Measuring boundaries in the dialect continuum

Péter Jeszenszky
University of Zurich /
Department of Geography
Winterthurerstrasse 190
Zürich, Switzerland
peter.jeszenszky@geo.uzh.ch

Robert Weibel
University of Zurich /
Department of Geography
Winterthurerstrasse 190
Zürich, Switzerland
robert.weibel@geo.uzh.ch

Abstract

GIScience has only rarely dealt with linguistic data so far despite its challenging nature with many peculiarities that make analysing spatial language variation a worthwhile endeavour. The two commonly used paradigms in dialectology to deal with dialectal areas in space, the dialect continuum and isoglosses, respectively, correspond to the dichotomy of fields vs. entities found in GIScience. These two ways of conceptualising dialectal boundaries engrain the problems that the quantitative analysis of delineations of language areas is facing. We present initial steps of a project that aims at quantitatively modelling language area formation and the influences of geographic factors on boundaries between dialect areas. We start by analysing the distinctive features of language data that set them apart from other types of data commonly dealt with in GIScience. We then phrase the key questions that guide the analysis of dialectal boundaries, and we propose a range of GIScience methods that can be used to answer these questions. We also present preliminary results of applying some of the proposed methods on Swiss German syntax data.

Keywords: dialectology; dialectometry; linguistic data; boundary statistics; fuzzy boundaries

1 Introduction

Studying the formation of language areas is one of the main concerns of dialectology. Linguists have traditionally used qualitative, descriptive methods for that purpose [1]. The use of quantitative methods, in particular the use of genuinely geospatial analysis techniques, is rather recent and still relatively rare. Furthermore, little work has been spent on quantifying the delineations of language areas themselves. Looking from the perspective of GIScience, we find that linguistic problems have not found a great deal of attention so far, even though the peculiarities found in linguistic data would make it an interesting challenge, and despite the fact that GIScience has a toolbox of methods available that could make potentially valuable contributions to linguistic research. In this paper, therefore, we set out to explore GIScience methods to assess boundary delineation within dialectology.

A spatial boundary (we will use the words 'boundary' and 'border' synonymously) within linguistics is a hard-to-grasp concept owing to the abundant uncertainties inherent to language data. The two key paradigms of dialectology to conceptualize dialectal boundaries are the *isogloss* and the *dialect continuum* [7], corresponding to the dichotomy of entities and fields, respectively, in GIScience. Isoglosses are theoretical lines delineating and separating occurrences of different variants used for a linguistic phenomenon, while the theory of dialect continua states that the change in dialectal spatial variation, be it a single phenomenon or aggregate variation, is gradual [10].

In our investigation we focus on individual syntactic phenomena occurring in dialect surveys, attempting to quantify the fuzziness and stability of dialectal boundaries. It is unique in the sense that the survey providing the data has multiple respondents per survey site. We provide a sensitivity analysis to assess how robust the boundaries of dialects are on examples of syntactic phenomena.

The main contribution of this paper is to describe the problems of intralinguistic fuzzy boundaries and offer measures that could be taken to solve them.

In the following:

- we describe why dialectological data is special;
- we propose GIScience methods addressing the problems related to boundaries in dialect continua;
- we present preliminary results with selected methods using Swiss German dialectal data

On Boundaries in Linguistics

Linguists have always been interested in studying the variation of languages over space, and to delineate linguistic areas and dialects. The theory of dialect continua is one of the most popular topics of variationist linguistics these days [10]. Dialectometry is dealing with discovering and measuring structures in spatial networks of dialects [3,5,11]. Quantification of language usage is thereby a natural need and thus connects spatial linguistics to other quantitative sciences.

In the past linguists were trying to find so called isogloss bundles using which they could delineate distinct dialect areas (e.g. [6]) and formulate further hypotheses. The theory of dialect continua (e.g. [7]) was introduced later. It has long been researched that speech variation mostly changes continuously rather than having geographically abrupt breaks [3] although to various linguistic phenomena certain physical boundaries may mean an abrupt change as well.

Ontological studies on boundaries in general do not leave any doubt about the linguistic boundaries being artificial (*fiat*) [13], implying their definition is always connected to scale. The *fiat* nature of linguistic boundaries means that they are always hard to grasp and will need to be defined by some decision. The most basic such line is the isogloss on the level of a single phenomenon.

Grieve, Spellmann & Geraerts [5] used three statistical techniques to delineate dialectal areas: spatial autocorrelation, factor analysis, and cluster analysis, which they assigned to the concepts of isoglosses, isogloss bundling and the analysis of relationship between the various bundles of isoglosses. Doing this they wanted to link the isogloss theory to the more realistic dialect continua.

