work in dialectometry is the degree to which it communicates with nondialectometric work. We review some of that related work here.

Given the close relation between dialectology and sociolinguistics, one would expect to see many applications of dialectometric techniques in sociolinguistics. Indeed, Ruette et al. (2014) refer to their work as sociolectometry, and there have been other studies that include not only typical geographic predictors but also social variables. We review some of these below in Section 3.3, and we include a focused section on analyses that include both social and geographical influences as independent variables (Section 3.2). An example of a mixed social dialectology analysis can be found in Nerbonne et al. (2013), which examined whether regional radio speakers are using regiolects in Auer & Hinskens's (1996) sense. We emphasize here, however, that dialectometry measures differences in language varieties with no primary commitment to geographic interpretation. Typically, the geographic interpretation has been the impetus behind the measurement, but it plays no role in the measurement itself. It is therefore safe to say either that dialectometric techniques make sense both for analyses of dialect geography and for those of social variation or that they make sense for neither. Standard sociolinguistics work certainly deals with the noise in language variation data effectively (through its application of logistic regression), but there is a second danger in variationist linguistics, as noted in Section 1, namely that researchers pick convenient variables rather than analyzing varieties in more general ways. We therefore suggest that dialectometric techniques might be deployed more often in sociolinguistic studies than they are now.

Although the use of an edit-distance measure in historical linguistics seems to have developed independently (Wichmann et al. 2010), there is now an active exchange of ideas on the best refinements supporting historical inference (List 2012, Wieling et al. 2012, Jäger 2014, Rama & Borin 2015). All of these papers explore enhancements of edit distance in which the cost of replacing a sound with another is systematically lower when the two sounds are phonetically

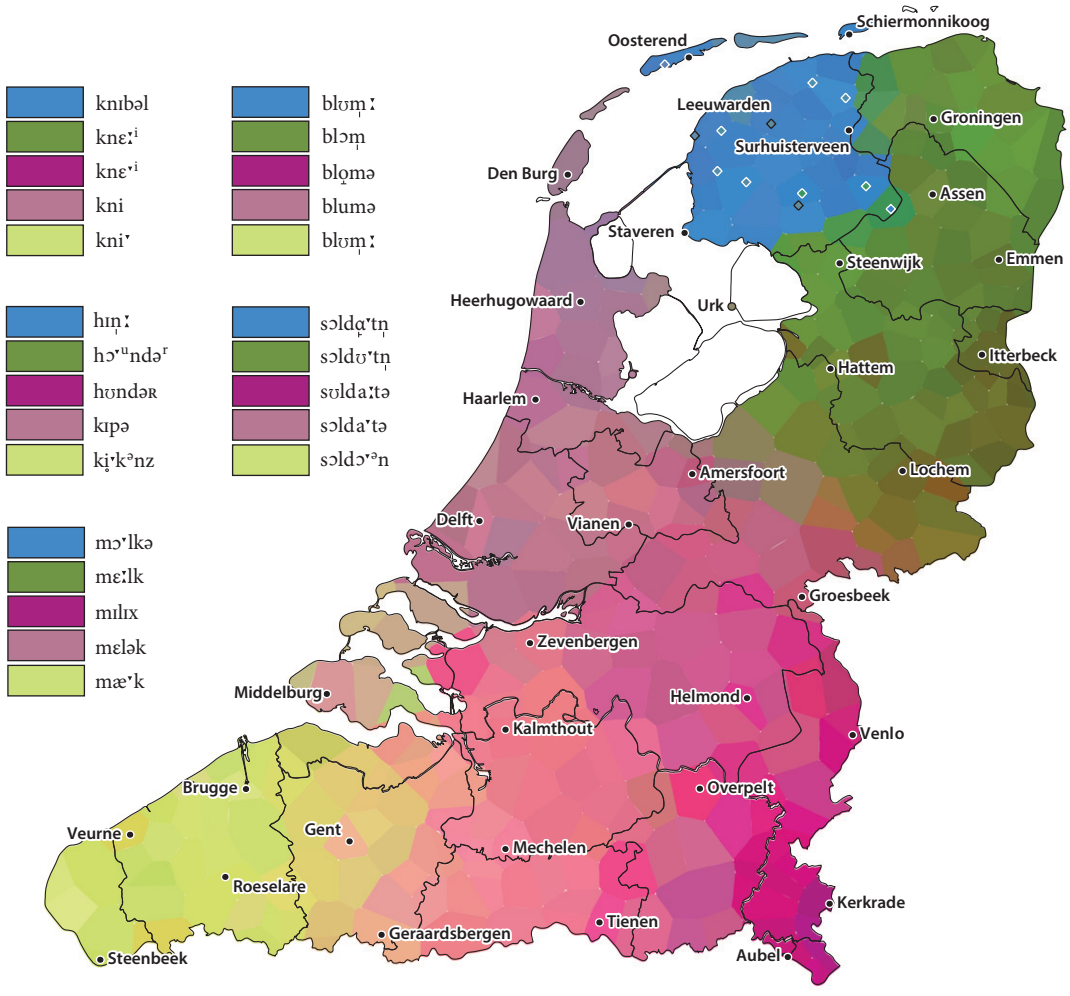

Figure 1

The three most important multidimensional scaling dimensions (together accounting for more than 85% of the variation in the location × location distance table) have been mapped to red, green, and blue, thereby providing a comprehensive visualization of Dutch phonetic variation. The five legends provide some typical pronunciations in the areas with the purest colors. Note that areas are genuine, even though borders are gradual.

similar. Whatever version is finally shown to be best, it will have applications in the automatic detection of cognates and in estimating whether languages are related at all. Historical linguists are, of course, wary of estimating relatedness on the basis of similarity, but as List (2012, pp. 247–48) notes, one may distinguish synchronic similarity from diachronic similarity, which always involves "similarity with respect to regular sound correspondences." At shallow time depths, similarity is likely to reflect genealogical relations, meaning that edit distance is often sufficient to identify relationships, and Snoek (2013) uses the online dialectometry application Gabmap (Nerbonne et al. 2011; also see Section 3.7, below) to investigate Athapaskan, whose time depth is estimated to be 1,300–2,000 years (Hale & Harris 1979, p. 175). Prokić & Moran (2013),

however, criticize efforts to employ edit distance, especially at greater time depths, that fail to take note of regular sound correspondences. But we observe that the algorithm that calculates the edit distance does so by aligning corresponding sounds (Kruskal 1999) and that there is substantial interest in using the alignments to identify sound correspondences (Prokić et al. 2009, List 2012, List & Moran 2013), a cornerstone of inference in historical linguistics. We hasten to add here that the pairwise alignment of word pronunciations is insufficient for the detection of regular sound correspondences, which is why these works are also concerned with multialignment, a topic that Prokić (2010) explores within dialectology.

Alignment (of phonetic transcriptions): procedure aimed at arranging the sounds in pairs of pronunciations (of cognates) so that corresponding sounds are identified

Pointwise mutual information (PMI): a measure of association strength

The degree to which the dialects of a given language (or closely related languages) are mutually intelligible has been examined in dialectology and dialectometry a good deal (Chambers & Trudgill 1998, p. 3ff), and Gooskens (2013) summarizes a great deal of work that has essentially applied dialectometric techniques to gauge the degree to which related varieties—but not dialectal varieties—are mutually comprehensible. Gooskens's work uses categorical procedures (percent of overlap or difference) to gauge lexical differences and edit-distance-based procedures to gauge pronunciation differences, and she often compares differences at various linguistic levels to determine which are more important (Gooskens et al. 2009). Focusing on language contact more than comprehensibility, Lauttamus et al. (2007) and Wiersma et al. (2011) use a syntactic distance measure to assay the degree of divergence in the English of Finnish emigrants to Australia (with respect to local Australian English). Their techniques were robust enough to be applied to corpora of spontaneous speech.

Wieling et al. (2014a) analyze foreign accents using a refinement of the edit-distance measure developed in dialectometry [i.e., using sensitive sound segment distances obtained via pointwise mutual information (PMI); Wieling et al. 2009, 2012] and validate the measure using native speakers' judgments of native-like versus foreign sounding, thereby obtaining a strong correlation ($r = 0.8$) for a logarithmically transformed version of the refined edit-distance measure. Wieling et al. (2014c) compare this measure with a completely different computational measure (i.e., one based on the theory of human discriminative learning) and show that the two methods achieve comparable performance. In effect, this observation suggests that when comparing pronunciations at an aggregate level by averaging the pronunciation distances of many words, the specific measure used might not be so important.

