Cartographic Visualization in Support of Dialectology

Pius Sibler, Robert Weibel, Elvira Glaser & Gabriela Bart

ABSTRACT: Using data from the Syntactic Atlas of Swiss German Dialects (SADS), several methods for the cartographic visualization of linguistic data are proposed, demonstrated, and evaluated against the requirements of linguistic research. After reviewing the challenges of linguistic data visualization, point symbol maps are introduced as a baseline. We then present alternative visualization methods that present linguistic data in new ways. The first uses Voronoi polygons about the data points to color in the dominant variant per location. The second uses kernel density estimation (KDE) to interpolate intensity values of all variants and thus infer and display the dominant variant per location (incl. at missing value locations). The KDE-based method helps to better see trends in the data and also automatically infer isoglosses. The third new technique uses 3-D visualization to support the exploration of spatial trends, as well as co-occurring variants. As a fourth alternative, measures and methods from geostatistics are used for the visualization of specific, global and local, variations and patterns.

KEYWORDS: Linguistic area formation, dialectology, cartographic visualization techniques, kernel density estimation, geostatistics

Introduction

In an attempt to document linguistic variation across geographic space and study the formation of linguistic areas (or, as geographers prefer to say, regions), such as those formed by different dialects of a language, linguists have been, and still are, routinely collecting linguistic data. These observation data provide evidence of the language spoken at selected locations, or sites. Linguistic data collections are then normally published in the form of atlases, such as LAMSAS (Linguistic Atlas of the Middle and South Atlantic States) in the USA (McDavid et al., 1980). Visualizing linguistic observations, however, poses formidable challenges. First, it is common for different variants of a linguistic feature to co-occur at the same geographic location. For instance, the "although" variant seems to be preferred over "though" in the Northeast of the USA (Grieve et al., 2011). However, in most places of the USA, both variants are possible. Second, data are typically only collected at relatively few selected locations, and these sites represented by points that are thought to act *en lieu* of entire municipalities (or other administrative areal units), with no observations available for most places. Third, from linguistic point data it is hard to infer clear-cut boundaries separating language areas (called isoglosses in linguistics, such as the 'boundary' between "though"/"although" occurrences in the USA), and there is also a certain arbitrariness to the definition of isoglosses. In the terminology of Smith and Varzi (2000) they thus clearly belong to the *fiat* type boundaries. Furthermore, isoglosses of different linguistic features very rarely overlap (though traditional

dialectological theory assumes the existence of bundles of isoglosses; Chambers and Trudgill, 1998). Thus, in an attempt of staying on the safe side, language atlases often rely on simple maps of the 'raw' point observations, possibly combined with selected isoglosses that have been visually ('manually') inferred.

This paper reviews several alternative methods for cartographic visualization of linguistic observations that have originally been collected at point locations. The techniques either use area-class maps, 3-D maps, or geostatistical mapping. Some of these have already been proposed elsewhere for dialectological mapping, while others are known in other fields but have not yet been applied to linguistic data. The main contribution of this paper is to bring these visualization techniques together and assess their utility in the context of dialectology.

Related Work

While many language atlases traditionally present dialect observations using point symbol maps (Fig. 1), a range of other cartographic methods has been used in linguistic research, though less widespread. Pi (2006) points to weaknesses of isoglosses and proposes an alternative technique called isograph. Rather than drawing boundaries, the isograph links areal units with the smallest percentage difference when compared to adjacent observations (e.g. using the percentages for "though" vs. "although" from the above example). Kretzschmar (2003) uses a similar technique, though using join-counts between neighboring sites over a Delaunay triangulation. The handbook edited by Lameli et al. (2010), gives a good overview of the various map types and visualization techniques used in concurrent linguistics and dialectology. In particular researchers working on dialectometry, that is, the quantitative analysis and mapping of dialect features, use visualization techniques that go beyond point symbol maps. Goebl (2010) (incl. also in his earlier work) advocates the use of 'honeycomb maps' and 'beam maps', that is, Voronoi diagrams and Delaunay triangulations, respectively, constructed about the linguistic point observations. Goebl's article also includes a map from the very early days of dialectology by Haag (1898), which depicts isoglosses and isogloss bundles in a Voronoi-like fashion. Voronoi-based maps are also used by dialectometrist Nerbonne (2009, 2010). Interestingly, however, while in both Goebl's and Nerbonne's work the Voronoi diagram is used as a spatial principle to structure geographic space and extend the point observations to form area-class maps, the actual attributes mapped in this Voronoi structure are not generated using geospatial principles such as spatial interpolation. Instead, both researchers focus on linguistic distances (as a measure of linguistic similarity) and multivariate, non-spatial methods to infer the mapped attributes. Distance measures such as the Levenshtein distance are used to express the similarity of linguistic features, and form the input to multivariate statistical techniques such as cluster analysis and multidimensional scaling (MDS), grouping sites according to aggregate linguistic similarity, hence yielding potential linguistic areas when re-projected to geographic space (i.e. mapped to the Voronoi diagram). The choice of multivariate methods is an effect of the interest of these researchers in mapping *aggregate* linguistic variation of many linguistic features (Nerbonne, 2009, 2010; Goebl, 2010).

Methods of spatial interpolation have been used by researchers in dialectometry with a focus on variations in single (or few) linguistic features. Rumpf et al. (2009) proposed the use of kernel density estimation (KDE) to interpolate the intensities of variants of a single linguistic feature, and mapping the intensities to a Voronoi diagram to yield an area-class map. Note that this method is one of the mapping techniques that will be used further down in this paper. Rumpf et al. (2010) extended their original work to explore geographical similarities between many individual area-class maps. From each individual area-class map, additional information regarding its structural composition in geographic space is extracted. Cluster analysis is then employed to obtain groupings of structurally similar maps. Wattel and van Reenen (2010) use an interpolation method called splashing on presence/absence data of linguistic variants, which is partially similar to the technique by Rumpf et al. (2009), but is used to interpolate a dense grid (rather than smoothed intensities at observation sites) to show transitions between variants.

It is interesting to note that while researchers in linguistics, in particular those in dialectometry, have used various techniques that represent core methods of GIScience, instances of the analysis and mapping of linguistic data in GIScience are hard to find. A notable exception is the paper by Lee and Kretzschmar (2003) (where Kretzschmar is a linguist) which introduces the point pattern analysis, in particular join-count statistics, used in Kretzschmar (2003). Another instance is the paper by Hoch and Hayes (2010), which advocates the use of GIS in linguistics and provides a review of related research in linguistics, and proposes the use of several GIS techniques (incl. interpolation by kriging and point pattern statistics such as Ripley's K), but does not really present empirical evidence for the proposals made.

Requirements

In our paper, we will focus on methods to map variants of single linguistics features, such as the though/although pair. Furthermore, we will use syntactic data as a basis (i.e. observations about the variation of grammatical features) rather than lexical data (i.e. data about variations in the words used). Syntactic features usually show less variation (i.e. a smaller number of variants) than lexical features. Following are the requirements for mapping methods (cf. also Bucheli Berger, 2008):

1) Must be capable of displaying variation in a single linguistic feature (rather than the co-variation of multiple features).

