Dialectometric concepts of space: Towards a variant-based dialectometry

Simon Pickl, University of Augsburg
Jonas Rumpf, Ulm University

to appear in: Hansen, Sandra / Christian Schwarz / Stoeckle, Philipp / Tobias Streck (eds.): Dialectological and folk dialectological concepts of space. Berlin: Walter de Gruyter.

Abstract
Outlining the development of quantitative methods in dialectology, especially dialectometry, we show that the conventional approaches are characterized by an exclusive focus on varieties. This is due to a "lect-based" concept of linguistic space. We argue that a different approach, taking into account the individual distributions of single variants of a geolinguistic corpus, is not only possible, but desirable as a complement to the traditional approach. By introducing a statistical method that allows us to classify linguistic variables according to their distributional patterns, we illustrate this "variant-based" approach. This method may reveal as yet unknown associations between those variables, while at the same time demonstrating the feasibility and the potential of an alternative quantitative approach to linguistic variation in space.

1. Introduction

In Nerbonne and Kretzschmar (2006: 387), the following characterization of dialectometry is given: "Dialectometric techniques analyze linguistic variation quantitatively, allowing one to aggregate over what are frequently rebarbative geographic patterns of individual linguistic variants […]. This leads to general formulations of the relation between linguistic variation and explanatory factors."

This definition implies that dialectometry employs "geographic patterns" ("rebarbative" though they may be) to formulate hypotheses about the determinants of spatial linguistic variation. Determining such geographical patterns quantitatively in the distribution of linguistic variants requires a concept of what linguistic space is: when do a number of samples at different points in space constitute a certain geographical pattern, and why?

The notion of linguistic space is, necessarily, conditioned by assumptions and perceptions that can be summed up as the "point of view" from which we look at language in space. The premises that determine the point of view that we take, which always involves a simplification (or "densification") of the data, vary according to the respective research objective (cf. Goebl 1994: 173). But what do we actually look at? Language in space consists of utterances that are made at different points in space (instances of *parole*) or, more abstractly, forms that can be expected to be uttered at different points in space (in their entirety forming an abstract *langue* that may differ from place to place, constituting local lects). The data that we have at hand in the form of dialect atlases therefore consist of utterances that are assigned to distinct, isolated coordinates – *a priori*, there are no dialect areas or boundaries. “Der Sprache als einer geistigen Tätigkeit kommt unmittelbar keine räumliche Ausdehnung zu. […] Es sind die Menschen, die Sprachschöpfer und -träger, die sich über den Raum verteilen. Die abkürzende Redeweise von der Verbreitung von Sprachen oder sprachlichen Erscheinungen läßt uns leicht vergessen, daß Menschen in ganz anderer Weise Raum einnehmen als z.B. Meere, Wüsten etc. […] Es gehört nun gerade zur Methode der Sprachgeographie […], so zu tun, als sei dies nicht der Fall” (Lang 1982: 63).

To achieve this, classical dialectometry looks at the lects aggregated from samples of single features and applies concepts regarding their linguistic and geographical relationship. To establish a perception of the geographical spread of dialects requires abstractions regarding the linguistic similarity of local lects and the geographical distance of the localities they belong to. Only when two localities feature a certain amount of matching language samples and geographical proximity, for example, can they be said to belong to the same dialect area. This requires the definition of a measure of linguistic similarity and a scale of how important geographical distance is considered for the delimitation of dialect areas (in classical dialectometry, geographical distance is usually not taken into account). Pairs of lects that do not exceed a certain level of discrepancy can then be regarded as belonging to the same dialect, or the deviations from one specified lect can be displayed as a dialect continuum in space. The underlying concept of linguistic space in this method is based on the premise that the lect as a whole (Lang’s “Sprache”) is what determines linguistic space, thereby defining the spatial configurations of the individual linguistic forms that constitute the lects (Lang’s “sprachliche Erscheinungen”) as mere deviations from an underlying pattern that is represented in the distribution of the lects. These configurations are neglected in that only patterns of similar lects can be detected, not patterns of linguistic forms.

If we read the above definition of dialectometry more carefully, however, it becomes clear that the linguistic forms are not aggregated to lects and then analysed further with respect to geographical patterns. Rather, patterns in the spread of “individual linguistic variants” are detected, which are then aggregated. This is clearly not the method of classical dialectometry, but it leads us to consider language in space from a different perspective. If we try to find patterns in the spreads of linguistic forms (“sprachliche Erscheinungen”) rather than of lects (“Sprachen”), and then “aggregate them” by assembling them into a corpus of geographical patterns, the concept of linguistic space is quite different. It is the single variant, then, that establishes linguistic space, which entails as many spatial configurations as there are linguistic variables. Why and how can this view of language in space be fruitful? The point is that the distributions of single variants are more than mere deviations from one underlying pattern constituted by the spread of lects; they differ so greatly that they deserve – and require – a closer, individual look. If ten out of a hundred linguistic variables show a recurrent, clear pattern which is not represented in the distribution of the lects as a whole, then relevant information is being ignored during the dialectometric process. A reversal of the sequence 1) *aggregation* 2) *detection of patterns* can prevent this information from being neglected and lead to a new methodology in dialectometry that is based on a different concept of linguistic space and directed at answering different questions than those typically asked by classical dialectometry.

In the following, we will have a closer look at the development of concepts of space in classical dialectometry and how they are motivated, and suggest what a variant-based dialectometry might look like.