Finally, there are papers using dialectometric techniques for a variety of other topics. Sanders & Chin (2009) use an edit-distance measure to gauge the deviance of the speech of cochlear implant bearers, and Kondrak (2013) uses another string similarity measure, longest common subsequence ratio, to explore orthographically similar words in different languages with an eye to distinguishing cognates and recognizing single-word translations.

3. RECENT ADVANCES IN DIALECTOMETRY

The past five years have seen a great number of improvements in the field of dialectometry. From a global perspective, these improvements mostly center on identifying the most important (diagnostic) individual linguistic items underlying aggregate dialect variation (Section 3.1), the realization that lexical and social factors may condition geographical variation (Section 3.2), new methods to assess linguistic change in dialects (Section 3.3), dialectological theory (Section 3.4), more attention to dialect grammar and morphosyntax (Section 3.5), the use of new data sources besides traditional dialect atlas data (Section 3.6), and the creation of new (online) applications enabling dialect researchers to more easily use dialectometric tools (Section 3.7).

In the following subsections, we discuss each of these seven areas in turn and highlight some of the most interesting studies conducted in each of them. Although we generally focus on research

from the past five years, in some cases we highlight (slightly) older studies to offer a more comprehensive overview.

Principal components analysis (PCA): a dimension-reduction technique for value matrices

Factor analysis: a dimension-reduction technique for value matrices related to PCA

Clustering: procedure that seeks groups in data sets

Geostatistics: techniques developed for spatial analyses

3.1. Identifying the Linguistic Basis of Aggregate Dialect Variation

Dialectometry has been rightly criticized for the lack of attention to the individual linguistic items underlying aggregate geographical patterns (Schneider 1988, Woolhiser 2005). Especially in recent years, however, various researchers have taken this criticism to heart and have investigated the linguistic basis of aggregate dialect variation. Many new approaches have been developed, in some cases building on earlier approaches that used the idea that principal components analysis (PCA) (e.g., Shackleton 2005) or factor analysis (e.g., Nerbonne 2006) can be used to detect linguistic items showing similar geographical patterns.

Pröll et al. (2014) indeed use factor analysis to detect linguistic patterns of variation. To visualize the resulting factors on a single map, however, they assign each location a color (corresponding to the factor with the strongest loading) and an intensity (corresponding to the factor loading). This visualization approach reveals distinct groupings, but Pröll et al. emphasize that it is not the borders between the groups, but rather the dark centers in each area that should be most informative.

Pröll (2013, 2014) elsewhere proposes an alternative approach. He starts from dialectometric intensity estimation (Rumpf et al. 2009), which visualizes the distribution of a single linguistic variable (i.e., a single-dialect atlas map) on a so-called area class map. Pröll then applies fuzzy clustering to (measures derived from) these area class maps to determine the linguistic basis of geographical patterns. Dialectometric intensity estimation is based on the idea that dialect atlas data also contain some random fluctuations due to having selected a specific speaker as being representative for the location. To better detect which variant is likely used in a location, the method takes into account the variants used in the locations within a certain radius around it [Pröll (2013, 2014) uses geographical distances as the basis for this radius, whereas Pickl et al. (2014) argue that using linguistic distances might be better]. The proposed method then uses these surrounding locations to assign a probability representing how likely it is that the variant is used in the location. Locations around the target location using the same variant increase this probability, whereas when all surrounding locations use a different variant, the probability might be so far reduced that another variant is actually assumed to be used in that location. In effect, this method smooths the distribution of variants. An area class map is used to visualize these probabilities using colors and intensities to denote the variant distribution and associated probabilities. In addition, numerical measures (i.e., complexity and homogeneity) can be derived from these area class maps, and these values appear to vary depending on the semantic category of the items (Pickl 2013).

Grieve et al. (2011) propose multivariate spatial analyses that have some similarities to Pröll's (2013, 2014) approach. As Grieve et al.'s approach builds on well-established spatial and geostatistical techniques (Chun & Griffith 2013), it is valuable for placing dialectometry within the larger field of geostatistics [also see Lameli (2013, chapter 5)]. Grieve et al. (2011) aim to obtain a quantitative counterpart to the traditional analysis of regional linguistic variation, which consists of identifying isoglosses, bundles of isoglosses, and finally dialect regions. Grieve et al. first use local spatial autocorrelation to identify patterns of regional linguistic variation for individual linguistic variables (i.e., lexical alternation variables). Similar to Rumpf et al.'s (2009) approach, this step involves taking into account the variant use in surrounding locations and therefore smoothing the geographical pattern. The most important difference between the two approaches is that Grieve et al. include only those variables that show a significant pattern of regional linguistic variation (as assessed by calculating a global autocorrelation score for each linguistic variable).

The second step taken by Grieve et al. is to apply factor analysis to the results of the local spatial autocorrelation analysis to identify groups of linguistic variables showing a similar regional pattern (i.e., determining the linguistic basis). The third and final step is to apply hierarchical cluster analysis to the factor scores per location to determine the resulting geographical clusters. In a comparison with traditional dialectometric analysis, Grieve (2014) finds that multivariate spatial analysis is able to identify more detailed regional patterns due to the removal of nonregional noise in the data set that was confounding the traditional aggregate dialectometric analysis. Not only has multivariate spatial analysis been applied to lexical alternation data, it has also been successfully used to distinguish the contraction rate in written standard American English (Grieve 2011). In addition, it has been applied to American English grammatical variation (Grieve 2012), American English vowel variation (Grieve et al. 2013), and Dutch phonology and morphology (Tamminga 2013). In a later study, Grieve (2013) compares the results of his group's phonetic analysis (Grieve et al. 2013) and lexical analysis (Grieve et al. 2011). Because the two data sets did not overlap in their locations, Grieve uses a geostatistical technique, ordinary kriging, to estimate the values of a variable at unobserved locations, effectively interpolating between data-collection sites. This technique uses the relationship between the geographical distance and the amount of spatial variability (i.e., a variogram) for a certain variable to determine the importance of the surrounding locations in estimating the value at a specific location. By calculating the values for both data sets at identical coordinates, Grieve (2013) compares the results of his group's phonetic analysis and lexical analysis for the United States and shows that these correlate strongly at a level of about $r = 0.7$.

In an attempt to improve on earlier studies that determined the linguistic basis of an aggregate geographical pattern in a post hoc fashion (e.g., Heeringa 2004, pp. 268–70; also discussed above), Wieling & Nerbonne (2009) tried to assess the varietal relatedness and the linguistic basis simultaneously. They used a technique from information retrieval, bipartite spectral graph partitioning (Dhillon 2001), to cluster Dutch dialect varieties together with their characteristic sound correspondences (using as reference a dialect variety whose pronunciations were close to standard Dutch). In subsequent studies, they adapted this graph-based method to allow for hierarchical clustering (Wieling & Nerbonne 2010), and they developed a measure to rank the most important sound correspondences in each cluster (Wieling & Nerbonne 2011a). Wieling & Nerbonne (2009) originally used this technique to investigate Dutch dialects, but the method was subsequently applied to Tuscan phonetic variation (Montemagni et al. 2013), English phonetic variation (Wieling et al. 2013), and contemporary English lexical variation (Wieling et al. 2014d).

The disadvantage of the clustering approach by Wieling and colleagues is that it works only for a two-dimensional matrix (i.e., location $\times$ linguistic features) and therefore needs a reference location for comparison when focusing on sound correspondences. Prokić & Van de Cruys (2010) use parallel factor analysis to analyze a three-dimensional matrix (i.e., location $\times$ location $\times$ sound correspondences) that does not require a reference location. The result of their analysis is a set of factors containing sound correspondences exhibiting relatively similar patterns. The associated geographical patterns can be identified by visualizing the distribution of the top-ranking sound correspondences in each factor.

Prokić et al. (2012) elaborate on the approach that Wieling & Nerbonne (2011a) used to rank the most important sound correspondences, generalizing it to identify characteristic items based on numerical dialect differences, rather than discrete features. Prokić & Nerbonne (2013) show that, in addition to being able to detect characteristic words (Prokić et al. 2012), their generalization can identify characteristic phones via phone-based dialect distances.

Recently, Ruette & Speelman (2014) developed another promising alternative, individual differences scaling. Their approach is an extension of MDS, but rather than using an aggregate

location × location distance table, it uses a series of these tables (e.g., a location × location distance table for each linguistic feature or group of features). Just as for MDS, the output is a dimension-reduced representation of the aggregate distance matrix. In addition, however, it returns the importance of every MDS dimension for each of the individual distance tables (i.e., linguistic features). Ruette & Speelman (2014) use their approach to illustrate that lexical convergence in Netherlandic and Belgian Dutch differs across semantic fields.