2) Should be capable of displaying co-occurrences of multiple variants of the same feature at the same geographic location (i.e. co-location).

3) Should be capable of filling in missing values (which often occur, since data are usually only collected at few locations, e.g. a subset of municipalities in a study area). That is, should be able to extra-/interpolate values.

4) Should support the delineation of linguistic boundaries as isoglosses.

5) Should be capable of displaying the gradients of spatial linguistic variation, and infer linguistic boundaries between variants.

6) Should depict the global spatial patterns of linguistic variation as well as the local patterns. These patterns should be easily perceivable, not necessitating cumbersome study of the map and legend.

7) Should allow displaying different variants with different visual weight (e.g. use light map symbols for a commonly occurring variant, and heavy, eye-catching symbols for infrequently occurring variants to better highlight them in the map).

8) If the number of responses per observation site is variable (e.g. 5 responses for one site, and 20 for another), it should be possible to display these quantities, in order to give an impression of the uncertainty involved in the responses (1 out of 5 is less reliable than 4 out of 20, though the percentage is the same).

9) Should be visually attractive.

10) Should include a base map (e.g. political boundaries, hydrography) that ensures spatial reference.

11) Should provide quantitative measures of spatial linguistic variation that may be statistically tested.

Presumably, it will be hard to satisfy all the above requirements equally well with one particular map design, as some of them represent trade-offs. We will now move to present different map design that might be used to meet the above requirements. We will start with the map type that may serve as a baseline — point symbol maps — and then present four alternative designs.

Background and Baseline

The review of related work, as well as the definition of requirements have been approached from a general perspective of mapping linguistic data. However, for the purposes of this paper, we will focus on the study of morphosyntax in dialectology, and we will do so using the example of the Syntactic Atlas of Swiss German Dialects (SADS).

The Syntactic Atlas of Swiss German Dialects (SADS)

The SADS project was initiated in the year 2000 to map and study syntactical phenomena of Swiss German dialects (Bucheli and Glaser, 2002). Close to 3,200 informants participated in the survey. They live in 383 municipalities, which represent approx. 25% of the German speaking municipalities in Switzerland. In other words, for about every fourth municipality, observation data are available, while for the other places no direct testimonies exist. Per observation site, between 3 and 26 informants were involved, with a median value of 5 to 6 informants per municipality. More details about the design of the survey and questionnaires that generated the database for the SADS can be found in Bucheli and Glaser (2002).

Use Case: Infinitival Complementizer

For the purposes of this paper, we will use—for the most part—a syntactical construct that is called 'infinitival complementizer'. For instance, in the English sentence "I don't have enough change *in order to* buy a ticket" the infinitival complementizer is '*in order to*' (or simply '*to*'). It introduces a so-called purposive infinitival clause (since the clause expresses a purpose). In the Standard German equivalent, the infinitival complementizer is '*um … zu*' ("Ich habe zu wenig Kleingeld, *um* ein Billet *zu* lösen"). In Swiss German dialects, two variants of this complementizer exist: *zum* and *für*. (As an aside, the noun 'Billet' is the Swiss variant for English 'ticket' (in 'proper' Standard German this would be 'Fahrkarte'. 'Billet' was used since the questionnaire was used in Switzerland.)

The two variants of this syntactic feature in Swiss German dialects show an East-West distribution, with the *für* variant predominantly occurring in the West, and the *zum* variant in the East. Due to its simple distribution pattern, we use this syntactic feature as our standard use case throughout the paper. Furthermore, and very importantly, linguistic research has also developed hypotheses about the variation of this feature: Seiler (2005), among others, describes the variation of the two variants as two inclined planes, with strike in E-W direction, and dip in easterly direction for *für*, and westerly direction for *zum*. Thus, using this use case we cannot only explore the utility of different map designs, but also apply further geostatistical analysis to test the hypotheses that linguistic research has generated. For this paper, we will focus on the inclined plane hypothesis alone.

The Baseline: Point Symbol Maps

Over the course of the SADS project experiments were made with several different map designs, including point symbol maps in several variations as well as area-class maps (in particular choropleth maps). Following a testing and evaluation phase, the decision was made to use point symbol maps for the production of the atlas (Bucheli Berger, 2008; Bucheli Berger et al., forthcoming). After the production phase of the atlas had started, and after the decision had been made for point symbol maps, a separate project was initiated (Sibler, 2011), which had as its objective to explore different alternative visualization and analysis techniques, some of which will be presented in the following section. This project is not in competition with the SADS atlas production process, and its results will also not directly influence the initial edition of the SADS; bear in mind that the production of such an atlas that comprises many maps, including associated commentaries, is a laborious and complex undertaking. Some of the results, however, might be included in later extensions of the atlas.

We thus use point symbol maps as a baseline to compare against. Figures 1 and 2 show two examples of point symbol maps used in the SADS. Both use base maps that depict the topographic relief, major hydrography, and political boundaries (Swiss *cantons*) to provide a spatial reference. Both also differentiate between single occurrences of a variant in a particular location ("Einzelnennung") and multiple occurrences ("Mehrere Nennungen"). Figure 1 depicts both variants of our infinitival complementizer example (*für* and *zum*), and is restricted to b/w symbology, while Figure 2 is restricted to a single complementizer (*für*), but it uses color and can thus also show the degree of dominance ("mehrheitlich").

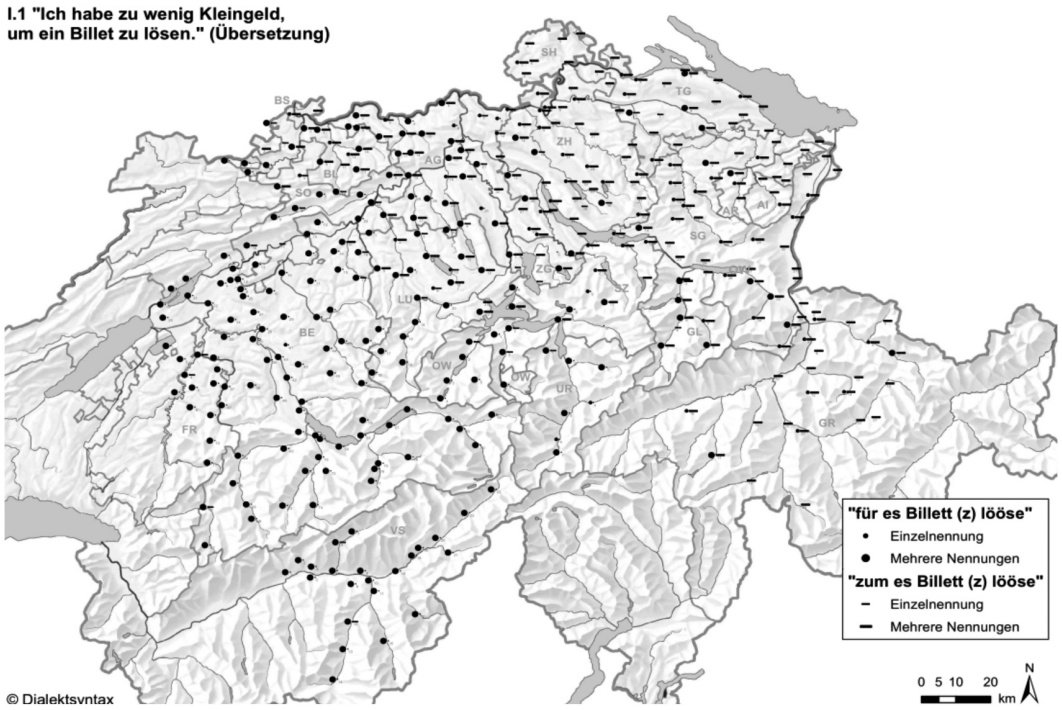

Figure 1: Black-and-white point symbol map in SADS. See text for details.