2. Concepts of space in dialectometry

2.1 Classical dialectometry

2.1.1 Karl Haag

The origin of classical dialectometry is connected with the search for dialect boundaries. The idea that there are distinct dialects and distinct dialect areas – conceding a certain amount of transition between them – made it seem like an easy task to determine where one dialect starts and another ends. The linguistic facts, however, impeded the success of this approach. The dialects were not as uniform within and distinct from one another as was thought. Even though many of the isoglosses between linguistic variants seemed to coincide more or less with the imagined dialect boundaries, there were plenty that deviated considerably from them. Still, it was thought that dialect boundaries were constituted by bundles of isoglosses, so that one simply had to add up multiple isoglosses to determine where enough of them cluster together to form a veritable dialect boundary. The idea to accumulate isoglosses in order to obtain borderlines of varying "thickness" between the locations on a map – ultimately to interpret the strongest of those borderlines as boundaries of dialects – was first implemented by Karl Haag in 1898.

2.1.2 Jean Séguy

Jean Séguy, who suggested the term "dialectométrie" for quantitative dialect analysis ("leurs [dialectologists'] recherches des méthodes numériques", Séguy 1973a: 1), established a related method in the 1970s which allows for the measurement of the linguistic distance between lects spoken at different locations. By comparing the records of two adjacent locations and counting the differences, an index for the linguistic distance of the locations' lects is obtained (cf. Séguy 1973b; for a concise account of his method cf. Francis 1983: 142–144, 155–158). This index is then used to determine "dialect boundaries" by the highest linguistic distance values between locations, and draw them on a map, thus dividing it into "dialect areas". These boundaries represent the above-mentioned “bundles of isoglosses”. Through his dialectometric studies, however, Séguy became convinced that distinct dialect areas delimited by bundles of isoglosses are a delusion: “les ‘aires dialectales’ que les frontières de cette carte semblent circonscrire ne sont que de fausses aires. […] Les conclusions de Lalanne quant à l’inexistence des aires dialectales sont indestructibles. […] [L]es bourrelets signifient qu’il existe une différence linguistique notable entre deux séries de localités contiguës, et ne signifient que cela” (Séguy 1973a: 22–23). This was the dismissal of dialectometry from the dialect-area-committed roots which lay in traditional dialectology: “traditional dialectology has progressed from the original rather naïve notion of markedly distinct, well-bounded dialect areas [...] to the concept of variable change across territory, a concept which can be substantiated objectively by the statistical methods of dialectometry” (Francis 1983: 158). As for linguistic space, this marks the first major change of concept in dialectometry: it ceased to be seen as a mosaic put together by distinct dialect areas, which dialectometry aimed to detect. Rather, it started to be viewed as a spatial continuum of gradually changing language, with some parts showing only very subtle differences between locations, and some parts exhibiting sharper transitions.[1] It was this topography of variation which dialectometry was meant to explore.

2.1.3 Hans Goebl

Hans Goebl was the first to use computers for the calculation of linguistic distances, thus being able to process much larger amounts of data. He inverted the point of view from linguistic distance to linguistic similarity and extended Séguy’s method to the calculation of all possible pairs of locations, not only neighbours; the result is a (symmetric) *place* × *place* matrix, which includes the degree of similarity between the lects of all measuring points on the map. Such a similarity matrix is the main instrument of all classical dialectometric research. According to Goebl, the generation of a *place* × *place* similarity matrix out of a *place* × *linguistic variable* identity matrix is an instance of “Daten-” or “Informationsreduktion”: the data available are reduced to better suit the theoretical concept; information that is dispensable for the research interest is discarded (cf. Goebl 1994: 171–173).

But what is the theoretical concept; what is the research interest here? Clearly, the research interest is to explore the relatedness between lects in order to chart the geographical arrangement of linguistic relationships on a map; it is based on the concept that the calculatory similarity between two diatopic lects is the decisive factor that establishes linguistic space through its spatial collocation. As a consequence, the n linguistic relationships for each pair of locations (where n is the number of linguistic variables, i.e. a total of $n \cdot l(l\text{-}1)/2$ relations per dataset, l being the number

[1] This, of course, is an abbreviated and simplified account of a long controversy, which has been discussed for almost 200 years and has not yet ended (cf. Lang 1982: 184–209; Heeringa and Nerbonne 2001: 375–377).

of locations included) are aggregated to only one relationship per pair of locations (i.e. $l(l\text{-}1)/2$ relations on the whole). Thereby, linguistic space is detached from the spatial variability that is exhibited in the distribution of individual variants and instead based on the overall similarity of lects.

In addition to the refined methods for the calculation of the similarity index that Goebl has achieved, he has developed a range of methods for the analysis of the data held by the similarity matrix. The results are usually depicted by generating thematic maps (for a more detailed description of the following analysis and mapping procedures, see Goebl 1984 and Goebl 2006).

One of the more basic applications of Goebl's methods is the generation of similarity maps, in which each cell (representing a measuring point) displays the degree of linguistic similarity to one given point of reference (which, of course, bears a value of 100 % of similarity to itself). Therefore, there are l possible, discrete similarity maps per dataset. The often asymmetric image that the distribution of similarity values in space reveals gives an impression of the continuum of language variation across space, although it is always limited to the viewpoint of a single location. This is a sort of representation that can be useful to illustrate the integration of a location into its surrounding space, but on the whole it provides only a rather constricted view. One of its disadvantages is the effect that two locations which exhibit an identical similarity value (in comparison to the reference location) may still differ greatly from one another.