Gini coefficient: measure of how nonuniform a statistical distribution is

Although attempting to determine the linguistic basis of geographical variation is certainly insightful, we should keep in mind that variation is never purely categorical. Whereas in one dialect one variant may be dominant, this does not mean that it is the only form being used. Kretzschmar (2012) argues that speech is a complex dynamical system (also see Kretzschmar 2010) and that the distribution of variants of a linguistic variable always shows a nonlinear distribution (i.e., an A-curve, or Zipf curve); only a few variants are highly frequent, and there are many infrequent variants. Recently, Kretzschmar et al. (2013) introduced the Gini coefficient (from the field of economics) as a way to identify the shape of the A-curve. The Gini coefficient can be used to assess whether the number of allowed categories for a specific linguistic variable is not too restrictive to identify the omnipresent nonlinear distribution.

In sum, dialectometrists now have the choice of a large number of different approaches to identify the linguistic basis of aggregate dialect variation. Hopefully, focusing only on aggregate geographical variation without any regard to the linguistic basis will soon become a relic of the past.

3.2. Determinants of Aggregate Dialect Variation

Whereas an aggregate approach averaging over a large number of linguistic variables is arguably more objective than (nonrandomly) selecting only a small number, this does not mean that no other factors need to be considered in assessing variation. For example, Speelman & Geeraerts (2008) illustrated that concept characteristics should be taken into account when studying lexical variation from an aggregate perspective. In a more recent corpus-based study on identifying lexical differences between Netherlandic and Belgian Dutch, Ruette et al. (2013) contrasted an approach that takes the semantic relationship between words into account with an approach that does not. The semantically controlled approach discussed by Ruette and colleagues is based on profiles (Geeraerts et al. 1999, Speelman et al. 2003) that contain the relative frequencies of a set of words from a specific conceptual category. For example, the relative frequencies of words such as 'subway,' 'metro,' and 'underground' may differ for one language variety compared with another. A profile-based distance between two language varieties is obtained by calculating a distance between the profiles for each domain. Aggregate distances between two language varieties are then obtained by averaging all profile-based distances and frequency-weighting them (i.e., a profile containing many highly frequent words is considered more important than a profile that consists of less frequent words). In contrast, the aggregate distance without semantic control is obtained simply by comparing the lists of relative frequencies of all words. After evaluating the two methods, Ruette et al. (2013) found that if the semantic relationship between the words was ignored, the aggregate approach did not show the expected distinction between Netherlandic and Belgian Dutch. When the semantic relationship was taken into account, however, the distinction was clearly identified. These results are also in line with a study by Heylen & Ruette (2013), who (using semantic vector space models, a different method suitable for automatically identifying lexical variables; see Peirsman et al. 2010) showed that taking the semantic relationship into account influences the results of an aggregate study of lexical variation. Whereas the studies by Ruette et al. (2013) and Heylen & Ruette (2013) did not really incorporate geography (other than the distinction between the Netherlands

versus Belgium), other studies showed that taking into account the semantic domain is also important when looking at the geographical distribution of regional lexical variation (Speelman & Geeraerts 2008, Pickl 2013).

Chambers & Trudgill (1998, p. 187) aimed to join traditional dialectological research with modern sociolinguistics, and their book is a major step in that program. But an analytical gap has remained. On the one hand, sociolinguistics has mostly employed factorial designs that aim to assess the importance of one or another social factor in the distribution of single linguistic features, such as the pronunciation of /r/ in syllable-final positions. Clearly there would be no place for geography in a factorial design, unless one reduced geography to a categorical distinction among a small number of values. Dialectometry, on the other hand, has emphasized the geographic analysis of aggregate linguistic differences, typically using regression designs with geography operationalized as distance (but see Wieling 2012).

Wieling (2012) has also attempted to bring sociolinguistics and dialectometry closer together methodologically by including both geography and candidate social factors in large-scale regression designs. This approach, generalized additive mixed-effects regression modeling, is able to include geography (represented two-dimensionally rather than simply as a distance), social and geographical predictors, and their interaction in a single statistical model. By including many items, the importance of lexical predictors can be assessed as well, and the mixed-effects regression framework is well tailored to allow a focus on individual items (i.e., words or concepts) in the data set.

Generalized additive mixed-effects regression: mixed-effects regression approach in which the relationship between predictors and the dependent variable may be nonlinear

Mixed-effects regression: regression approach that takes into account the structural variation associated with individual speakers and words

Wieling et al. (2011) simultaneously took social predictors (such as the average age of the speakers in a location), lexical predictors (such as word frequency and word type), and geography into account in predicting the edit-distance-based pronunciation distance for hundreds of Dutch dialects from the standard Dutch language. In addition to identifying a clear effect of geography, which was moderated by word frequency (**Figure 2**) and word category, they found that smaller communities and older communities differed more from the standard Dutch language than did larger and younger communities. The direction of the effect was definitely expected given many earlier studies, but the innovative step was to measure the strength of these effects in a model wherein the nonlinear influence of geography had already been incorporated.

A more recent study by Wieling et al. (2014b) used a comparable model to investigate lexical variation in Tuscan dialects. Also in this case, they showed that the geographical pattern of lexical variation was not constant but rather varied depending on concept frequency and speaker age. In sum, the approach used by Wieling and colleagues is able to combine social, lexical, and geographical influences, bringing Chambers & Trudgill's (1998) vision of a unified discipline devoted to the study of language variation one step closer.

3.3. Language Change

Investigating linguistic change via dialectometry is not new, but recently more and more dialectometric studies have started to focus explicitly on this question. Many of these studies have looked at pronunciation distances and combined them with the apparent time construct to show the pattern of linguistic change. For example, Leinonen (2010) analyzed the vowel pronunciations of roughly 1,200 Swedish speakers in Sweden and Finland, comparing the old (about 65 years old) with the young (about 27 years old). She analyzed the two groups together, but then projected them separately onto two maps, which strikingly show the rapid leveling of dialects we see throughout most of Europe (**Figure 3**).

Maguire et al. (2010) examined pronunciation in English varieties spoken on the British Isles, proceeding from the aggregate pronunciation differences in a 100-word sample. They discovered

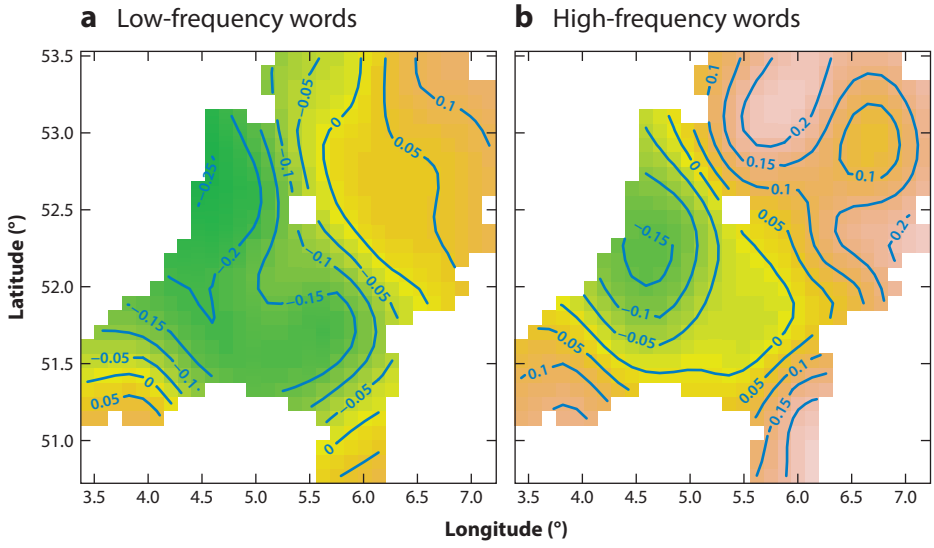

Figure 2

The pronunciation distance from standard Dutch as a function of longitude and latitude, and dependence on word frequency. (*a*) Low-frequency words. (*b*) High-frequency words. For the high-frequency words, the pronunciation distance from standard Dutch is largest, a finding that Wieling et al. (2011) interpreted as resistance to the influence of the standard Dutch language. Modified with permission from Wieling et al. (2011).

neither overall convergence nor overall divergence but rather complex developments. Valls et al. (2013) compared dialectometric maps based on data from older and younger Catalan speakers and concluded that dialect leveling had taken place due to the introduction of a standard language. Perea (2013) also focused on the Catalan language and introduced "dialectal stratigraphy" to show cartographically how language variation was patterned across time. The approaches used by Valls et al. (2013) and Perea (2013) are limited in the sense that they merely compare maps visually, but do not provide further analysis. In contrast, the aforementioned studies by Wieling et al. (2011, 2014b) statistically showed that age (of the speaker, or the average age of the community) needed to be included in the analysis and that younger speakers and communities were converging toward the standard language.