In reality, single linguistic phenomena analyzed do rarely display the type of clear-cut regional patterns that are often exhibited in traditional dialect classification studies. Quantifying differences of dispersions can be approached from several directions, as shown further on. For instance, *homogeneity* of an area-class dialect map was used by [11] to quantify the distribution of certain dialectal phenomena along with the total length of boundaries formed by dominant variants, as a measure of complexity of the map.

2 Analysis of the Problem

2.1 Characteristics of linguistic data

As mentioned above, we can conceptualize boundaries in the dialect continuum in two forms, either as entities (*isoglosses*) or as fields (*dialect continuum*). A boundary (in space) is a linear phenomenon where a given property or variable is changing. We can approach the boundary problem in linguistics from two sides, the boundary being a geometric object or being a gradient. If we choose to place a crisp, entity-like boundary, we will do so by discretising the continuum of spatial language variation, e.g. by classifying according to a certain dominance threshold of one or more dialect variants. If we don't discretise, we can regard 'boundaries' (or rather, transition zones) as gradients, with steep gradients implying a stronger and less fuzzy boundary.

Due to their human nature linguistic data are burdened with different kinds of peculiarities and uncertainties that are unlike those that GIScientists normally encounter. For instance, linguistic data differs greatly from other spatially sampled data that is used for detecting boundaries. For example in soil science variables (e.g. soil pH) are single-valued, with only one value per survey site. Generally physical variables can be measured on a numerical scale and can be easily interpolated using physical laws. Since many of the linguistic variables have nominal scale (e.g. in syntax), and since the variation is not governed solely by physical processes, *interpolation* is challenging.

Linguistic data have different *scales of measurement*, the aforementioned syntactic level is nominal in most cases, while the phonological level can be turned into interval data in order to calculate Levenshtein-distances between transcriptions of pronunciations which are in turn burdened with subjectivity. The *representativeness* of the data can be questioned as long as linguistic surveys have only one or a few meticulously chosen respondents per survey site, thus possibly artificially reducing linguistic variation. In modern dialect surveys, it is common to use multiple respondents per site. As a consequence, *co-occurrence* of different variants per site is commonplace. This heterogeneity of linguistic data may be further confounded by other sources of *uncertainty*, such as differences in phonetic transcription, semantic issues etc. Other challenges are related to the *sampling scheme* used (e.g. number and quality of answers per survey site, distribution of data points, type of collection method). One way to deal with this uncertainty is to "aggregate the differences in many linguistic variables in order to strengthen their signals" [10]. A further strategy that has recently found increasing attention is to use large text corpora, such as those provided by social media (e.g. [2]), aiming to overcome the limitations associated with traditional language atlases.

2.2 Requirements and research questions

In order to assess boundaries in a dialectal space, we need to define what we mean by a 'crisp' boundary and a transition zone, and define requirements for any methods that could be used to quantitatively determine these concepts. As a general requirement, we have a need for testing the *statistical significance* [8]. Furthermore, all methods are *scale-sensitive*, as we need the define thresholds above which we consider something a boundary. If we approach linguistic variation as a continuum, we can characterize boundaries as changes in gradient. Estimating the steepness of a gradient, however, is a matter of scale, as is well-known from surface analysis: we have to define an analysis window, and that will also affect the resulting gradient values. The threshold might be higher when we look at a boundary between two adjacent survey sites and lower when we consider boundaries at the global level. The crispness of a boundary shouldn't depend on how many survey sites bear the given variant, only the *homogeneity* of its cluster should count respective to the scale of analysis. (For example a boundary between an area with 100 % and 0 % of variant usage should be crisper than between 75 % and 25 %).

The research questions considered in our study are:

A) Can we find dialectal boundaries considered as 'crisp'? How can this crispness (or conversely, fuzziness) be assessed?

B) How robust are these boundaries?

C) How appropriately placed (i.e. meaningful in subdividing the geographic space for the given variants) are linguistic boundaries, i.e. isoglosses?

D) Do they correspond to geographic boundaries?

2.3 Data

We use the Syntactic Atlas of German-speaking Switzerland (SADS; [1]). This database is unusual among linguistic surveys because multiple respondents exist per survey site, and a respondent is even allowed to use different variants of a dialect phenomenon (or variable). Between 2000 and 2002 close to 3,200 respondents participated in a series of four surveys in 383 survey sites (i.e. one quarter of Swiss German municipalities), responding to questions about syntactic variables. Having multiple respondents (3-26 with a median of 7) per survey site gives us the chance to better grasp the linguistic diversity that is present within a settlement, thereby being able to see how gradual the change is from a dominance area of one variant to the other. The data sample in our case is

dense enough to test whether the spatial change is abrupt (isogloss-like) or rather a transition zone and gives us the chance to quantify the given transition. To estimate values between survey sites we used Voronoi-polygons as a tessellation, which is a common method in dialectometry to interpolate between survey sites in area class maps [6,7,9,10,11,12].