Parameter maps, as Goebl calls them, go beyond mapping data from the similarity matrix in that they provide a synopsis of certain characteristics obtained from the matrix. Maximum distribution maps display the highest similarity value (lower than 100 %) in each cell for the respective location. Despite the different approach, the kind of insights it yields does not differ greatly from that of "linguistic distance" maps in Séguy's manner (cf., for instance, the linguistic distance map in Goebl 2005: 529 and the maximum distribution map in Goebl 2006: 431). Even more than Séguy's method, Goebl's is very sensitive to the primary choice of locations: a denser mesh will produce higher maxima. Less sensitive to the mesh scale, but more so to the size and shape of the area under investigation, are so-called skewness maps. On these maps, the "skewness" of the similarity distributions for each location is charted, meaning that locations which have more similarity values above the arithmetic mean than below it are assigned a negative skewness value, and vice versa. Goebl interprets these values as indicators for the degree of isolation or integration of a location within the whole investigation area (cf. Goebl 2006: 419). The values depend greatly on the choice of the investigation area: if, for example, a segment is added in which locations have relatively low similarity values compared to the original area, the whole area will have higher skewness values. This arises from the fact that all locations are taken into consideration for each value, no matter the geographic distance. In this way, the size and shape of the whole map are decisive factors for the values on it. One cannot zoom in to a detail and expect the values in it to remain the same.

Goebl's main achievement in dialectometry, however, is arguably the introduction of cluster analysis as a means of numerical taxonomy (for a more detailed introduction to the application of cluster analysis in dialectometry, cf. Goebl 1983: 17–29). Cluster analysis is a popular taxonomic method for the classification of elements with pairwise similarity values, as provided by the dialectometric similarity matrix. In our case, the elements are the lects identified with the respective locations. They are agglomerated successively into clusters of growing size. At the beginning of the procedure, there are *l* clusters containing one location each; at the (theoretical) end of the procedure, there is one cluster containing *l* locations. At each step, those two clusters which bear the greatest similarity to one another are clustered together, thus reducing the number of clusters by one.[2] At a point where it seems reasonable,[3] the procedure can be stopped and the clusters accumulated up to that point can be displayed on the map, where they appear as distinct areas. Providing a non-geographical view of the results, they can be depicted as a tree (dendrogram), with each instance of clustering appearing as a ramification. There are numerous variations of this method, especially regarding the measurement of similarity between two clusters. As Goebl states, the choice of method is not always unequivocal: "Keines dieser Verfahren kann in Anbetracht des gegebenen Klassifikationszieles als ‚richtig' oder ‚falsch', sondern stets nur als ‚mehr oder weniger brauchbar' qualifiziert werden" (Goebl 1983: 17); however, "[f]ast alle der in der einschlägigen Literatur dazu beschriebenen Methoden ergeben brauchbare Resultate" (Goebl 2005: 511).

There are several implications arising from this method. Firstly, in theoretical terms, it appears to be a step back to the perception that there are distinct dialect areas because of the distinct clusters that are obtained. The interpretation as to what these areas actually signify, however, is difficult. "The synchronic interpretation concentrates on the determination of dialectal landscapes of different size and on their reciprocal dependence, i.e. similarity" (Goebl 2006: 421) – what the "reciprocal dependence, i.e. similarity" of two clusters (i.e. areas) actually is, however, remains obscure. Depending on the number of clusters chosen, the similarity within two pairs of clusters on one map can differ considerably; at the same time, two locations within one cluster can differ to a greater extent than two locations that belong to different clusters.

Secondly, Goebl aims at a diachronic interpretation looking at the dendrograms. "A diachronic interpretation simulates, as in a theoretical 'game', the progressive fragmentation of a

---

[2] We will not elaborate on the various methods for measuring the similarity between two clusters which have more than one element.

[3] "Die Problematik dieses Verfahrens liegt in der Kappung der Dendrogrammstränge, die jeweils auf unterschiedlichem Werteniveau vorgenommen wurde. Da diese Verclusterung nicht völlig automatisch durchzuführen ist, geht das Vorwissen des Dialektologen in seine Entscheidungen ein und steuert somit auch den gesamten Klassifizierungsprozeß" (Putschke 1993: 429).

given linguistic area, beginning at the first bifurcation after the root. These views […] depend on the basic assumption that ca. 1900 years ago [in this case] Galloromania represented a linguistically homogeneous area which diversified progressively over time" (Goebl 2006: 421). But "[d]ie Annahme einer solchen initialen Homogenität […] ist […] glücklicherweise stets kontrovers geblieben" (Goebl 1983: 23), so that the avail of such an interpretation is doubtful, as it is based on an assumption that Goebl himself rejects.

2.1.4 John Nerbonne and Wilbert Heeringa

The works of Nerbonne, Heeringa, and others have brought forth advances in mainly two fields of dialectometry. The first regards the calculation of the similarity values, for which they introduced a measurement that uses phonetic and hence genuinely gradual distances (cf. Nerbonne and Siedle 2005; Heeringa 2004: 27–143). Former methods, by contrast, rely on the counting of (dis)agreements, which can be 1 or 0, and become gradual only when aggregated. The other innovations concern data reduction techniques which deal with the problem of how to cope with vectors of an *n*-dimensional space (where *n* is the number of linguistic features) and display the results in the two dimensions of paper or computer screens. To this end, they introduced two new methods to dialectometry: multi-dimensional scaling, which, for instance, allows for the representation of dialect continua as colour transitions (cf. Heeringa 2004: 156–163), and factor analysis, which "proceeds from a matrix of correlations among variables, and, based on these, postulates common factors, which may be responsible for the correlations" (Nerbonne 2006: 468).