Using the apparent time construct, Heeringa & Hinskens (2014) statistically assessed change at three different linguistic levels: lexical, morphological, and pronunciation. Lexical and morphological distances were calculated by counting the proportion of linguistic items in which two dialects used a different lexical or morphological variant, whereas pronunciation distances were based on the edit distance. These authors found that the lexical level changed the most, whereas the morphological level was the most stable. Furthermore, they showed that Netherlandic dialects converged toward standard Dutch, whereas Belgian dialects did not show the same pattern.

Prokić & Cysouw (2013) studied a more specific question. Using a novel approach, they tried to identify sound correspondences showing regular change in a synchronic set of Bulgarian dialect pronunciations. Their approach started by calculating a sound similarity measure, the Poisson association, which is different from (for instance) the PMI-based measure employed by Wieling et al. (2012) in that it takes the nonnormal distribution of linguistic frequencies into account. On the basis of a set of multiply aligned Bulgarian dialect pronunciations, they obtained the sound similarity between every pair of sounds for every pair of dialects. For each individual sound position in each word of the multiply aligned corpus, they could then assess whether the sound

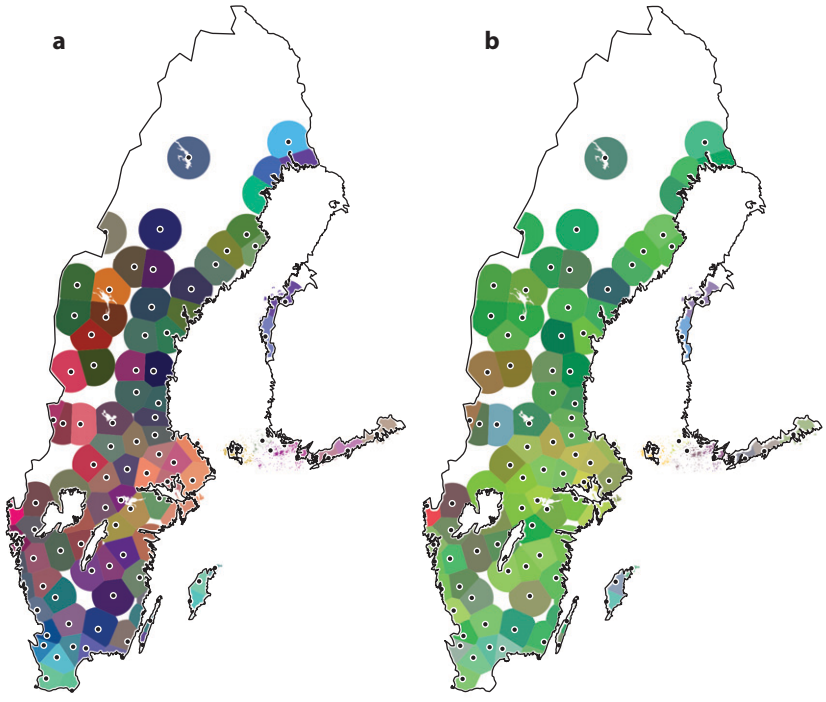

Figure 3

Leinonen (2010) analyzed more than 1,000 Swedish speakers using aggregate differences in vowel quality and subjected the result to multidimensional scaling. She then projected the vowel quality variability of (*a*) older (65 years) and (*b*) younger (27 years) speakers to these two maps, dramatically capturing how much variation is being lost in the ongoing process of dialect leveling. A great deal of the variation of the older speakers' speech, reflected by the range of colors in panel *a*, has been lost by the younger speakers. Modified with permission from Leinonen (2010).

substitution was regular or not on the basis of the Poisson association strength. By aggregating over all positions and visualizing these on a map, they were able to identify centers of innovation (i.e., being in the center of an area of regular changes), as well as transitional areas (i.e., containing many irregular sound correspondences).

In line with Prokić & Cysouw (2013), Montemagni et al. (2013) also used dialect atlas data to track the diachronic change of sound correspondences. Using bipartite spectral graph partitioning (mentioned above), they contrasted an older group of speakers with a younger group and showed that the younger speakers used more innovative sound correspondences.

The studies summarized above represent a sample of inquiries into linguistic changes in progress using dialectometric methods. They mark a relatively recent initiative in dialectometry, one that shows considerable promise both in elucidating large-scale changes and in engaging sociolinguists in the adjacent methodologies.

3.4. Dialectological Theory

Dialectometrists have eagerly pursued theoretical issues in dialectology. Language change is obviously related to how changes diffuse through a population, which has been a venerable research

goal in dialectology. For example, the diffusion of linguistic variation from a dialectometric perspective has been explicitly compared with Trudgill's (1974) "gravity theory," in which one expects population to act as gravitational mass and linguistic differences to rise with the square of geographic distance. Nerbonne & Heeringa (2007) examined the Low Saxon area in the Netherlands, concluding that population played a negligible role and that aggregate linguistic differences rose only sublinearly—and definitely not quadratically—with respect to geography. Heeringa et al. (2007) replicated this study on an area with a broader distribution of population sizes and saw the same sublinear dependence on geographic distance, but they also concluded that population indeed played the role suggested by the gravity hypothesis, adding significantly to the explained variance of the model. Nerbonne (2010) added two elements. The first was a comparative perspective, in which he examined the dependence of aggregate linguistic differences as measured by edit distance for the Netherlands (again), Germany, Norway, the eastern United States, Bulgaria, and Bantu Gabon, confirming in each case the sublinear rise shown earlier. Second, Nerbonne (2010) asked whether the dialectometric picture might be different only due to its focus on aggregate differences rather than differences among individual features (Trudgill's focus), which turned out not to be the case.

Spruit et al. (2009) examined phonetic, lexical, and syntactic differences as a function of geography in the Netherlands, showing that pronunciation differences depended most strongly on geographic distances, followed by syntactic differences and then by lexical ones. The best-fitting regression line in their syntactic analysis was linear, not sublinear, incidentally, but it definitely was not quadratic, as the gravity model predicts. Stanford (2012) suggests that geographic distance might be replaced by a more culturally specific variant, such as, for the small, clan-based Chinese indigenous society he studied, the distance between the rice paddies cultivated by the dialect speakers. Stanford's best-fitting line was not sublinear, however, but linear (and, again, definitely not quadratic). Naturally, his paper suggests that we might become much more creative in thinking about geographic distance, travel distance (Gooskens 2005), or other potential determinants of the degree to which dialects differ. Britain (2002) advocates a deeper engagement with human geography for dialectology, but dialectometry has yet to strike out firmly in this direction. The nonlinear model of geography put forward by Wieling (2012, chapters 6–8) seems promising in this sense because it allows one to directly model geography, rather than using more abstract geographical distances. Furthermore, the geographical pattern might be visually compared with georeferenced data sources (De Vriend et al. 2010).

Dialectometrists have also tried to estimate how much of linguistic variation might be explained by geography. This topic was implicit in Séguy's earliest paper (Séguy 1971), in which he graphed linguistic differences as a function of geographical distance, effectively plotting a regression line. Heeringa & Nerbonne (2001) compared areas and continua as organizing elements in geography, and Shackleton (2007) used both geographical distance and areal distinctions in a multiple regression design, showing that both contribute to the explanation of aggregate pronunciation differences in English dialects (in spite of some obvious collinearity). Nerbonne (2013) compared the dialects of several languages, demonstrating that pure geographical distance accounts for between 14% and 38% of the aggregate pronunciation differences. In addition, by examining German more closely, he showed that adding variables for areal distinctions increased the explanatory effect from 32% to 45%.

Falck et al. (2012; also see Lameli 2013, chapter 10) address the question of whether culture influences mobility, in particular whether people were more likely to move to another city or town if a very similar dialect was spoken there. They thus use dialect similarity as an operationalization of culture, and they assay dialect similarity in a dialectometric manner, proceeding from 66 thematic maps from the venerable Wenker atlas (Wrede et al. 1927). Administrative districts were

measured as linguistically similar to the degree that they coincided in these 66 features. Falck and colleagues use a standard model from economics to analyze mobility that assumes that people's readiness to move depends on the distance to the new domicile. We mention this point explicitly because the geographical distance is naturally inversely correlated with dialect similarity, and these authors take care to focus on the added explanatory contribution of dialectal similarity, which turns out to add approximately 10% to the effect of geographic distance alone. The studies mentioned here clearly indicate that dialectometry helpfully contributes to dialectological theory from a quantitative perspective.