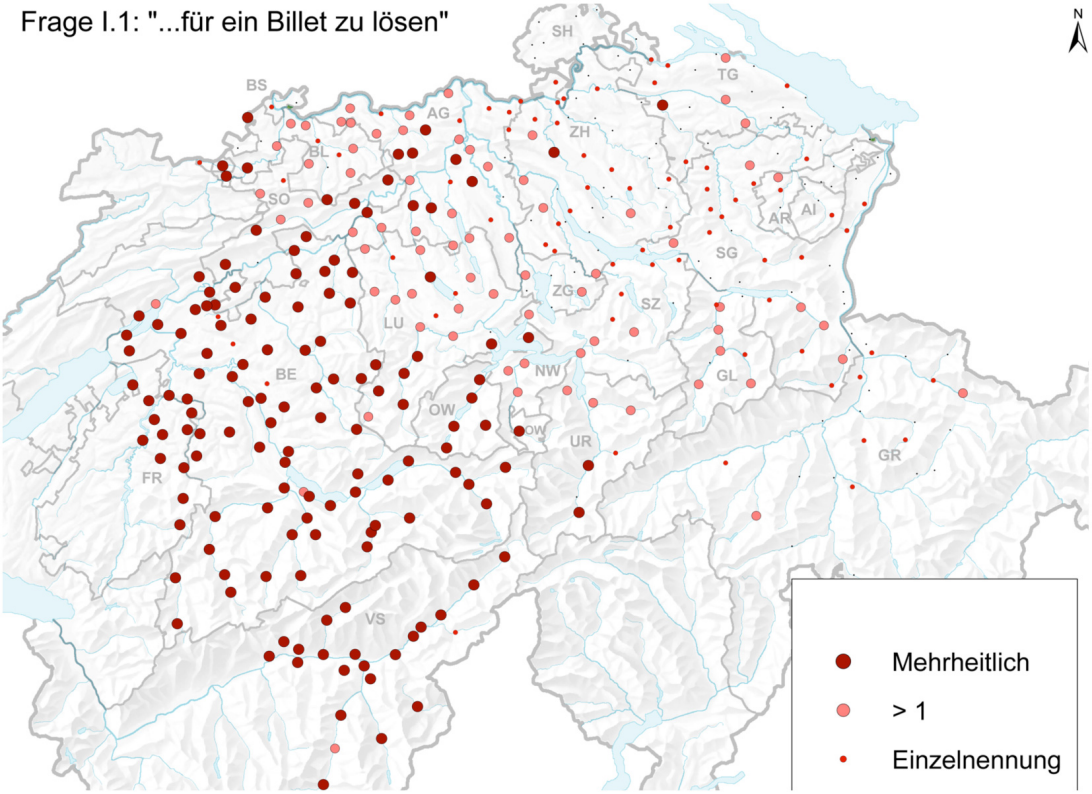

Figure 2: Color point symbol map in SADS. See text for details.

In both Figure 1 and Figure 2, it is possible to see both the global pattern (an E-W trend) as well as local variations and concentrations or outliers. Furthermore, the data that was collected in the linguistic survey is displayed directly, without further modification or processing by some analytical procedure; the map reader is not influenced by the effects of processing imposed by some quantitative method. On the other hand, point symbol maps rely on the visual perception and cognition, intuition, and linguistic expertise of the map reader to infer meta structures implicitly contained in the point data. Boundaries between variants, or language areas, must be inferred visually. To a great extent, this will be alleviated in the published SADS version by the fact that the maps are accompanied by commentaries that offer a description and interpretation of the linguistic phenomena displayed. It is, however, also possible to draw isoglosses on such maps, in order to visually clarify an interpretation or linguistic hypothesis. An example of this is shown with the isoglosses of Figure 3, which have been overlaid on a map from another language atlas, the dialect atlas of Swiss German (SDS; Hotzenköcherle, 1962-1998).

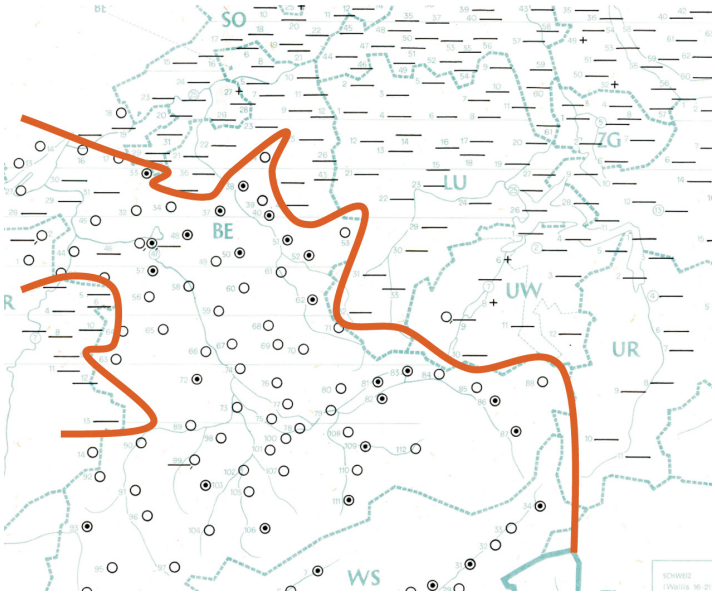

Figure 3: Section from an SDS map with isoglosses overlaid
(base map from Hotzenköcherle et al. 1975, Vol. III: 261).

Alternative Visualization Techniques

We will now present four alternative mapping methods. They have been developed and tested in the course of the project by Sibler (2011). For an evaluation of different types of maps for the purposes of a language atlas, see also Bucheli Berger et al. (forthcoming).