Nerbonne and Heeringa's theoretical orientation is twofold, as they accept "that both the area view and the continuum view are useful for gaining insight in the nature of the dialect landscape" (Heeringa and Nerbonne 2001: 399). Accordingly, Heeringa uses clustering alongside multidimensional scaling in his dissertation, generating distinct areas as well as continua (Heeringa 2004). In this respect, Nerbonne and Heeringa are in line with Goebl's view that the notions of "Stammbaum" (as represented in cluster analysis) and "Welle" (constituting dialect continua) are not mutually exclusive (cf. Goebl 1983).

2.1.5 Restrictions and limitations of lect-based dialectometry

What all dialectometric techniques discussed up to this point share, however, is that they rely exclusively on lects that are aggregated from a large number of linguistic features. Only recently have a number of studies appeared that look more closely at features individually (cf., for example, Clopper and Paolillo 2006; Cichocki 2006). A similar restriction concerns the extent to which space is actually part of the dialectometric calculations. Since Séguy's *relation entre la distance spatiale et la distance lexicale* (Séguy 1971), very little research (with the exception of some recent studies) concerning the correlation between linguistic and geographical distance has used geographical information for analysis (cf. Goebl 2006: 421–423; Goebl 2007; Spruit 2006). The fact that the coordinates of the locations play a role only when the results are charted on a map has already been attested for traditional dialectology: “although traditional dialectology is often (always?) portrayed as one of the earliest forms of geographical linguistics, in fact there is virtually no geographical contribution to the work at all. The role of space is reduced to that of data presentation on a map” (Britain 2002: 607).

2.2 “New” dialectometry: a “bottom-up” view of language in space

The restrictions to classical dialectometry that have been touched upon in the last paragraph lead to a situation where, *a priori*, certain dialectological questions cannot be answered. This is due to persistence on the theoretical approach that lects, not variants, constitute linguistic space. Necessarily, in data densification processes information is neglected that does not serve the actual research interest, although this information could be useful for the examination of other problems of dialectological interest. Naturally, these questions would concern the spatial distribution of single variants, as this is what is excluded from classical dialectometry by its reliance on the similarity matrix.

2.2.1 Spatial distribution of single linguistic features

As mentioned above, the spatial configurations of variants are more than mere deviations from one underlying pattern, but rather exhibit highly diverse structures and patterns in themselves (cf. Rumpf et al. 2009). Despite this high variance, some recurrent patterns can be discerned that can be used as the basis for a classification of maps. This has been attempted by several scholars (e.g. Wenzel 1930: 107–110; Frings 1956; Bach 1969: 39–226), although to date no quantitative geolinguistic (i.e. dialectometric) approach has been undertaken. The importance of a substantial quantitative approach to this subject lies in the fact that the variants’ distributions are to be seen as the results – or snapshots – of spatial diffusion processes. Each variant exhibits a different spatial spread, which is due to a different development: “chaque mot a son histoire” (attributed to Jean Gilliéron; cf. Christmann 1971). A classification of the spatial distributions, then, equals a classification of patterns of diffusion, which should allow us to gain insight into the mechanisms that are responsible for the way that variants develop in space.

This approach is quite different from that of classical dialectometry, as the research interest is a different one. While the former seeks to investigate the way in which dialects are distributed in space, we aim to analyse the way that single linguistic features, i.e. the smallest elements that constitute these dialects, are distributed in space. This bottom-up view of language in space not only recognizes the fact that each linguistic feature has a distribution of its own, but utilizes these distributions to establish a different notion of linguistic space. It is not considered a single partitioning of the two dimensions of the map, but a multitude of partitionings according to all linguistic features that show a spatial distribution.

The methods that we will discuss in the following represent results of the joint research project "New Dialectometry Using Methods from Stochastic Image Analysis" of the Chair of German Linguistics at the University of Augsburg and the Institute of Stochastics at Ulm University.[4] The main difference that distinguishes this new dialectometric approach from classical dialectometry is that lects play virtually no role in it – in other words, there is no similarity matrix. Instead, we look at the spatial distributions of single linguistic features to find out what caused these distributions. A quantitative approach to this question requires a quantification of the structural characteristics of the linguistic feature maps, which – as mentioned above – consist solely of geographically isolated records. From these data, it is possible to obtain values that are able to express intuitive concepts like "complexity" or "homogeneity" of a map. Thus, from a corpus of maps, a dataset containing these values for each map can be generated, which is the basis for a range of quantitative statistical tests concerning interrelations between the maps. Apart from these values, which reflect the "overall" geographic constitution of a map, it is possible to sort the maps according to their actual similarity, which can be measured by comparing the shapes, sizes, and positions of the areas on them.