3.5. New Data Sources

Because dialect atlases take many years of work to construct (and may sometimes suffer from inconsistencies caused by the large number of people involved in the construction of the atlases; the sidebar titled Correcting Transcription Inconsistencies Dialectometrically discusses the use of dialectometric methods to alleviate such inconsistencies), researchers in dialectometry have also tried to use other sources of data, especially corpora, to investigate dialect variation. Corpora have several advantages as data sources, which we examine in this section. In particular, the past five years have seen many interesting innovations in this respect. Szmrecsanyi (2011) has dubbed one such marriage of these research approaches "corpus-based dialectometry."

When studying pronunciation variation, dialectometrists generally use data of the sort collected by the compilers of linguistic atlases—pronunciations of the same word in different locations (and when studying lexical variation, they compare realizations of the same concept). Although speech productions are elicited in various ways for atlas compilation, the data are organized in such a way that the pronunciations and/or lexicalizations are linked to the specific words or concepts. In many situations, however carefully controlled, commensurable data are lacking. For example, one may obtain transcriptions of spontaneous speech by various dialect speakers. Because such data are not nicely aligned, in many cases they are ignored by dialectometrists. Scherrer (2012), however, proposes to determine the pronunciation distance between two speakers for this type of data by first identifying cognates (focusing on words similar with

CORRECTING TRANSCRIPTION INCONSISTENCIES DIALECTOMETRICALLY

The Belgian and Dutch field-workers in the Goeman–Taeldeman–van Reenen project (Goeman & Taeldeman 1996) gathered and transcribed the pronunciations of more than 1,800 words and phrases in 613 towns and villages throughout the Netherlands and Flanders (the northern half of Belgium, where Dutch is the native language). Unfortunately, the Dutch transcriptions were based on an alphabet of 83 phonetic symbols, whereas the Belgian field-workers used only 56 (Wieling et al. 2007), a circumstance that clearly threatens systematic comparison.

Wieling & Nerbonne (2011b) systematically sought pairs of phonetic symbols that were most similar by identifying those whose alignments showed that they were frequently used as alternatives. They used pointwise mutual information (PMI) for this purpose, a technique that Wieling et al. (2012) have shown to be capable of inducing acoustic differences from sheer distributional similarities.

By iteratively finding the symbol pair that contributed least to pronunciation distinction and merging the pair into a single symbol (thereby removing the distinction), Wieling & Nerbonne (2011b) reduced the combined phonetic inventory to 42 symbols. The aggregate edit distances of phonetic transcriptions obtained in this way correlated almost perfectly with the original distances ($r = 0.99$), enabling a reliable comprehensive comparison.

respect to the edit distance) and then determining the proportion of cognates that are identical. Scherrer applies his approach to Swiss dialects and shows that his method indeed detects patterns in line with common dialectological knowledge. Streck & Auer (2012) apply dialectometric techniques to pronunciations in spontaneous speech, resolving a dispute about which isoglosses ought to be taken as diagnostic in dividing the German Alemannic area. Schwarz (2014) includes a dialectometric chapter in his dissertation, which mainly applies other techniques to the same spontaneous speech data in the analysis of vocalism in the same German Alemannic area.

In contrast to pronunciation variation, Grieve (2009) focused on variation in written text and compiled a large, 25-million-word corpus of letters to the editor extracted from the websites of 200 regional newspapers in the United States. Grieve et al. (2011) then used this data set to investigate American regional lexical variation.

Grieve et al. (2014) also focused on data from the web and introduced site-restricted web searches to dialectometry. Their approach consisted of searching (e.g., via Google) the websites of local newspapers in the United States for the occurrences of certain lexical alternations (e.g., 'bag' and 'sack'). By obtaining the frequencies of these lexical alternation variables for a large set of local newspapers in the United States, Grieve and colleagues obtained a clear view of lexical dialect variation in the United States, generally corresponding with dialect boundaries on the basis of a previously conducted large-scale dialect survey. Of course, the advantage of Grieve et al.'s approach is that it is relatively fast compared with conducting a dialect survey, because the data are readily available.

In a technical paper, Eisenstein et al. (2010) focused on social media data from the microblogging website Twitter. They included only data containing the geographical coordinates from the user when a tweet (i.e., short message) was sent. By developing a model that incorporated two sources of lexical variation, topic and geographical region, Eisenstein et al. were able to detect regional differences, especially in the use of slang.

In addition to extracting data from various online corpora, it is also possible to develop new ways of obtaining data. In fact, Kolly & Leemann (2014) have proposed a promising new approach for dialectology. They created a smartphone application with which users could indicate their own (dialectal) pronunciation variant for 16 different words, after which the application predicted their dialect. In addition, users could upload their own pronunciations. This type of crowdsourcing proved to be extremely popular: More than 42,000 Swiss German speakers provided variant selection data, and almost 2,600 speakers provided pronunciation data. Even though only a limited number of words were included, clear dialectal patterns emerged on the basis of these words. In sum, corpora complement traditional dialect data sources and are a valuable resource to assess contemporary dialect variation.

3.6. More Attention to Morphosyntax

Numerous variationist studies have focused on pronunciation and lexical variation. In large part, this focus arose from the greater availability of this type of data in dialect atlases but also from the assumption that syntactic variation is not patterned geographically (noted by Szmrecsanyi 2014). Given the increased availability of new data sources, however, dialectometrists have become increasingly interested in studying language variation at the level of the lexicon and morphosyntax.

This development had already started by the turn of the century, and it is only accelerating. Spruit (2008) examined the variation in 500 binary morphosyntactic features in the *Syntactic Atlas of the Dutch Dialects* (*SAND*) (Barbiers et al. 2005) by using simple mismatch values in (categorical) features, which were summed to obtain a measure of syntactic difference. Spruit showed, among other things, that syntactic similarity occasionally disagreed with phonetic and lexical

measures, most strikingly in the north of the Netherlands, where the very sharp boundary between Lower Saxony (Groningen) and Friesland (whose language variety enjoys recognition as an independent language) blurred significantly. Intriguingly, Longobardi & Guardiano (2009) used similar measures to detect genealogical relationships. Sanders's (2010) dissertation, building on Lauttamus et al. (2007), explores a range of statistical techniques for their value in analyzing syntactic difference measures. More recently, Szmrecsanyi (2013) studied the geographical distribution of English morphosyntactic variation, Grieve (2012) investigated American English syntactic variation (i.e., adverb position), and Glaser (2013) studied Swiss German syntactic variation. Whereas clearly more attention has been paid to morphosyntax, it is still an area in which more dialectometric research is warranted.

Aggregate analysis: analysis based on averaging over many items

3.7. Tools

Although the open-source statistical package R (**http://www.R-project.org**) lacks standard facilities for dialectometric analyses, many of the more sophisticated analyses described above have been (fully or partly) conducted in R. For example, the spatial statistics approach used by Grieve (e.g., Grieve et al. 2011) and the individual differences scaling approach employed by Ruette & Speelman (2014) are based on existing R packages and custom-made R code. Unfortunately, however, the precise commands are not described in their publications, making these tools somewhat hard for others to use. In contrast, the generalized additive mixed-effects regression approach employed by Wieling et al. (2011, 2014b) also uses R, but these authors supply paper packages (available at the Mind Research Repository: **http://openscience.uni-leipzig.de**) that contain all data and all R commands used to replicate the results reported in those publications, making it easier for other researchers to follow the same approach. Of course, it takes practice to become skilled in R, but we emphasize that the investment is worthwhile, especially if more researchers make their R code and data publicly available.

For many years, there have been two standard computational applications in dialectometry: one developed at the University of Salzburg, VisualDialectoMetry (VDM) (Goebl 2004, Haimerl 2006), and the other developed at the University of Groningen by Peter Kleiweg, RUG/L04 (**http://www.let.rug.nl/kleiweg/l04**). Both tools conduct aggregate analyses and support a range of visualizations of aggregate dialect differences. The focuses of these two applications are slightly different. RUG/L04 focuses on determining the geographical distribution of aggregate differences in dialect pronunciations (on the basis of edit distance), whereas VDM bases its aggregate patterns on categorical distinctions between linguistic items. VDM is more user friendly, given that it has a graphical user interface, whereas RUG/L04 relies on separate command-line tools. VDM has seen extensive use in the analysis of Romance languages. Bauer (2009) provides a book-length study on the Dolomite Ladin dialects using VDM, and Goebl (2006) makes extensive use of VDM's analyses and attractive visualizations.