Area-class: Voronoi Polygons

When the aim is to generate area-based maps, two questions need to be answered: What should be the areal unit used? And, which attribute is mapped? To answer the first question, one might simply choose the administrative areal units that are associated with the linguistic observations (e.g. zipcode areas, municipalities). However, as mentioned

above, linguistic surveys rarely ever cover a study area exhaustively. The example of the SADS, where only a quarter of municipalities has been surveyed, is not uncommon; in fact, it even represents an example of dense data coverage. Thus, a map based on administrative units would be quite 'perforated', with most units depicting missing values. The Voronoi diagram offers a way out of this problem. Voronoi (or Thiessen) polygons have advantageous properties that are well-known in GIScience: they generate by definition a complete subdivision of the plane (which removes the above problems of holes), and they create a proximal map about the points used to generate the cells of the Voronoi diagram. In our case, we choose the centroids of the SADS sites (i.e. the municipalities that were surveyed for SADS), as shown in Figure 4.

As an attribute to be mapped, we choose the intensity of the dominant variant. For each site the percentage of occurrence of each variant is calculated. Next, for each site the dominant variant (i.e. the one that yields the highest percentage) is chosen and mapped, while all others are suppressed. Hence, in Figure 4 we see a similar picture as in Figure 1, but now the observations have been extended to areas. Like Figure 1, which shows an overall trend, but which also has outliers, Figure 4 also looks a bit patchy, though the trend between the two variants of the infinitival complementizer is still visible. This type of map follows the choropleth model: a quantitative attribute is mapped to areal units, using lightness variations. Note also that we could have mapped other attributes that can be extracted from the SADS database, such as number of respondents per site (and thus an indicator of reliability). However, with this map type only a single attribute can be displayed at one time.

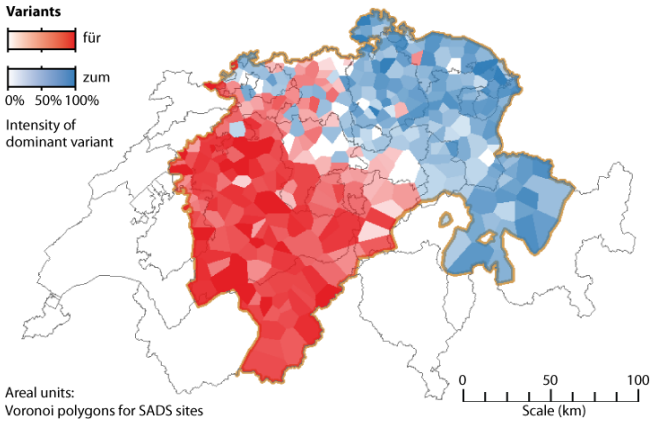

Figure 4: Voronoi polygons for SADS sites. Cells colored according to dominant variant of the infinitival complementizer variants *für* vs. *zum* (Sibler, 2011: 41).

Area-class: Intensities from KDE

The patchiness of maps like the one of Figure 4 may be an effect of short-range spatial variation of a linguistic feature, or it may simply be the effect of unreliable responses that create noise and uncertainty. We now want to retain the choice of areal units (Voronoi polygons for SADS sites) and the choice of mapped quantity (intensities of linguistic variants), but we want to change the way that the intensities are mapped, removing short-range variations. Thus, we would like to replace the original intensity values of Figure 4

with intensities that represent an estimate of a local neighborhood. For this purpose, interpolation methods can be used, but since we are dealing with counts data, the use of a method that can deal with that type of data is warranted. Hence, we use the method by Rumpf et al. (2009), which uses kernel density estimation (KDE) to infer smooth intensity values. In the following examples, we use KDE with a bandwidth of 10 km. This value was established after calibration (details see Sibler, 2011). Figure 5 shows the working principle of the method. For each variant of a feature—in our case the infinitival complementizer variants *für* and *zum*—KDE-based interpolation yields a smooth intensity surface. These separate surfaces are then merged to form one layer, with only the dominant variant remaining per location.

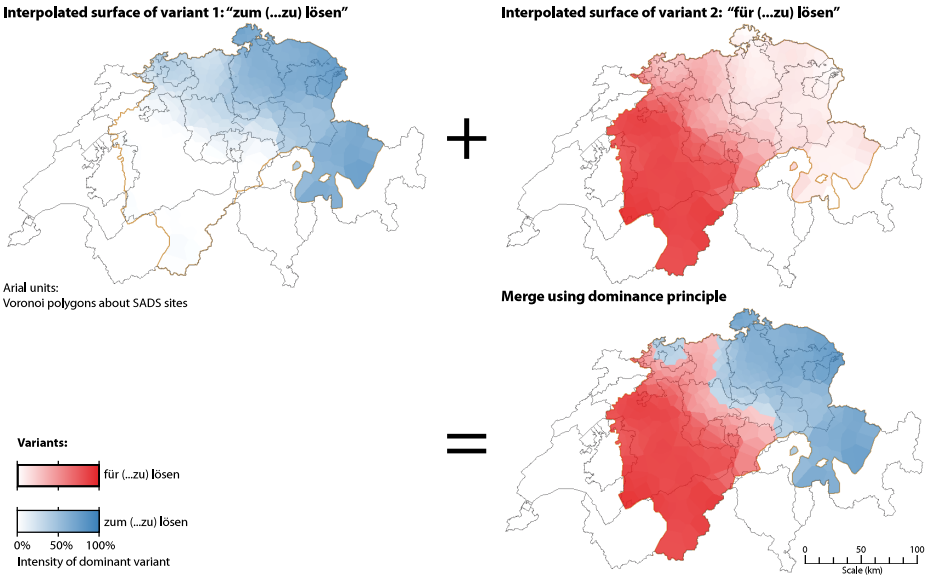

Figure 5: Working principle of KDE-interpolated intensities for dominant variants, on the example of the infinitival complementizer variants *für* and *zum* (Sibler, 2011: 29).

Figure 6 shows how this method cannot only be used to estimate smoothed intensities at the original SADS locations, but also at locations where intensity values are missing (i.e. for the ¾ of municipalities with no SADS observations). The visual comparison of the two maps of Figure 6 suggests that they largely resemble each other. The map with the Voronoi polygons formed for the centroids of the Swiss municipalities is somewhat smoother, but the interpolation is apparently sufficiently robust to yield an almost equivalent result even when four times more interpolation points are used.

Most importantly, however, we can now see much more clearly the East-West trend in the mapped linguistic feature, as it was described by Seiler (2005), compared to the maps of Figures 1 and 4, respectively. In Figure 1, it takes some time (and perhaps also experience) to see the trend emerge from the pattern of points symbols. In Figure 4, the effect is more clearly noticeable, due to the area-class representation. But it is only in Figure 6 where we can clearly see how the red *für* surface and the blue *zum* surface gently slope towards each other, forming high intensities at the eastern and western end, and a transition zone in between. There is only one small blue area in the region around Basel that forms an exception.