2.2.2 Area-class maps

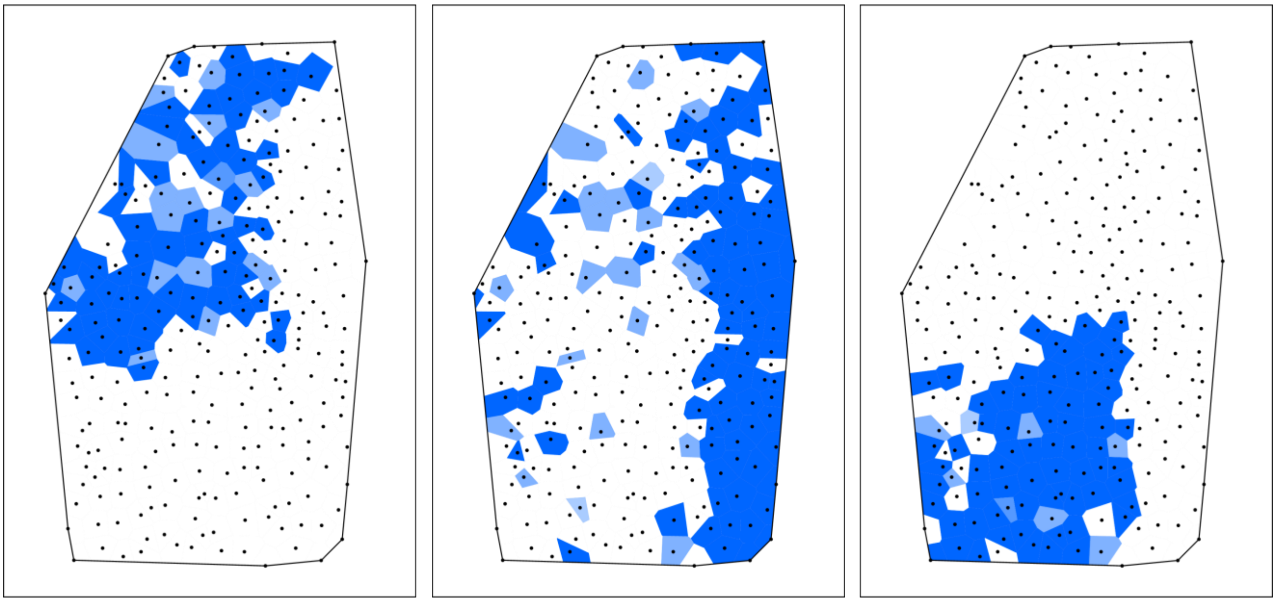
For the calculation of the structural characteristics, it is convenient to dissect the maps into prevalence areas of the single variants. Representing a conversion of pointwise measurement data into area-class maps, this approach is a response to Lang's claim that geolinguistics pretend that there is actually something like coherent linguistic space rather than discontinuous measuring points. To achieve this, it is necessary to relate the geographically isolated records to one another. Here, we rely on what was mentioned above as a constituent of linguistic space: the co-occurrence of linguistic similarity and geographic proximity. If two records are instances of the same variant and are found at more or less adjacent locations, then they can be seen as belonging to one variant area. If, however, a third, dissimilar record lies between them, this will weaken the other locations' cohesion.[5] The mathematical method that is employed to achieve this goal is intensity estimation. Inspired by point process statistics (cf. e.g. Diggle 2003; Illian et al. 2008), it views the set of all

---

[4] For a more detailed account of the methods, cf. Rumpf et al. (2009, 2010). All results and examples were obtained from data from the *Sprachatlas von Bayerisch-Schwaben* (König 1996–2009).

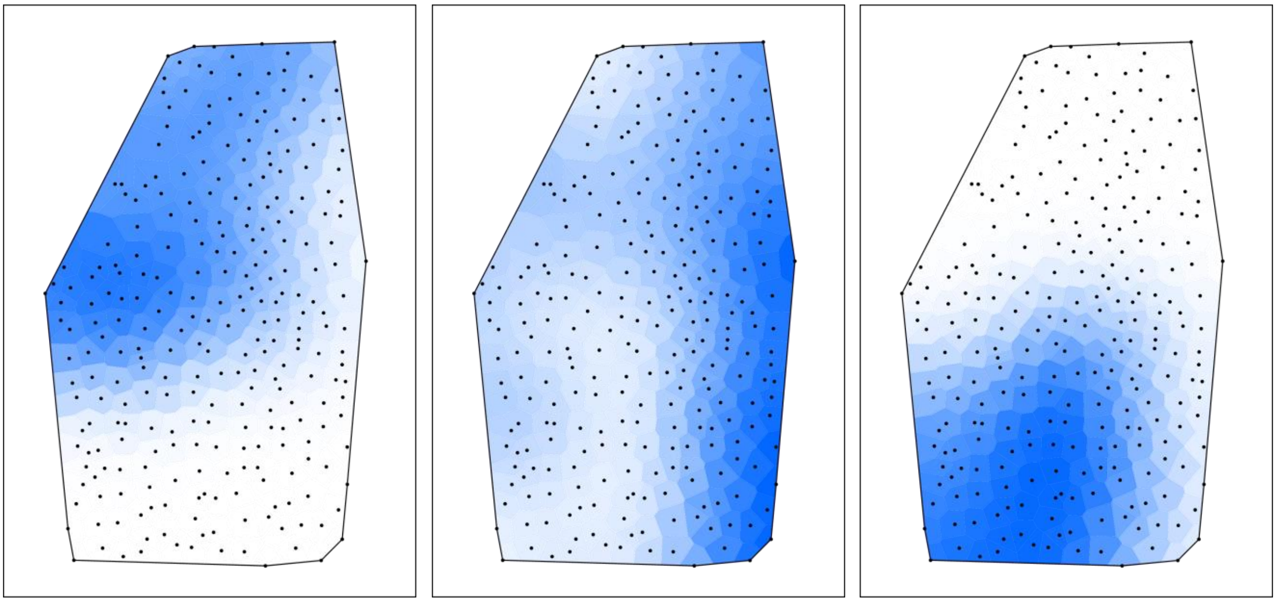
[5] In the first stage of our research, we only dealt with lexical maps, which feature nominal-scale data, i.e. the scale of linguistic similarity between variants is restricted to an opposition of 0 and 1.

locations where an occurrence of a certain variant *x* is recorded (cf. Figs. 1–3) as a spatial point pattern. By taking into account the spatial configuration of these points, intensity estimation creates an estimate of their spatial distribution. Simply put, while for locations in regions with many occurrences of variant *x* the estimated intensity for *x* should be high, the opposite should be the case for areas with few or no occurrences of that variant.