In addition, new online dialectometric tools have recently been developed, simplifying dialectometric analysis for many dialectologists. Nerbonne et al. (2011) developed Gabmap (**http://www.gabmap.nl**), which began as an online user-friendly version of RUG/L04. Gabmap is a web application that can work with categorical data such as lexical or morphological data, or numerical data such as formant frequencies, and it can compare transcribed pronunciations via edit distance. It also contains data-inspection tools intended to help find errors in the data and determine the geographical distribution of single features. Aggregate dialect patterns can be investigated using (noisy) cluster analysis and MDS, and there is a visualization option enabling a close comparison between the two. Since 2013, Gabmap has allowed users to detect characteristic features of dialect regions in a fairly general fashion (see Section 3.1). Wieling (2013) has

used Gabmap to allow researchers and laymen to explore data from the BBC Voices project, and Mathussek (2013) has used it to explore potential field-worker confounds in Middle Franconia (northwest Bavaria). A comprehensive review of Gabmap is offered by Snoek (2014).

Another web application with similar functionality as Gabmap, but with a greater focus on the types of analysis present in VDM, is DiaTech (Aurrekoetxea et al. 2013; also see **http://eudia.ehu.es/diatech**). A particular focus of DiaTech is the analysis of multiple responses, namely multiple values attributed to single sites or even respondents. Given that its development started more recently than that of Gabmap, it is still less polished. We suspect that dialectometry will be explored and criticized more deeply and will be able to improve more rapidly if additional tools are made (easily) accessible to working dialectologists and variationist linguists, and if existing tools are presented more frequently in tutorial form.

4. CONCLUSIONS AND OUTLOOK

After a brief historical introduction to dialectometry, we have discussed encouraging signs that work in dialectometry is engaging researchers in related fields. We have concentrated on various interesting dialectometric developments that have occurred during the past five years (also see Summary Points, below).

Chambers & Trudgill (1998, p. 140) optimistically closed a brief section on dialectometry in their book, *Dialectology*, as follows:

> It is not too soon to ask larger questions [. . .]. What does [dialectometry] reveal about [. . .] diffusion, about limits on differences and similarities among neighbors, and about common or universal patterns of gradience? The answers to those questions will require a comparative dialectometry, and a theoretical sensitivity commensurate with the data-handling and map-making that have so far provided the focus of activity.

As the section on dialectological theory demonstrates, dialectometrists have indeed made headway with respect to the larger questions, such as how much linguistic variation may be explained by geography, and with respect to comparative dialectology, in particular in characterizing the sort of gradience one finds with respect to geography. Even so, we respectfully suggest that Chambers & Trudgill's (1998) second requirement—theoretical sensitivity—should remain high on the list of research lines in dialectometry that deserve much more attention. Dialectometrists have made enormous progress in measuring dialect difference but have been less successful in turning this technical progress into a theoretical advantage. Lord Kelvin famously quipped that "When you can measure, [. . .] you know something; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind." (Thomson 1889). But he never suggested that measuring was sufficient.

SUMMARY POINTS

1. Dialectometry has made significant progress in developing techniques for identifying the most important (diagnostic) individual linguistic items underlying aggregate dialect variation.
2. Dialectometrists have developed various techniques to simultaneously analyze the linguistic and social factors conditioning geographical variation, and to gauge their relative strengths.
3. Dialectometry has ventured further afield to assess linguistic change in dialects.

4. Dialectometry has contributed to dialectological theory, in particular to comparative dialectology and to the theory of dialect diffusion.
5. Dialectometry is now paying increasing attention to morphosyntax, supplementing earlier work in lexical and phonological variation.
6. Dialectometry is tapping into data sources beyond the traditional dialect atlas, most notably dialect corpora constructed from online sources.
7. The creation of new (online) dialectometrical applications is enabling more dialectologists to use dialectometric tools.

FUTURE ISSUES

1. Dialectometry should focus more on theoretical questions in dialectology.
2. Dialectometry should be subjected to more critical analysis by other dialectologists and sociolinguists.
3. The time is ripe for sorting out which computational and statistical procedures are best suited for detecting and analyzing synchronic variation, and which for diachronic change.
4. Dialectometrists should devote more effort to making their tools publicly available to allow other researchers access to them.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Auer P, Hinskens F. 1996. The convergence and divergence of dialects in Europe. New and not so new developments in an old area. *Sociolinguistica* 10:1–30

Aurrekoetxea G, Fernandez-Aguirre K, Rubio J, Ruiz B, Sánchez J. 2013. 'DiaTech': a new tool for dialectology. *LLC J. Digit. Scholarsh. Humanit.* 28:23–30

Barbiers S, Bennis H, Devos M, de Vogelaer G, van der Ham M. 2005. *Syntactic Atlas of the Dutch Dialects* (*SAND*). Amsterdam: Amsterdam Univ. Press

Bauer R. 2009. *Dialektometrische Einsichten: Sprachklassifikatorische Oberflächenmuster und Tiefenstrukturen im lombardo-venedischen Dialektraum und in der Rätoromania*. St. Martin in Thurn, Italy: Ist. Ladin Micurà de Rü

Black P. 1976. Multidimensional scaling applied to linguistic relationships. *Cah. Inst. Linguist. Louvain* 3:43–92

Bloomfield L. 1933. *Language*. New York: Holt, Rhinehart & Winston

Borin L, Saxena A, ed. 2013. *Approaches to Measuring Linguistic Differences*. Berlin/Boston: Walter de Gruyter

Britain D. 2002. Space and spatial diffusion. In *The Handbook of Variation and Change*, ed. JK Chambers, P Trudgill, N Schilling-Estes, pp. 603–37. Oxford, UK: Blackwell

Chambers JK, Trudgill P. 1998. *Dialectology*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.

Chun Y, Griffith DA. 2013. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks, CA: Sage

De Vriend F, Boves L, van Hout R, Swanenberg J. 2010. Visualization as a research tool for dialect geography using a geo-browser. *LLC J. Digit. Scholarsh. Humanit.* 26:17–34

Dhillon I. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery*, pp. 269–74. New York: ACM

Eisenstein J, O'Connor B, Smith NA, Xing EP. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processes*, ed. J Li, L Màrquez, pp. 1277–87. Stroudsburg, PA: Assoc. Comput. Linguist.

Embleton S. 1993. Multidimensional scaling as a dialectometric technique: outline of a research project. In *Contributions to Quantitative Linguistics*, ed. R Köhler, B Rieger, pp. 267–76. Dordrecht, Neth.: Kluwer

Embleton S, Uritescu D, Wheeler ES. 2013. Defining dialect regions with interpretations: advancing the multidimensional scaling approach. *LLC J. Digit. Scholarsh. Humanit.* 28:13–22

Falck O, Heblich S, Lameli A, Südekum J. 2012. Dialects, cultural identity, and economic exchange. *J. Urban Econ.* 72:225–39

Geeraerts D, Grondelaers S, Speelman D. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Inst.

Glaser E. 2013. Area formation in morphosyntax. In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, ed. P Auer, M Hilpert, A Stukenbrock, B Szmrecsanyi, pp. 195–221. Berlin: Mouton de Gruyter

Presents the first book-length introduction to dialectometry.

Goebl H. 1982a. *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Österr. Akad. Wiss.

Goebl H. 1982b. Ansätze zu einer computativen Dialektometrie. In *Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, ed. W Besch, U Knoop, W Putschke, HE Wiegand, pp. 778–92. Berlin/New York: Mouton de Gruyter

Goebl H. 2004. VDM—visual dialectometry. Vorstellung eines dialektometrischen Software-Pakets auf CD-ROM (mit Beispielen zu ALF und Dees 1980). In *Romanistik und neue Medien*, ed. W Dahmen, G Holtus, J Kramer, M Metzeltin, W Schweickard, O Winkelmann, pp. 209–41. Tübingen, Ger.: Narr

Goebl H. 2006. Recent advances in Salzburg dialectometry. *Lit. Linguist. Comput.* 21:411–35

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

Gooskens C. 2005. Travel time as a predictor of linguistic distance. *Dialectol. Geolinguist.* 13:38–62

Gooskens C. 2013. Methods for measuring intelligibility of closely related language varieties. In *Handbook of Sociolinguistics*, ed. R Bayley, R Cameron, C Lucas, pp. 195–213. Oxford, UK: Oxford Univ. Press

Gooskens C, Beijering K, Heeringa W. 2009. Phonetic and lexical predictors of intelligibility. *Int. J. Humanit. Arts Comput.* 2:63–81

Grieve J. 2009. *A corpus-based regional dialect survey of grammatical variation in written Standard American English*. PhD thesis, North. Ariz. Univ., Flagstaff. 340 pp.