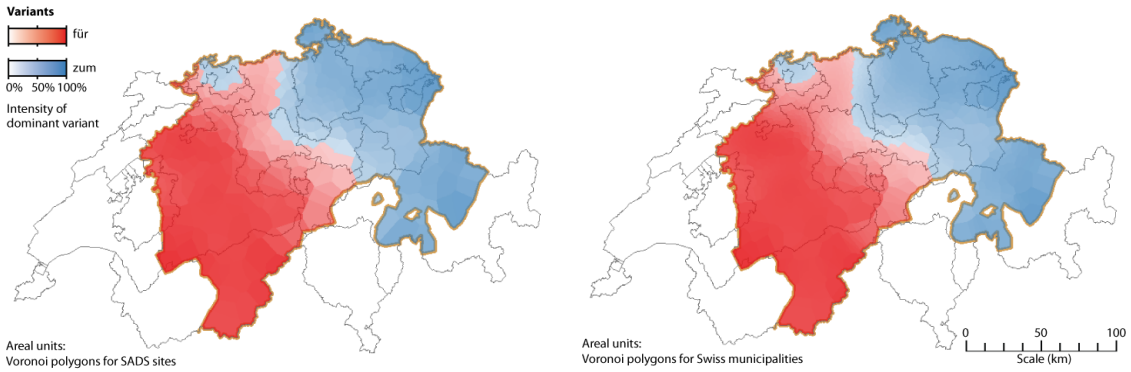

Figure 6: Interpolated intensities from KDE for dominant variants of infinitival complementizers *für* and *zum*, for different areal units. Left: Voronoi polygons for SADS locations. Right: Voronoi polygons for the centroids of Swiss municipalities (Sibler, 2011: 41).

Figure 7 shows an example from another linguistic feature, the position of the indefinite article in conjunction with a complex adjective phrase. The example uses the sentence (in Standard German): 'Susi wäre *eine ganz liebe* Frau für Markus' ('Susi would be *a very nice* woman for Mark'). In Swiss German, this feature yields three variants: 1) preponed, as in Standard German; postponed, as in *ganz ä liebi Frau* ('very a nice woman'); and doubled, as in *e ganz e liebi Frau* ('a very a nice woman'). Figure 7 presents the three variants for this linguistic feature in two ways: using Voronoi polygons without interpolation (as in Fig. 4), and using intensities that have been interpolated using KDE (as in Figures 5 and 6). The difference is very clearly pronounced: while in the un-interpolated Voronoi map, relatively little structure is noticeable, a pattern emerges from the interpolated intensities that forms a corridor of the doubled indefinite article variant (*e ganz e liebi*) from Basel in the Northwest towards the area of Grisons in the Southeast. The preponed variant has disappeared completely (note, however, that even on the simple Voronoi map it only existed in small, isolated pockets). It is interesting to note that previous dialectological research (Richner-Steiner, 2011) had provided hints for the corridor of the doubling variant between Basel and Grisons. Those hints were not strongly pronounced, though: the corridor would only become noticeable if only the places with strong preference for the doubling variant (i.e. > 75% preference) were considered. Thus, it came as a positive surprise when this corridor of the doubling variant was extracted by the KDE-based method. In this case, the dialectometrical approach helped finding a syntactic area where the doubling of the indefinite article is quite common. Conversely, the traditional point symbol maps of the SADS present a rather chaotic distribution that can be interpreted only with difficulty.

While Figure 7 showed an example where KDE-based interpolation helps to reveal a pattern of spatial variation of a linguistic feature, Figure 8 shows an example where KDE interpolation conceals rather than reveals. The example is about the complementizer in comparative clauses, as in the following sentence in Standard German: 'Sie ist grösser *als* ich' ('She is taller *than* me'). In Swiss German, this feature has four variants: *als*, *weder*, *wie*, *wa(n)*. The *als* variant is dominating over the whole area, while the others prevail only in very restricted zones, so that KDE interpolation sweeps them away. So, if there was any variation pattern noticeable in the original data, it has now disappeared completely.

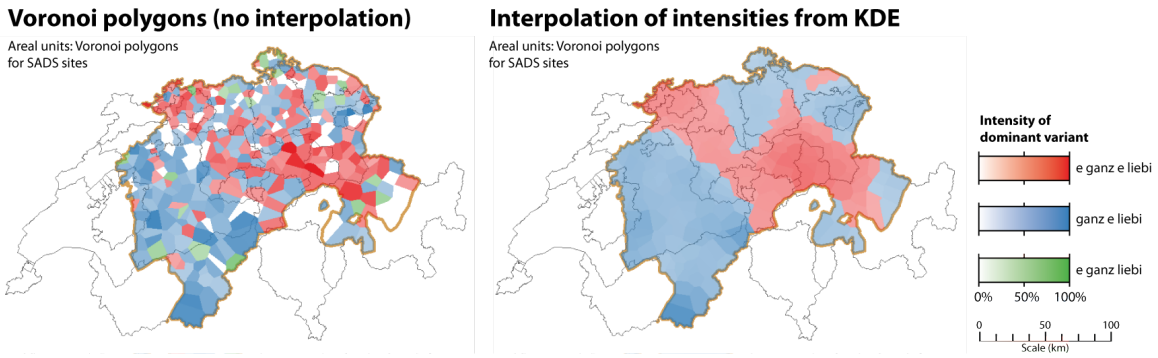

Figure 7: Three variants of the position of indefinite article in conjunction with an adjective. Left: Dominant variants mapped to Voronoi polygons, without interpolation. Right: Interpolated intensity values of dominant variant from KDE (Sibler, 2011: 46).

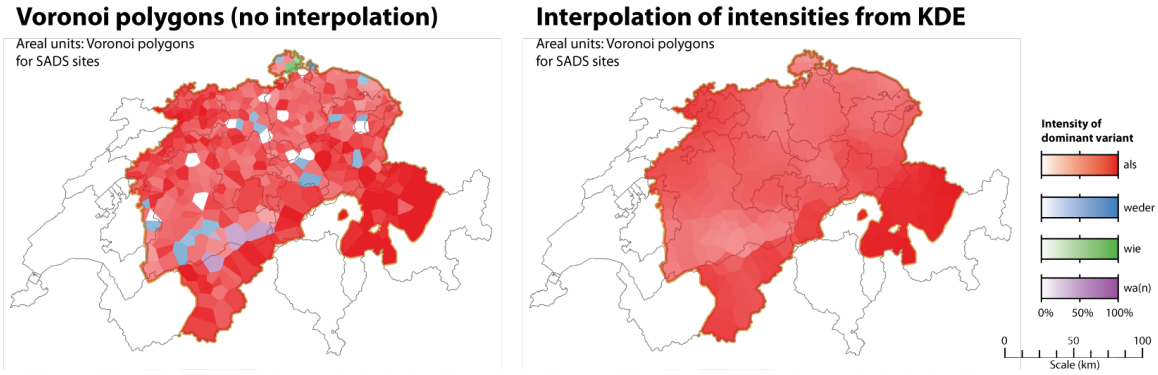

Figure 8: Four variants for complementizer in comparative clauses. Left: Dominant variants mapped to Voronoi polygons, without interpolation. Right: Interpolated intensity values of dominant variant from KDE (Sibler, 2011: 44).