Figures 1–3 Records of three variants for 'woodlouse' in the investigation area of the *Sprachatlas von Bayerisch-Schwaben* (König 1996–2009, vol. 8, p. 230–233). The darkest shade of blue indicates that the respective variant is the only one occurring at a location, while the lighter shades of blue signify that (one or two of the) other variants are present at the same location.

Various techniques for intensity estimation exist (cf. e.g. Scott 1992 and Silverman 1986). The most common are so-called kernel estimation techniques: each location where a certain variant *x* is recorded is assigned a certain "mass", which extends into the location's surroundings, representing the "influence" of that occurrence of *x* on the locations in its environment. It is obvious that this influence should be largest at the exact location where *x* was recorded, and, relying solely on geographical distance for now, it should decline equally in all directions, but apart from that, the exact shape of this mass (the so-called "kernel") must be selected according to the specific application. Frequently, it is chosen to have the shape of a bell curve, i.e. the standard normal distribution. Furthermore, it is important to select an adequate "bandwidth", which is a parameter that determines how far the mass is stretched out. Then, by simply adding up the influences of variant *x* from all its occurrences, we arrive at the estimated intensity of *x* at any location of interest (cf. Figs. 4–6).



Figures 4–6 The estimated intensities of the three variants.

In this way, distinct intensity estimates for each variant that occurs on a map can be obtained. By combining these intensity estimates, we can assign any location *t* to a variant area, namely the area of the variant *x*(*t*), which is the variant with the highest estimated intensity at location *t*; we denote the area of a variant *x* by *T*(*x*). Applying this procedure to all locations will create a map with distinct variant areas, i.e. an area-class map (cf. Fig. 7), where the variant areas can, for example, be distinguished by varying colour hues. To refine this type of map, the value *b*(*t*), which denotes the proportion of the total estimated intensity at *t* that is taken up by *x*(*t*), i.e. the dominance of *x* at *t*, can be displayed as the brightness of *t*'s colour. For a more extensive and mathematically rigorous account of the creation of area-class maps through variant density estimation, see Rumpf et al. (2009).

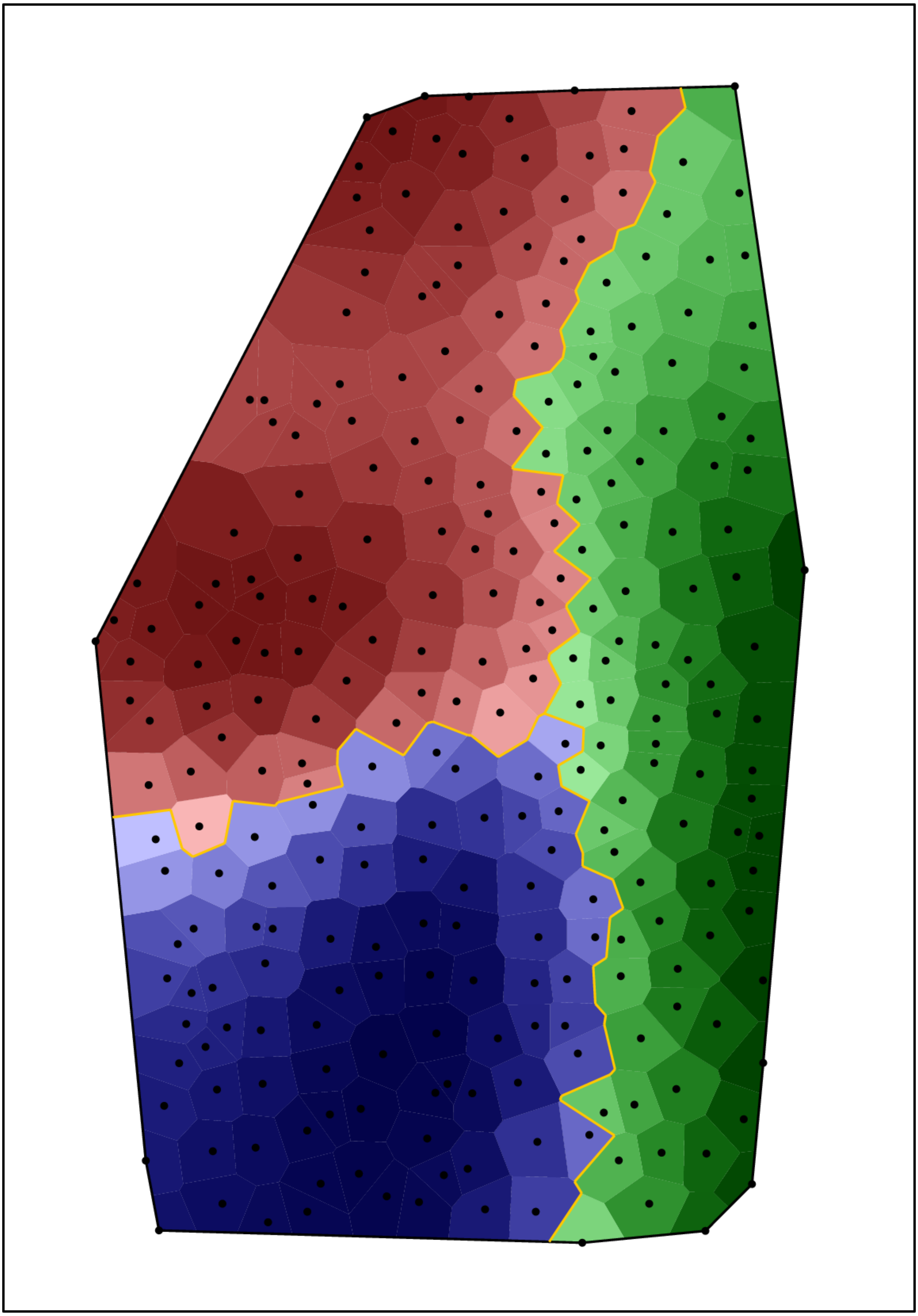


Figure 7 The intensities of the three variants combined to one area-class map. The different colour hues correspond to the respective variants.