Grieve J. 2011. A regional analysis of contraction rate in written Standard American English. *Int. J. Corpus Linguist.* 16:514–46

Grieve J. 2012. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguist. Linguist. Theory* 8:39–72

Grieve J. 2013. A statistical comparison of regional phonetic and lexical variation in American English. *LLC J. Digit. Scholarsh. Humanit.* 28:82–107

Grieve J. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, ed. B Szmrecsanyi, B Wälchli, pp. 53–88. New York: Walter de Gruyter

Grieve J, Asnaghi C, Ruette T. 2014. Site-restricted web searches for data collection in regional dialectology. *Am. Speech* 88:413–40

Presents the first application of geostatistical techniques in dialectometry.

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

Grieve J, Speelman D, Geeraerts D. 2013. A multivariate spatial analysis of vowel formants in American English. *J. Linguist. Geogr.* 1:31–51

Haag K. 1898. *Die Mundarten des oberen Neckar- und Donaulandes* (*schwäbisch-alemannisches Grenzgebiet: Baarmundarten*). Reutlingen, Ger.: Buchdruckerei Hutzler

Haimerl E. 2006. Database design and technical solutions for the management, calculation, and visualization of dialect mass data. *Lit. Linguist. Comput.* 21:437–44

Hale K, Harris D. 1979. Historical linguistics and archaeology. In *Handbook of North American Indians*, Vol. 9, ed. A Ortiz, pp. 170–77. Washington, DC: Smithsonian Inst.

Heeringa W. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, Univ. Groningen. 315 pp.

Extensively examines the use of edit distance to measure pronunciation differences.

Heeringa W, Hinskens F. 2014. Convergence between dialect varieties and dialect groups in the Dutch language area. In *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, ed. B Szmrecsanyi, B Wälchli, pp. 26–52, 452–53. New York: Walter de Gruyter

Heeringa W, Nerbonne J. 2001. Dialect areas and dialect continua. *Lang. Var. Change* 13:375–400

Heeringa W, Nerbonne J, van Bezooijen R, Spruit MR. 2007. Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied. *Ned. Taalkd. Letterkd.* 123:70–82

Heylen K, Ruette T. 2013. Degrees of semantic control in measuring aggregated lexical distances. See Borin & Saxena 2013, pp. 353–73

Jäger G. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Lang. Dyn. Change* 3:245–91

Kessler B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 60–67. San Francisco: Morgan Kaufmann

Presents the first application of edit distance to dialectometry.

Kolly M-J, Leemann A. 2014. Dialäkt Äpp: Communicating dialectology to the public—crowdsourcing dialects from the public. In *Trends in Phonetics and Phonology in German-Speaking Europe*, ed. A Leemann, M-J Kolly, V Dellwo, S Schmid. In press

Kondrak G. 2013. Word similarity, cognation, and translational equivalence. See Borin & Saxena 2013, pp. 375–85

Kretzschmar WA Jr. 2010. Language variation and complex systems. *Am. Speech* 85:263–86

Kretzschmar WA Jr. 2012. Variation in the traditional vowels of the Eastern states. *Am. Speech* 87:378–90

Kretzschmar WA Jr, Kretzschmar BA, Brockman IM. 2013. Scaled measurement of geographic and social speech data. *LLC J. Digit. Scholarsh. Humanit.* 28:173–87

Kruskal J. 1999. An overview of sequence comparison. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. D Sankoff, J Kruskal, pp. 1–44. Stanford, CA: Cent. Study Lang. Inf.

Lameli A. 2013. *Strukturen im Sprachraum. Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Berlin/Boston: Walter de Gruyter

Lauttamus T, Nerbonne J, Wiersma W. 2007. Detecting syntactic contamination in emigrants: the English of Finnish Australians. *SKY J. Linguist.* 21:273–307

Leinonen T. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. PhD thesis, Univ. Groningen. 237 pp.

Levenshtein V. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Dokl. Akad. Nauk SSSR* 163:845–48

List J-M. 2012. Phonetic alignment based on sound classes. In *New Directions in Logic, Language, and Computation*, ed. M Slavkovik, D Lassiter, pp. 32–51. Berlin/Heidelberg: Springer

List J-M, Moran S. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the 51st Conference of the Association for Computational Linguistics*, pp. 13–18. Stroudsburg, PA: Assoc. Comput. Linguist.

Longobardi G, Guardiano C. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* 119:1679–706

Maguire W, McMahon A, Heggarty P, Dediu D. 2010. The past, present, and future of English dialects: quantifying convergence, divergence, and dynamic equilibrium. *Lang. Var. Change* 22:69–104

Mathussek A. 2013. Sprachgrenzen und Sprachräume im Mittelfranken. In *Handbuch zum Sprachatlas von Mittelfranken. Dokumentation und Auswertung*, ed. HH Munske, A Mathussek, pp. 244–72. Heidelberg, Ger.: Winter

Montemagni S, Wieling M, de Jonge B, Nerbonne J. 2013. Synchronic patterns of Tuscan phonetic variation and diachronic change: evidence from a dialectometric study. *LLC J. Digit. Scholarsh. Humanit.* 28:157–72

Nerbonne J. 2006. Identifying linguistic structure in aggregate comparison. *Lit. Linguist. Comput.* 21:463–76

Nerbonne J. 2009. Data-driven dialectology. *Lang. Linguist. Compass* 3:175–98

Nerbonne J. 2010. Measuring the diffusion of linguistic change. *Philos. Trans. R. Soc. B* 365:3821–28

Nerbonne J. 2013. How much does geography influence language variation? In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, ed. P Auer, M Hilpert, A Stukenbrock, B Szmrecsanyi, pp. 220–36. Berlin: Mouton de Gruyter

Nerbonne J, Colen R, Gooskens C, Kleiweg P, Leinonen T. 2011. Gabmap—a web application for dialectology. *Dialectologia* II(spec. issue):65–89

Nerbonne J, Heeringa W. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In *Roots: Linguistics in Search of Its Evidential Base*, ed. S Featherston, W Sternefeld, pp. 267–97. Berlin: Mouton de Gruyter

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

Nerbonne J, van Ommen S, Gooskens C, Wieling M. 2013. Measuring socially motivated pronunciation differences. See Borin & Saxena 2013, pp. 107–40

Peirsman Y, Geeraerts D, Speelman D. 2010. The automatic identification of lexical variation between language varieties. *Nat. Lang. Eng.* 16:469–91

Perea M-P. 2013. Dynamic cartography with diachronic data: dialectal stratigraphy. *LLC J. Digit. Scholarsh. Humanit.* 28:147–56

Pickl S. 2013. Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry. *LLC J. Digit. Scholarsh. Humanit.* 28:63–81

Pickl S, Spettl A, Pröll S, Elspaß S, König W, Schmidt V. 2014. Linguistic distances in dialectometric intensity estimation. *J. Linguist. Geogr.* In press

Prokić J. 2010. *Families and resemblances*. PhD thesis, Univ. Groningen. 186 pp.

Prokić J, Van de Cruys T. 2010. Exploring dialect phonetic variation using PARAFAC. In *Proceedings of the 11th Meeting of the Association of Computational Linguistics, Special Interest Group on Computational Morphology and Phonology* (*ACL-SIGMORPHON*), ed. J Heinz, L Cahill, R Wicentowski, pp. 46–53. Stroudsburg, PA: Assoc. Comput. Linguist.

Prokić J, Çöltekin Ç, Nerbonne J. 2012. Detecting shibboleths. In *Proceedings of the 2012 Joint Workshop of LINGVIS and UNCLH*, pp. 72–80. Stroudsburg, PA: Assoc. Comput. Linguist.

Prokić J, Cysouw M. 2013. Combining regular sound correspondences and geographic spread. *Lang. Dyn. Change* 3:147–68

Prokić J, Moran S. 2013. Black box approaches to genealogical classification and their shortcomings. See Borin & Saxena 2013, pp. 429–45

Prokić J, Nerbonne J. 2013. Analyzing dialects biologically. In *Classification and Evolution in Biology, Linguistics and the History of Science: Concepts—Methods—Visualization*, ed. H Fangerau, H Geisler, T Halling, W Martin, pp. 149–61. Stuttgart, Ger.: Steiner

Prokić J, Wieling M, Nerbonne J. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* (*LaTeCH-SHELT&R*), ed. L Borin, P Lendvai, pp. 18–25. Athens: Assoc. Comput. Linguist.