3-D Views

Both preceding map types have the disadvantage that a single attribute (e.g. the dominant variant per location) can be displayed. Variants that are not dominant are concealed, and the interplay of multiple variants is hardly perceivable. A possible way out of this problem is to view the intensities of multiple variants in three dimensions, as it is done in Figure 9 for the well-known infinitival complementizer example. The 2-D Voronoi polygons have been assigned the 'elevation' of the intensity value, colored in according to the usual scheme (red for the *für* variant, blue for *zum*), and projected into 3-D space. If an interactive system is used capable of handling 3-D data (this example was done in a standard commercial GIS software system) then the user can manipulate the 3-D view interactively and explore the data from changing perspective, thus getting a better idea of how the different variants interact, how steep the 'inclined planes (Seiler, 2005) are, etc.

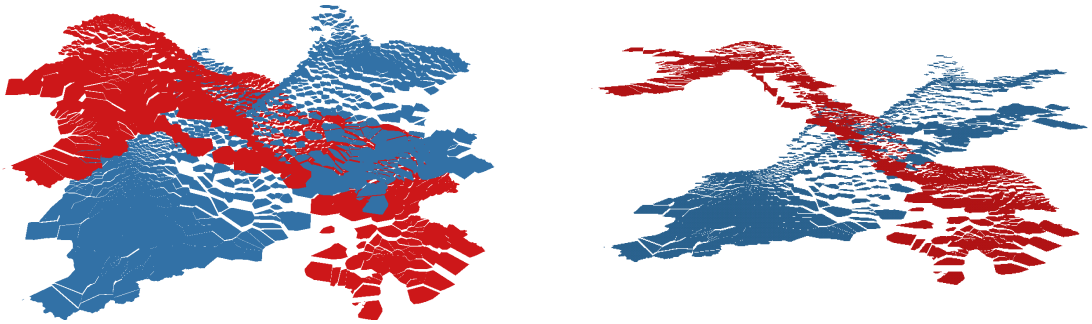

Figure 9: Two 3-D views of infinitival complementizer variants *für* (red) and *zum* (blue) (Sibler, 2011: 51).

Geostatistics to Highlight Specific Patterns

The methods discussed so far all focus on cartographic visualization. However, it should be pointed out that the intensity values—both without and with KDE interpolation—can also be processed further, in an analytical sense. As is well known in GIScience, there are plenty of geostatistics methods available that can be applied to linguistic data. Our linguistic data are originally counts data, which limits the options for geostatistical analysis. However, if these counts are converted to intensities (see above), then many more options become available.

Sibler (2011) conducted several geostatistical analyses. Among others, in order to test the hypothesis by Seiler (2005) who postulated that the two variants of the infinitival complementizer would form two opposing inclined planes, trend surface analysis was carried out. Trend surfaces of first to fourth order were fitted to the unsmoothed intensities and the residuals tested with an F-test, showing a highly significant trends ($p < 0.01$). This analysis also revealed that the general direction of the trend is not from West to East, but rather rotated by 45 degrees, that is SW to NE. For the same intensity data, Moran's *I* (to test for global autocorrelation) as well semivariogram fitting were used, with semivariograms fitted to two stripes of data, one in the direction parallel to the strike of the inclined planes (SW-NE), the other one perpendicular (NW-SE). This revealed a clear direction dependency of the semivariograms, and thus also supports the inclined plane hypothesis.

Another linguistic feature that was further explored and tested was the complementizer in comparative clauses. As Figure 8 showed, the *als* variant is so wide-spread and dominant over the entire area that only in few cases do the other three variants become dominant. As soon as KDE interpolation is applied, the small pockets of dominance of other variants disappear completely. The question, then, is whether below this 'blanket' of the dominant *als* variant the other variants show some distinct patterns. On the point symbol maps for this feature, it is very hard to extract clear patterns; only careful study give hints to that extent. On the area-class maps that focus on the display of dominant variants, it is simply impossible to see anything, due to the overwhelming dominance of *als*. Therefore, the intensity data were subjected to an analysis of local spatial autocorrelation using the Getis-Ord G_i* statistic (Ord and Getis, 1995; Getis, 2010), which has also been used in dialectology by Grieve et al. (2011). The result can be seen in Figure 10. Very clearly, it

paints a much more differentiated picture than the maps of Figure 8 do. In Figure 10, we can see hot spots (clusters of highly positive Z-scores) and cold spots (clusters of highly negative Z-scores) for all variants, except for the *wan* variant, which shows only a hot spot in the Valais / Bernese Alps region, where it is the dominant variant. The *als* variant does no longer appear as overwhelmingly dominant as would seem from Figure 8. The *wie* variant appears with a prominent hot spot along the Northern border of the Swiss border towards Germany. These quantitatively extracted patterns support the qualitative findings of Friedli (2005), who had reached his conclusions based on the study of point symbol maps.

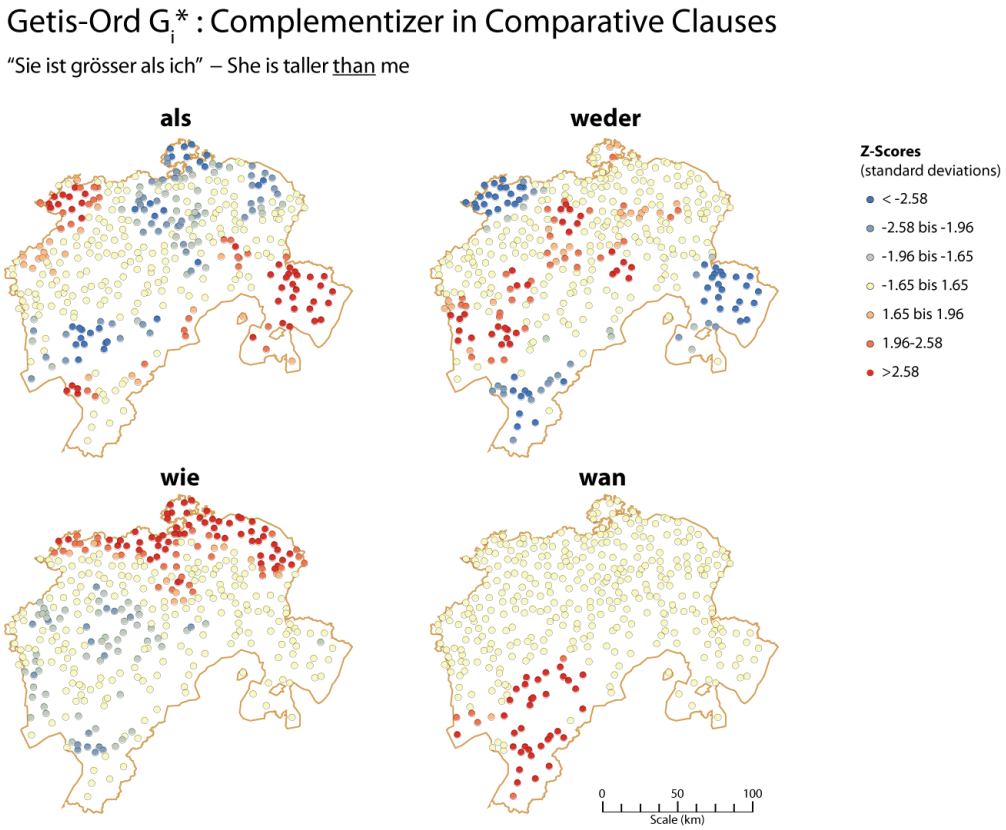

Figure 10: Four variants for complementizer in comparative clauses, displayed with hot/cold spot maps using the Getis-Ord G_i* statistic (Sibler, 2011: 76).