3 Methods

When proposing methods we remain testing them on the level of single phenomena for the time being. Some of these variables were investigated on an individual level by Sibler et al. [12] and their patterns discussed in relation to geographic distance variation on an aggregate level by Jeszenszky & Weibel [9].

A) Responding to research question A we propose multiple methods to assess the *crispness* (or *fuzziness*, conversely) of the boundary between two variants. Mapping the intensity values (the proportions of the most dominant variant) at each survey site can yield transition zones. The "width" such transition zones is an indicator of the gradient between two variants, while the "depth" (i.e. degree of variation) tells us something about the relation of the dispersions. The smaller the intensity of the dominant variant, the more probable that other variants have a large share too (Fig. 1 and 2). Trend surface analysis [12] can be employed for estimating the gradient across the transition zone, and for analysing the variation of residuals.

Also the crispness of a transition from dominance of one variant to another can be evaluated by taking cross-sections and plotting the intensity of each variant at the survey sites along this line (Fig. 3). Regression analysis can be used to further analyse gradients and residuals.

B) To test the *robustness* or *significance* of boundaries testing procedures based on Monte Carlo simulation can be used such as described in [8], randomizing the underlying data to a certain degree and assess how much it moves the position of the boundary under evaluation from the original state.

C) To assess the meaningfulness of boundaries we propose a *homogeneity* measure in which we use externally sourced, geographic boundaries dividing the survey sites into two groups. 'Geographic boundaries' are those that may act as potential barriers to language contact, including political or administrative boundaries, religious borders (e.g. between protestant and catholic areas), natural borders such as waterbodies or topographic ridge lines, etc. Note that our homogeneity definition is different from the homogeneity measure of [11], which does not take into account geographic boundaries. We then measure what proportion of the variants occur on either side of a boundary. If we repeat this with multiple geographic boundaries, we can determine the best fitting boundary, i.e. the one that keeps both sides most homogeneous, having the highest possible number of respondents of one variant on one side while having the lowest possible number of respondents of the other variant(s) on the other side. Geographic subdivisions can then also be tested against the most dominant border between two variants, that is, the isogloss. This also leads us to research question D. Additionally, using a smoothing function such as kernel density estimation [11,12] suppresses infrequently occurring variants and thus accentuates differences.

D) *Correspondence* of dialectal boundaries with boundaries of geographic importance can be tested, such as the ones mentioned above, in order to assess to what extent geographic factors might influence linguistic variation. Possible methods for this purpose include overlap statistics [8], intersection with nested buffer zones, line density computation of linguistic borders (which may give an indication of isogloss bundling), and earth mover's distance (EMD) [4].

4 Experiments

To illustrate the above methods, we present some results of preliminary experiments for research questions A to C, which will be continued in future work.

Figure 1: Example of an intensity map, where the proportion of the most dominant variant (ranging from 38 % to 100 %) for *Word order in causative* phenomenon is mapped.

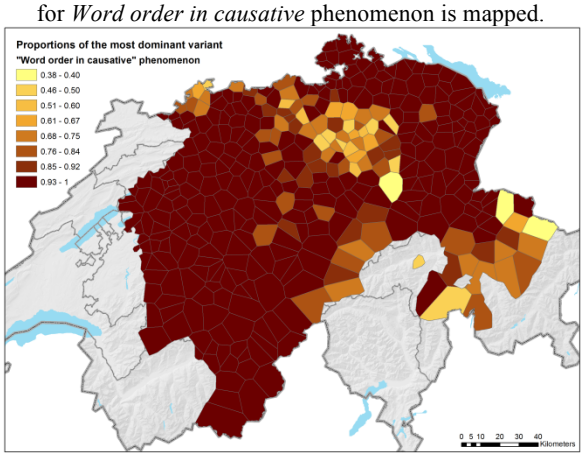
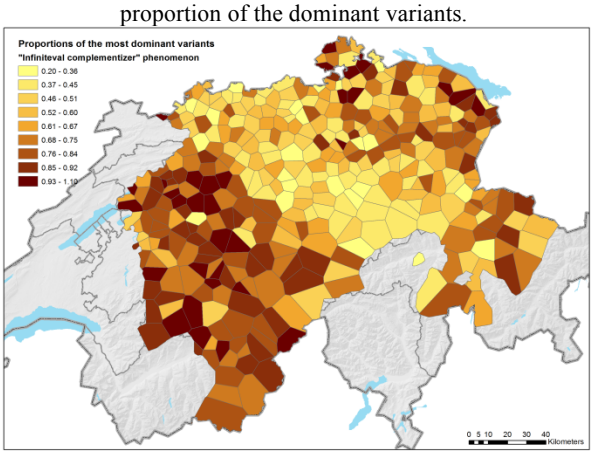


Figure 2: Proportions of the most dominant variant for *Infiniviteval complementizer*. Darker brown means higher proportion of the dominant variants.