Pröll S. 2013. Detecting structures in linguistic maps—fuzzy clustering for pattern recognition in geostatistical dialectometry. *LLC J. Digit. Scholarsh. Humanit.* 28:108–18

Pröll S. 2014. Stochastisch gestützte methoden der Dialectdifferenzierung. In *Dialekte im Kontakt. Beiträge zur 17. Arbeitstagung zur alemannischen Dialektologie*, ed. D Huck, F Bogatto, P Erhart. Stuttgart, Ger.: Steiner. In press

Pröll S, Pickl S, Spettl A. 2014. Latente Strukturen in geolinguistischen Korpora. In *Deutsche Dialekte—Konzepte, Probleme, Handlungsfelder*, ed. M Elmentaler, M Hundt, JE Schmidt. In press

Briefly summarizes dialectometry's motivations and goals.

Rama T, Borin L. 2015. Comparative evaluation of string similarity measures for automatic language classification. In *Sequences in Language and Text*, ed. GK Mikros, J Macutek. Berlin: Walter de Gruyter. In press
Ruette T, Speelman D. 2014. Transparent aggregation of variables with individual differences scaling. *LLC J. Digit. Scholarsh. Humanit.* 29:89–106
Ruette T, Geeraerts D, Peirsman Y, Speelman D. 2014. Semantic weighting mechanisms in scalable lexical sociolectometry. In *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, ed. B Wälchli, B Szmrecsanyi, pp. 205–30. Berlin: Walter de Gruyter
Ruette T, Speelman D, Geeraerts D. 2013. Lexical variation in aggregate perspective. In *Pluricentricity: Linguistic Variation and Sociocognitive Dimensions*, ed. A Soares da Silva, pp. 95–116. Berlin: Walter de Gruyter
Rumpf J, Pickl S, Espass S, König W, Schmidt V. 2009. Structural analysis of dialect maps using methods from spatial statistics. *Z. Dialectol. Linguist.* 76:280–308
Sanders NC. 2010. *A statistical method for syntactic dialectometry.* PhD thesis, Indiana Univ., Bloomington. 155 pp.
Sanders NC, Chin SB. 2009. Phonological distance measures. *J. Quant. Linguist.* 16:96–114
Scherrer Y. 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the 2012 Joint Workshop of LINGVIS and UNCLH*, pp. 63–71. Stroudsburg, PA: Assoc. Comput. Linguist.
Schneider E. 1988. Qualitative vs. quantitative methods of area delimitation in dialectology: a comparison based on lexical data from Georgia and Alabama. *J. Engl. Linguist.* 21:175–212
Schwarz C. 2014. *Phonologischer Dialektwandel in den alemannischen Basisdialekten Südwestdeutschlands im 20. Jahrhundert. Eine empirische Untersuchung zum Vokalismus*. Stuttgart, Ger.: Franz-Steiner. In press
Séguy J. 1971. La relation entre la distance spatiale et la distance lexicale. *Rev. Linguist. Rom.* 35:335–57
Séguy J. 1973. La dialectométrie dans l'atlas linguistique de la Gascogne. *Rev. Linguist. Rom.* 37:1–24
Shackleton RG Jr. 2005. English–American speech relationships: a quantitative approach. *J. Engl. Linguist.* 33:99–159
Shackleton RG Jr. 2007. Phonetic variation in the traditional English dialects. A computational analysis. *J. Engl. Linguist.* 35:30–102
Snoek C. 2013. Using semantically restricted word-lists to investigate relationships among Athapaskan languages. See Borin & Saxena 2013, pp. 231–48
Snoek C. 2014. Review of Gabmap: doing dialect analysis on the web. *Lang. Doc. Conserv.* 8:192–208
Speelman D, Geeraerts D. 2008. The role of concept characteristics in lexical dialectometry. *Int. J. Humanit. Arts Comput.* 2:221–42
Speelman D, Grondelaers S, Geeraerts D. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Comput. Humanit.* 37:317–37
Spruit MR. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects.* PhD thesis, Univ. Amsterdam. 157 pp.
Spruit MR, Heeringa W, Nerbonne J. 2009. Associations among linguistic levels. *Lingua* 119:1624–42
Stanford JN. 2012. One size fits all? Dialectometry in a small clan-based indigenous society. *Lang. Var. Change* 24:247–78
Streck T, Auer P. 2012. Das raumbildende Signal in der Spontansprache: Dialektometrische Untersuchungen zum Alemannischen in Deutschland. *Z. Dialectol. Linguist.* 79:149–88
Szmrecsanyi B. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6:45–76
Szmrecsanyi B. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry.* Cambridge, UK: Cambridge Univ. Press
Szmrecsanyi B. 2014. Methods and objectives in contemporary dialectology. In *Contemporary Approaches to Dialectology: The Area of North, Northwest Russian and Belarusian Vernaculars*, ed. IA Seržant, B Wiemer. In press
Tamminga M. 2013. Phonology and morphology in Dutch indefinite determiner syncretism: spatial and quantitative perspectives. *J. Linguist. Geogr.* 1:115–24
Thomson W. 1889. Electrical units of measurement. In *Popular Lectures and Addresses*, Vol. I, pp. 73–136. London: Macmillan

Effectively launches the field of dialectometry.

Trudgill P. 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Lang. Soc.* 3:215–46

Valls E, Wieling M, Nerbonne J. 2013. Linguistic advergence and divergence in north-western Catalan: a dialectometric investigation of dialect leveling and border effects. *LLC J. Digit. Scholarsh. Humanit.* 28:119–46

Wichmann S, Holman EW, Bakker D, Brown CH. 2010. Evaluating linguistic distance measures. *Physica A* 389:3632–39

Presents a series of methods that aim to integrate dialectology and dialectometry.

Wieling M. 2012. *A quantitative approach to social and geographical dialect variation*. PhD thesis, Univ. Groningen. 178 pp.

Wieling M. 2013. Voices dialectometry at the University of Groningen. In *Analysing 21st-Century British English: Conceptual and Methodological Aspects of the BBC 'Voices' Project*, ed. C Upton, B Davies, pp. 208–18. London: Routledge

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

Wieling M, Bloem J, Mignella K, Timmermeister M, Nerbonne J. 2014a. Automatically measuring the strength of foreign accents in English. *Lang. Dyn. Change*. In press

Wieling M, Margaretha E, Nerbonne J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *J. Phon.* 40:307–14

Wieling M, Montemagni S, Nerbonne J, Baayen RH. 2014b. Lexical differences between Tuscan dialects and standard Italian: accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*. In press

Wieling M, Nerbonne J. 2009. Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processes*, pp. 14–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Wieling M, Nerbonne J. 2010. Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In *Proceedings of the 2010 Workshop on Graph-Based Methods for Natural Language Processes*, pp. 33–41. Stroudsburg, PA: Assoc. Comput. Linguist.

Wieling M, Nerbonne J. 2011a. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Comput. Speech Lang.* 25:700–15

Wieling M, Nerbonne J. 2011b. Measuring linguistic variation commensurably. *Dialectologia* II(spec. issue): 141–62

Presents the first dialectometric analysis combining lexical and social predictors while also taking geography into account.

Wieling M, Nerbonne J, Baayen RH. 2011. Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLOS ONE* 6:e23613

Wieling M, Nerbonne J, Bloem J, Gooskens C, Heeringa W, Baayen RH. 2014c. A cognitively grounded measure of pronunciation distance. *PLOS ONE* 9:e7574

Wieling M, Prokić J, Nerbonne J. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* (*LaTeCH-SHELT&R*), ed. L Borin, P Lendvai, pp. 18–25. Athens: Assoc. Comput. Linguist.

Wieling M, Shackleton RG Jr, Nerbonne J. 2013. Analyzing phonetic variation in the traditional English dialects: simultaneously clustering dialects and phonetic features. *LLC J. Digit. Scholarsh. Humanit.* 28:31–41

Wieling M, Upton C, Thompson A. 2014d. Analyzing the BBC Voices data: contemporary English dialect areas and their characteristic lexical variants. *LLC J. Digit. Scholarsh. Humanit.* 29:107–117

Wiersma W, Nerbonne J, Lauttamus. T. 2011. Automatically extracting typical syntactic differences from corpora. *LLC J. Digit. Scholarsh. Humanit.* 26:107–24

Woolhiser C. 2005. Political borders and dialect divergence/convergence in Europe. In *Dialect Change: Convergence and Divergence in European Languages*, ed. P Auer, F Hinskens, P Kerswill, pp. 236–62. New York: Cambridge Univ. Press

Wrede F, Mitzka W, Martin B. 1927. *Deutscher Sprachatlas. Auf Grund des von Georg Wenker begründeten Sprachatlas des Deutschen Reichs*. Marburg, Ger.: Elwert