Discussion

We have presented five cartographic visualization methods: 1) point symbol maps (our baseline), 2) area-class maps based on Voronoi polygons, 3) area-class maps using intensities from KDE interpolation, 4) 3-D views, and 5) geostatistical analysis and mapping techniques. In the following discussion we will call these five methods short M1 to M5, in order to save space. Likewise, the requirements will be abridged to R1-R11.

Naturally, other methods could be imagined than the five methods presented here. However, given our experience with language and dialect atlases, we argue that they consti-

tute a representative sample that stands for a broad range of different characteristics of visualization and GIScience methods.

R1 (focus on single features): Since we focus on the study of individual language features (rather than aggregate variation; Nerbonne, 2010), all visualization and analysis methods that assume multivariate input data are automatically ruled out. Conversely, requirement R1 is met by all methods of this paper.

R2 (display co-occurring variants; co-location): M1 obviously can display multiple variants at the same spot, though with increasing number of these variants, and increasing density sample sites, the map will become less and less legible, particularly with respect to R6. On the other hand, M2 and M3, with their focus on dominant variants, do not meet R2. M4 can deal with spatial co-occurrence (i.e. co-location), but similarly to M1 legibility decreases rapidly with increasing number of variants. Like all 3-D displays, M4 also suffers from visibility problems (part of the display may be hidden). Finally, M5 focuses primarily on analysis not visualization and is not capable of dealing with co-location.

R3 (fill in missing values): Both M1 and M2 focus on the 'raw' data and have no interpolation facility; they cannot infer missing values at locations where no data was collected. M3 interpolates intensity values from KDE (possibly at arbitrary locations) and thus meets R3. M4 relies on intensities generated by M3 (but renders them in 3-D) and this meets R3. And some of the M5 geostatistics methods (kriging, trend surfaces, etc could also interpolate values).

R4 (support of isogloss/boundary delineation): M1 supports the delineation of language boundaries graphically. The advantage here is the map reader can visually infer boundaries but is not influenced by the interpretation of an automated, quantitative method. M2 already introduces a model (the Voronoi diagram). Thus, delineation of language boundaries is possible, but will be influenced by the Voronoi structure, which pre-supposes the position of boundaries. M3 clearly helps finding potential language boundaries (sometimes even non-obvious ones, see Figure 7), but it also imposes a computational model that is bound to have an effect on the result (cf. Fig. 8). M4 is less suited for R4, due to visual overlaps and unfamiliar perspective. M5 offers several geostatistical methods that can help in boundary delineation.

R5 (gradient display and boundary inference): The requirement is an extension of R4. In M1, gradients must be 'perceived' visually (like the gradients that 'feel' smoother in Fig. 1 and 2 than in Fig. 3). M2, the Voronoi model assumes a discrete surface and can thus not infer gradients. M3, on the other hand, has excellent capabilities for gradient computation, and based on gradients also boundary inference. Since intensity values are the same in M4 as in M3, and in M5, both M4 and M5 also have gradient mapping and delineation capability.

R6 (easy-to-read depiction of global/local patterns): As mentioned similarly for R4, M1 allows to visually explore patterns. The human visual and cognitive system is capable of seeing highly complex patterns that are hard to describe. On the other hand, visual interpretation is subjective and may not lead to the same result, if different map readers are

given the same task of map reading. M2 is similar regarding fulfillment of R6, though no longer with point data (which might have a perceptual and cognitive effect). M3 in general helps to detect patterns, as through smooth KDE interpolation noise in the input data is suppressed (fig. 7). But it may sometimes also plane away local patterns (cf. Fig. 8). M4 visualizes global trends very well, but local patterns may disappear, not the least due to perspective foreshortening and visibility problems. M5 may support both, the detection of global trends (through Moran's *I*, semivariograms or trend surfaces) as well as local patterns (local point pattern statistics such as G_i*).

R7 (weighting of display with frequency of occurrence): As the symbology of the map in Figures 1 and 2 shows, M1 meets R7. M2 to M4 all rely on the display of dominant variants. Thus it is not possible to visualize directly the frequency of occurrence of the variants. However, it is possible to use the frequencies as weights in the computation of the dominant variant in M2 and the intensities of dominant variants in M3 and M4, respectively. Similarly, M5 can also not directly display count frequencies, but can use them as weights in geostatistical analysis.

R8 (display reliability / uncertainty of observations): The situation for this requirement is similar to that of R7. M1 is the only method that has the potential to construct point symbols that could directly display reliability of observations (e.g. through change of saturation), although no example of this is shown in this paper. M2 to M5, again will have a problem displaying reliability as a separate visual variable (particularly the area-class methods M2 and M3), but in all methods, reliability / uncertainty could be used as a weight in establishing quantitative values such as variant intensities.

R9 (attractiveness of map): All maps seem attractive, each in its own way. M3, for instance, might be particularly attractive in the interactive version, which allows visual interactive exploration. M3 and M5 may be particularly attractive because they very vividly show global vs. local patterns of linguistic variation.

R10 (include base map): All examples of M1 (Fig. 1 to Fig. 3) include base maps. While base maps would be theoretically possible (and advisable!) for all methods, only point symbols can be easily combined with base maps, while it is definitely more demanding to overlay linear or areal data on a base map. The geostatical map (M5) of Figure 10 is essentially also a point symbol map. Hence, the same applies as to M1.

R11 (provide quantitative measures for statistical testing): M1 to M4 focus on visualization, and hence are useful primarily for visual, exploratory analysis rather than confirmatory, statistical analysis. Since M3 and M4 use KDE, which generates smooth gradients, fitting calculating derivatives of the intensity surface (e.g. gradient, curvature) could be envisioned. M5 is built for analysis: maps are only a by-product of geostatical analysis. Hence, M5 (geostatistics) really represents the culmination in an analytical sense: After visualization of the type of M1 to M4 have been used, and hypotheses formulated, analysis methods from geostatistics and statistics are used to falsify/verify these hypotheses. Further visualizations such as the one shown in Figure 10 may then be shown to further explore and communicate the results of geostatistical analysis.

Conclusions

Starting off from the identification of the peculiarities of linguistic data we have defined an extensive set of requirements, which visualizations of linguistic data should meet (with the constraint, though, that they are used to map variations of single linguistic features). We have then presented five different cartographic visualization techniques. The five methods were chosen so they jointly span a broad scope of visualization options, and so they can be used to explore linguistic data in multiple ways, hypothesize about spatial patterns of linguistic variation, and finally test and confirm such linguistic hypotheses. On the basis of data from the Syntactical Atlas of Swiss German Dialects (SADS) we then demonstrated these visualization techniques, which provided the basis for an in-depth evaluation and discussion of the methods regarding the requirements defined initially.