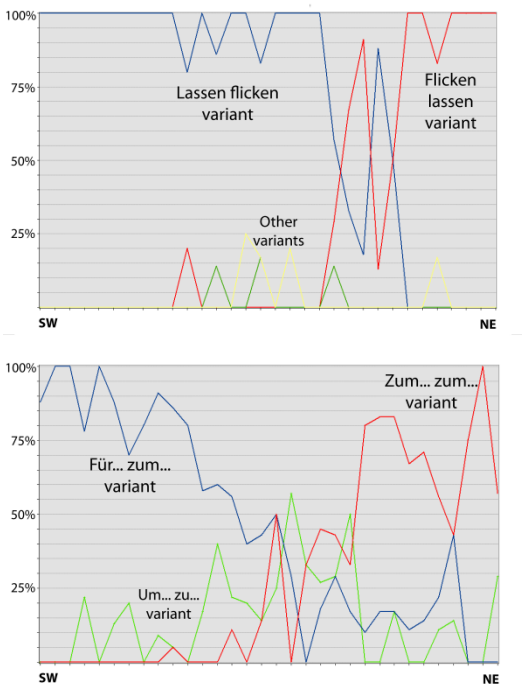


Above, we posited that we could find transition zones between two main areas of variants by mapping the maximum intensities present at each survey site (the proportion of the most dominant variant), as shown in Fig. 1 and 2. In Fig. 1 we

can only see a small area where the intensity of the most dominant variant is not close to maximum, meaning that the transition is relatively abrupt from one dominant variant to the other. In Fig. 2, on the other hand, a fuzzier NW-SE transition zone surfaces where the majority of the polygons receive low values. That means that the transition between dominant variant areas is quite gradual (with a low gradient), or even a third potential variant is in play (as it is the case here).

We constructed section profiles along a SW-NE line for the linguistic phenomena depicted in Fig. 1 and 2, as shown in Fig. 3. The upper graph (*word order in causative*) depicts a steep gradient from one dominant variant to the other, and the two main variants exist almost exclusively in their respective areas. On the other hand, the lower graph (*infinitival complementizer*) shows a gradual transition from one dominant variant to the other, with a presence of a third variant that becomes dominant at points.

Figure 3: The intensity profiles for *Word order in causative* and *Infiniviteval complementizer* respectively. The position of the cross-section line is shown in the lower maps of Figure 4.



We conducted a *sensitivity* test where we changed 20 % of the answers at each survey site randomly, which models asking 20 % new correspondents while discarding 20 % existing ones (Fig. 4). The lower row features the original raw data maps where the colour hue represents the different dominant variants, while the colour intensity represents the proportion of the given variant. The right-hand columns, featuring data from *Word order in causative*, shows almost no change in dominance in the upper map, while the left column featuring *Infiniviteval complementizer*, shows more change, with the green and blue variant gaining more polygons. This indicates that if an isogloss border was placed, it would be of a lower degree of robustness / significance.

Finally, for the syntax variable *infinitival complementizer* we took variant proportions on two sides of an arbitrary boundary. This allows quantifying how well a boundary delineates two areas based on the variants' proportions, and thereby quantifies the areas' homogeneity (Table 1). 85 % of the "Für…zum…" variant's respondents are contained in Area 1 while the "Zum… zum…" variant hits 88 % in Area 2. On the other hand the third most important variant "Um… zu…" and the aggregate of other variants are contained about 50:50 in the two areas which suggests they are more randomly distributed than the main variants.

References

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

[2] J. Eisenstein. Identifying regional dialects in online social media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *Handbook of Dialectology* (Vol. 2013, pp. 1–15). Wiley, 2015.

[3] C. Gooskens. Norwegian Dialect Distances Geographically Explained. In *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE Vol. 2.*, pages 195–206. Uppsala, 2004.

[4] K. Grauman & T. Darrel. Fast contour matching using approximate earth mover's distance. Proceedings IEEE In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington DC, 2004.

[5] J. Grieve, D. Speelman, & D. Geeraerts. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(02), pages 193–221. 2011.