As the term "intensity estimation" suggests, the results are estimations of each variant's actual spatial distribution, based on the geographical configuration of the observed occurrences. In a synopsis of the intensity fields of all variants, $b(t)$ can be interpreted loosely as the probability with which variant $x(t)$ will appear at location t, or as the degree of conclusiveness with which location t can be assigned to variant area $T(x)$. The boundaries that are obtained with this method are likewise the result of these values, indicated by the lines where it is expected that the probabilities of the occurrence of two variants counterpoise each other. The results that these estimations yield provide a useful tool for further feature map analysis and classification. In representing the geographic interrelation of linguistic elements, they allow for an assessment of the maps' structural characteristics, which are the basis for their classification according to their geographic constitution.

The (large-scale) *complexity* of a map can be measured by adding the lengths of all boundaries on it. Many smaller areas will result in higher complexity than fewer larger areas, but also borderlines that are jagged rather than smooth will increase the complexity of a map. A map's (small-scale) *heterogeneity* can be measured by calculating the arithmetic mean of all $b(t)$ on the map, which means that maps where variants do not interfere much with each other will be less heterogeneous. In other words, stray records in another variant's area will increase the overall heterogeneity. What we have dubbed the *fidelity* or *area compactness* of a map is a similar value, giving an index of how well the actual records are represented by the areas. This is important for the assessment of the map's suitability to be displayed as an area-class map. It is measured by calculating the mean fraction of stray records in an area. For details on these characteristics and example maps illustrating correspondence between certain values and visual features of the area class maps, see Rumpf et al. (2009).

2.2.3 Groupings of linguistic feature maps

The values mentioned above allow for a classification of linguistic feature maps according to their structural characteristics. This facilitates the search for linguistic factors that determine the variants' spatial distributions, be it frequency, semantic field, or something else. For the influence of geographical conditions such as rivers, roads, mountains etc. on the distribution of linguistic variants, however, a different approach must be taken, as the above characteristics such as *complexity* or *heterogeneity* give no account of the actual geographic layout of a map. If a considerable number of maps, for instance, show boundaries that are concurrent to a significant degree, this suggests that they are conditioned by some geographically defined circumstance that affects the amount of communication taking place between two parts of the investigation area. At the same time, this raises the question of why some maps are affected and others not. Therefore, we have developed methods for the classification of linguistic feature maps according to their actual similarity. As there is always a pairwise similarity between single maps, cluster analysis is an appropriate means of doing this. The clusters found, then, would contain maps that share certain characteristics, for example a diagonal boundary in the upper right corner, or a rectangular-shaped area in the middle, etc. (cf. Rumpf et al. 2010).

Thus, the groupings that are obtained with these methods can help identify geographical conditions that influence the distributions of variants. If structures are found that appear only in maps of certain types of variables (e.g. semantically defined), then there is strong evidence for a connection between a linguistic and a geographic fact. Finding out what these variables have in common can lead to an understanding of what factors determine whether certain geographical conditions influence the spread of variants.

For the purpose of clustering linguistic feature maps, their similarities must be quantified. Based on the dialectological area-class maps obtained with the method discussed above, various measures of similarity are plausible. Two of the most expedient measures will briefly be mentioned here. Preliminary investigations employing these clustering methods have yielded promising results; details will be reported in Rumpf et al. (2010).

The first method we propose to calculate the similarity between two maps draws on the boundaries between different variant areas. If two maps have identical boundaries, this entails that all pairs of points that are assigned to the same variant area/to different variant areas on one map are assigned to the same variant area/to different variant areas on the other map. A simple way to quantify the differences between two maps is to count the pairs of points that violate this condition. The measure of similarity is then easily obtained by subtracting this number from the total number of point pairs. Obviously this measure could be refined in various ways, for example by not counting each pair of points equally, but rather weighting them with the inverse of the geographical distance between them.

Another way to quantify the differences between two maps is to rely not on the affiliation of measuring points to variant areas, but on the estimated variant intensities. By calculating the differences of $b(t)$ and $b(t')$ for all pairs of points t and t' of one map, the structure of heterogeneity on the map is quantified. Subtracting the values of these characteristics of one map from the corresponding values of another map will yield a numerical value for each pair of points. The sum of all absolute values can then be seen as a measure of dissimilarity, with a small total sum indicating a high similarity. Again, various refinements of this measure are conceivable (cf. Rumpf et al. 2010).

Apart from the methods suggested here, various other applications of a variant-based dialectometry are possible. For example, once a measure of similarity between maps is fixed, it is not only possible to employ cluster analysis to obtain sets of similar maps. Also, maps with a specified pattern can be easily detected by creating an “artificial map” that exhibits the prototype of this pattern and then simply finding the maps that are most similar to the prototype. Furthermore, by restricting the measures described above to certain subsets of the points of measurement, one can easily obtain clusters of maps that show similarities, especially in certain sub-regions that are of particular interest, while differences in other regions are disregarded.

3. Conclusions

It has been shown that classical dialectometry does not exhaust the possibilities that quantitative geolinguistics has to offer, which is mainly due to its reliance on the similarity matrix, accompanied by a lect-based concept of linguistic space. A variant-based approach, which examines the geographic interrelation of records for one linguistic feature at a time before they are aggregated, can contribute to a wider range of dialectometric techniques, providing a means of comparing maps rather than accumulating them. In this way, the approach looks at the very information which is systematically excluded from classical dialectometric analyses, i.e. the variation among the spatialities of linguistic features.