Not surprisingly, none of the presented techniques meets all of our requirements (which partially represent trade-offs). However, each of them has its distinctive strengths, so that in combination of several visualizations, optimal results should be achieved. We used *point symbol maps as* a baseline, since they are used in the current production of the SADS publication, which is nearing completion. The key advantage is that the original observations can be mapped without further processing and alteration by some statistical procedure; the map image is not biased by the potential effects imposed by some quantitative method (e.g. the excessive smoothing effect visible in Figure 8). Also, point symbols can be designed and configured in almost unlimited ways, and they can be easily combined with other map information (e.g. base map). *Voronoi polygons* of dominant variants provide an easy way to generate area-class maps and to get a first picture of the areal coverage of a linguistic feature. As in points symbol maps, no further processing takes place. On the other hand, they are susceptible to noise in the original data (again like point symbol maps). Area-class maps that are based on *intensities from KDE* introduce a smooth interpolation function that allows extending local trends across neighborhoods, panes away small outliers and noise, and particularly brings out the big picture. Also, it offers a method to interpolate values where they are missing. *3-D views* are based on the intensities calculated by the previous technique, but they eliminate the problem of the 2-D techniques that the only dominant variants are rendered. In 3-D trends and breaks in the spatial variation of a linguistic feature, as well as the interplay of different variants, become better visible and can be explored interactively. Finally, *geostatistical analysis and mapping techniques* represent a whole family of methods that can be used to analyze and later visualize specific patterns of linguistic variation in geographic space on the global and local level.

References

Bucheli, C. and Glaser, E. (2002) The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems. In: Barbiers, S., Cornips, L. and van der Kleij, S. (eds.) *Syntactic Microvariation, Vol. 2*. Amsterdam: Meertens Institute Electronic Publications in Linguistics, pp. 41-73

Bucheli Berger, Claudia (2008) "Neue Technik, alte Probleme: auf dem Weg zum Syntaktischen Atlas der Deutschen Schweiz (SADS)". In St. Elspass, W. König (eds.) Sprachgeographie digital – die neue Generation der Sprachatlanten. Hildesheim: Olms (= Germanistische Linguistik 190-191). 29-44.

Bucheli Berger, Claudia, Elvira Glaser and Guido Seiler (forthcoming) "Is a syntactic dialectology possible? Contributions from Swiss German". In A. Ender, A. Leemann & B. Wälchli (eds.) Methods in Contemporary Linguistics. Accepted for publication in the Trends in Linguistics Series. Berlin: de Gruyter Mouton.

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

Friedli, M. (2005) Si isch grösser weder ig! Zum Komparativanschluss im Schweizerdeutschen. In: Christen, H. (ed.) Dialektologie an der Jahrtausendwende. *Linguistik Online*, Vol. 24, No. 3. Available at http://www.linguistik-online.de/24_05/friedli.html (last accessed on 15 August 2012).

Getis, A. (2010) Spatial Autocorrelation. In: Fischer, M. M. and Getis, A. (ed.) Handbook of Applied Spatial Analysis. Berlin/Heidelberg/New York: Springer.

Goebl, H. (2010) *Language and Space*, *Vol. 2: Language and Mapping*. Berlin: De Gruyter Mouton, pp. 433-457.

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

Haag, K. (1898) Die Mundarten des oberen Neckar- und Donaulandes. Reutlingen: Buchdruckerei Egon Hutzler.

Hotzenköcherle, R. et al. (eds.) (1962-1998) *Sprachatlas der Deutschen Schweiz*. Volumes I-VIII. Tübingen: Francke.

Hotzenköcherle, R. et al. (eds.) (1975) *Sprachatlas der Deutschen Schweiz*. Volume III. Tübingen: Francke.

Hoch S. & Hayes J. J. (2010) Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin*, 51, 1, pp. 23-36.

Lameli, A., Kehrein, R. and Rabanus, S. (2010) *Language and Space. Vol. 2: Language Mapping*. Berlin: De Gruyter Mouton.

Lee, J. & Kretzschmar, Jr., W. (1993) Spatial analysis of linguistic data with GIS functions. *Int. Journal of Geographical Information Systems*, 7, 6, pp. 541-560.

McDavid, R.I., Jr. and O'Chain, R. (1980) *Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Nerbonne, J. (2009) Data-driven Dialectology. *Language and Linguistics Compass*, 3, 1, pp. 175-198.

Nerbonne J. (2010) *Language and Space. Vol. 2: Language Mapping*. Berlin: De Gruyter Mouton. pp. 476-495.

Ord, J. K. and Getis, A. (1995) Local Spatial Autocorrelation Statistics. Distributional Issues and an Application. *Geographical Analysis*, 27, pp. 286-306.

Pi, C.-Y. (2006) Beyond the Isogloss: Isographs in Dialect Topography. *Canadian Journal of Linguistics*, 51, pp. 177-184

Richner-Steiner, Janine (2011) »*E ganz e liebi Frau« – Zu den Stellungsvarianten in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen*. Eine dialektologische Untersuchung mit quantitativ-geographischem Fokus. PhD Dissertation, German Department, University of Zurich.

Rumpf, J. Pickl, S. Elspass, S. König, W. and Schmidt, V. (2009) Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, 76, 3, pp. 280-308.

Rumpf, J. Pickl, S. Elspass, S. König, W. & Schmidt, V. (2010) Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-class Maps. *Dialectologia et Geolinguistica*, 18, pp. 73-98.

Seiler, G. (2005) *Moderne Dialekte – neue Dialektologie*. Stuttgart: Steiner. pp. 313-341.

Sibler, P. (2011) *Visualization and Geostatistical Analysis with Data of the Syntactic Atlas of Swiss German Dialects (SADS)*, Master's Thesis , Zurich: Department of Geography, University of Zurich.

Smith, B. and Varzi, A. (2000) Fiat and bona fide boundaries. *Philosophy and Phenomenological Research*, 60, pp. 401-420.

Wattel, E. and van Reenen, P. (2010) *Language and Space. Vol. 2: Language Mapping*. Berlin: De Gruyter Mouton, pp. 495-505.

Pius Sibler, Application Engineer, Intergraph (Switzerland) AG, Neumattstrasse 24, 8953 Dietikon (Switzerland). Email <pius.sibler@intergraph.com>

Robert Weibel, Professor, Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich (Switzerland). Email <robert.weibel@geo.uzh.ch>

Elvira Glaser, Professor, German Department, University of Zurich, Schönberggasse 9, 8001 Zurich (Switzerland). Email <eglaser@ds.uzh.ch>

Gabriela Bart, Research Assistant, German Department, University of Zurich, Schönberggasse 9, 8001 Zurich (Switzerland). Email <gabriela.bart@ds.uzh.ch>