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

[7] W. Heeringa & J. Nerbonne. Dialect Areas and Dialect Continua. *Language Variation and Change*, 13(03), pages 375–400. 2001.

[8] G. M. Jacquez, S. Maruca & M. Fortin. From fields to objects: A review of geographic boundary analysis. *Journal of Geographical Systems*, 2, pages 221–241. 2000.

[9] P. Jeszenszky & R. Weibel. Correlating morphosyntactic dialect variation with geographic distance: Local beats global. In *Extended Abstract Proceedings of the GIScience 2014*, pages 186–191. Vienna, 2014.

[10] J. Nerbonne. Mapping Aggregate Variation. In *Language and Space. An international Handbook of Linguistic Variation. Vol 1. Theories and Methods*, pages 476 – 495. Mouton de Gruyter, Berlin/New York, 2010.

[11] J. Rumpf, S. Pickl, S. Elspaß, W. König, & V. Schmidt. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, 76(3), 280–308, 2009.

[12] P. Sibler, E. Weibel, E. Glaser & G. Bart. Cartographic Visualization in Support of Dialectology. In *Proceedings - AutoCarto 2012* - Columbus, Ohio, 2012.

[13] B. Smith & A. C. Varzi. Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research*, 60(2), pages 401–420. 2000.

Figure 4: Sensitivity analysis that models surveying 20 % new respondents at each survey site. Original values in the lower row, changes mapped in the upper row. *Infiniviteval complementizer* in the left hand column, *Word order in causative* in the right hand column.

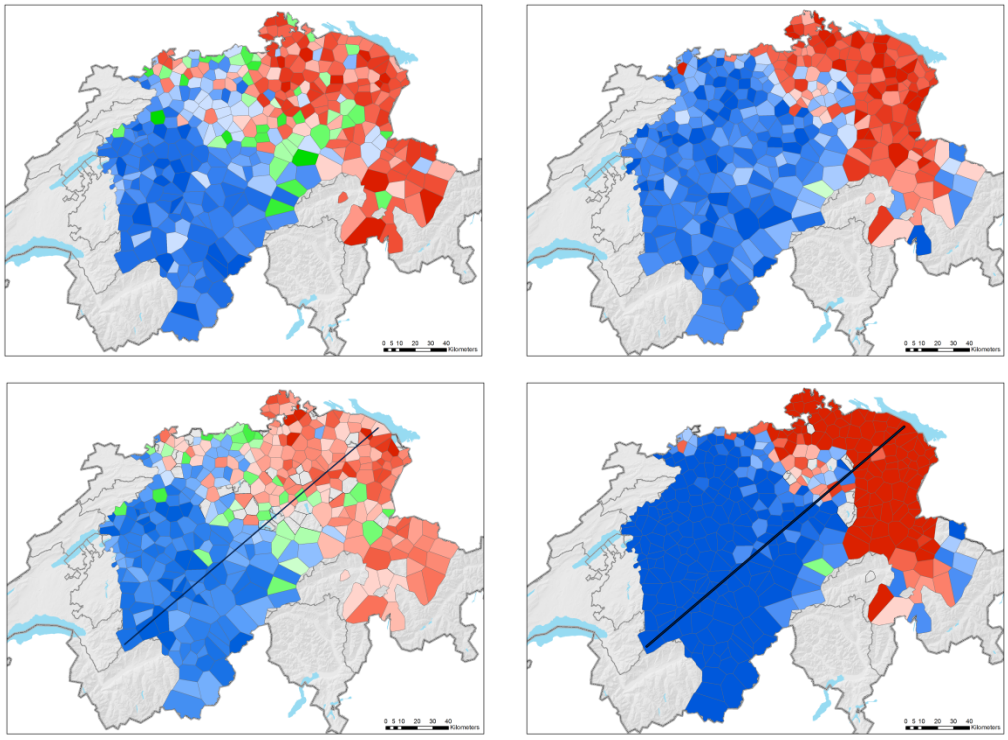


Table 1. Homogeneity test for the dominance areas of the aggregate main variants for a survey question (*Infiniteval complementizer*), to assess what proportion of the respondents of respective answers is included in the dominance area.

NUMBER OF RESPONDENTS	**Für... variants**	**Zum… variants**	**Um… variants**	**Other variants**
Area 1	978	116	218	178
Area 2	166	841	254	212
SUM	1144	957	472	390
PROPORTIONS				
Area 1	85 %	12 %	46 %	45 %
Area 2	15 %	88 %	54 %	55 %

Area 1 = region where *Für… zum…* variant is dominant, Area 2 = region where *Zum… zum…* variant is dominant