This perspective is directed towards discerning the factors that are responsible for the various different layouts that can be found in the maps. Characteristics of these layouts are quantified so that patterns in them can be identified. These patterns can lead to an understanding of the factors that are responsible for the way in which linguistic features develop in space.

References

Bach, Adolf 1969 *Deutsche Mundartforschung. Ihre Wege, Ergebnisse und Aufgaben.* (Germanische Bibliothek. Dritte Reihe: Untersuchungen und Einzeldarstellungen.) Heidelberg: Winter.

Britain, David 2002 Space and spatial diffusion. In: J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (Blackwell Handbooks in Linguistics.), 603–637. Malden/Oxford: Backwell.

Cichocki, Wladyslaw 2006 Geographic variation in Acadian French /r/: what can correspondence analysis contribute toward explanation? In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 529–541.

Christmann, Hans Helmut 1971 Lautgesetze und Wortgeschichte. Zu dem Satz „Jedes Wort hat seine eigene Geschichte“. In: Eugenio Coseriu and Wolf-Dieter Stempel (eds.), *Sprache und Geschichte. Festschrift für Harri Meier zum 65. Geburtstag*, 111–124. München: Fink.

Clopper, Cynthia G. and John C. Paolillo 2006 North American English vowels: A factor-analytic perspective. In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 445–462.

Diggle, Peter J. 2003 *Statistical Analysis of Spatial Point Patterns*. 2nd Edition. London: Arnold.

Francis, W. Nelson 1983 *Dialectology. An Introduction.* London: Longman.

Frings, Theodor 1956 *Sprache und Geschichte II.* (Mitteldeutsche Studien 17.) Halle (Saale): Niemeyer.

Goebl, Hans 1983 „Stammbaum“ und „Welle“. Vergleichende Betrachtungen aus numerisch-taxonomischer Sicht. In: *Zeitschrift für Sprachwissenschaft* 2/1: 3–44.

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

Goebl, Hans 1994 Dialektometrie und Dialektgeographie. Ergebnisse und Desiderate. In: Klaus Mattheier and Peter Wiesinger (eds.), *Dialektologie des Deutschen. Forschungsstand und Entwicklungstendenzen*, 171–191. Tübingen: Niemeyer.

Goebl, Hans 2005 Dialektometrie. In: Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook* (Handbooks of Linguistics and Communication Science 27.), 498–531. Berlin/New York: de Gruyter.

Goebl, Hans 2006 Recent advances in Salzburg dialectometry. In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 411–435.

Goebl, Hans 2007 Kurzvorstellung der Korrelativen Dialektometrie. In: Peter Grzybek and

Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday* (Quantitative Linguistics 62.), 165–180. Berlin/New York: Mouton de Gruyter.

Haag, Karl 1898 *Die Mundarten des oberen Neckar- und Donaulandes. Schwäbisch-alemannisches Grenzgebiet: Baarmundarten*. Reutlingen: Hutzler.

Heeringa, Wilbert and John Nerbonne 2001 Dialect areas and dialect continua. In: *Language Variation and Change* 13/3, 375–400.

Heeringa, Wilbert 2004 *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen: Univ. Diss.

Illian, Janine, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan 2008 *Statistical Analysis and Modelling of Spatial Point Patterns.* Chichester: Wiley.

König, Werner1996–2009 *Sprachatlas von Bayerisch-Schwaben*. (Bayerischer Sprachatlas: Regionalteil 1.) 14 volumes. Heidelberg: Winter.

Lang, Jürgen 1982 *Sprache im Raum. Zu den theoretischen Grundlagen der Mundartforschung. Unter Berücksichtigung des Rätoromanischen und Leonesischen.* (Beihefte zur Zeitschrift für Romanische Philologie 185.) Tübingen: Niemeyer.

Nerbonne, John 2006 Identifying linguistic structure in aggregate comparison. In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 463–475.

Nerbonne, John and William Kretzschmar, Jr 2006 Progress in dialectometry: toward explanation. In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 387–397.

Putschke, Wolfgang 1993 Zur Kritik dialektologischer Einteilungskarten. In: Wolfgang Viereck (ed.), *Proceedings of the International Congress of Dialectologists, Bamberg 29.7.–4.8.1990. Plenary lectures, Computational data processing, Dialect structure and classification*, 421–443. Stuttgart: Steiner.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König, and Volker Schmidt 2009 Structural analysis of dialect maps using methods from spatial statistics. In: *Zeitschrift für Dialektologie und Linguistik* 76/3, 280–308.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König, and Volker Schmidt 2010 Quantification and statistical analysis of structural similarities in dialectological area-class maps. In: *Dialectologia et Geolinguistica* 18, 73–98.

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

Séguy, Jean 1973a La dialectométrie dans l'*Atlas linguistique de la Gascogne*. In: *Revue de Linguistique Romane* 37, 1–24.

Séguy, Jean (ed.) 1973b *Atlas linguistique de la Gascogne*, vol. VI. Paris: Centre National de la Recherche Scientifique.

Scott, David W. 1992 *Multivariate Density Estimation: Theory, Practice, and Visualisation.* New York: Wiley.

Spruit, Marco René 2006 Measuring syntactic variation in Dutch dialects. In: *Literary and Linguistic Computing* 21/4: Special Issue on Progress in Dialectometry, 493–506.

Wenzel, Walter 1930 *Wortatlas des Kreises Wetzlar und der umliegenden Gebiete.* (Deutsche Dialektgeographie XXVIII.) Marburg: Elwert.