A measure for linguistic coherence in spatial language variation

Anonymous ACL submission

Abstract

Based on historical dialect data we introduce a local measure of linguistic coherence in spatial language variation aiming at the identification of regions which are particularly sensitive to language variation and change. Besides, we use a measure of global coherence for the automated detection of linguistic items (e.g., sounds or morphemes) with higher or lesser language variation. The paper describes both the data and the method and provides analyses examples.

1 Introduction

Dialectometric work typically focuses on the co-occurrence of the distribution of variants in different sites (see Goebl 1984). From these co-occurrences, reasonably coherent regions of linguistic similarity can be identified. These regions then provide, for example, clues to the aggregated structuring of higher-level linguistic areas (e.g., within a nation). Alternatively, they show to what extent individual sites of a given corpus are integrated into the region under discussion in terms of their similarity or distance to other sites (e.g., Heeringa 2003). Such analyses, which at the same time constitute the classical field of dialectometry, thus benefit from the aggregation of all linguistic phenomena of a given corpus.

However, if the interest is not in the overall structuring of a region, but in the distribution patterns of individual variants, non-aggregating procedures must be applied. For a single phenomenon, spots of variation may be identified in most cases by visual inspection (see Ormeling 2010 for a critical account). However, in order to capture this variation quantitatively, more recent studies have considered a number of solutions, for example based on resampling techniques (e.g., Wieling & Nerbonne 2015), Kernel Density Estimation (e.g., Rumpf et al. 2009) or the concept of entropy (e.g., Prokić et al. 2009).

This paper presents a diagnostic measure for the detection of coherence or heterogeneity in spatial language variation aimed at identifying those regions that are particularly prone to variation or particularly sensitive to language change. We perform an approach based on nearest neighbor comparison. We exemplify our measure based on historical German dialect data.[1]

In the remainder, we provide information on the data and introduce the technique. In what follows we present example analyses and discuss the introduced procedure.

2 Data

The study makes use of a data set collected by the German linguist Friedrich Maurer during the year 1941 in the Upper German dialect region within the boundaries of the national territory at the time. The survey was based on a questionnaire with 113 individual words (most of them nouns, but also adjectives and verbs) and 10 sentences together with biographic information of the participants. In contrast to both the earlier survey by Wenker (Wenker 2013) and the contemporaneous investigation by Mitzka (cf. Wrede et al. 1926–1956), Maurer focused more strongly on social and biographic information. Thus, in addition to the age of the informants, for example, their gender as well as the origin of their parents or their preferred market towns are documented.

[1] The study builds on R programming (R Core Team 2021), using the packages spatstat (Baddeley & Turner 2005) and Rvision (Garnier et al. 2021). A R package is currently under development. A first version is available at **LINK**.

We focus on the Alemannic part of the Maurer data which is mainly related to the southwestern part of nowadays Germany (the Baden region) and the Alsace in France (see Strobel 2021 for further information). In total, the data document 2344 locations, providing a quasi-total coverage of the region under discussion (Figure 1). The hand-written questionnaires of this area have been typewritten and therefor digitalized by student assistants. The data is stored in *.csv files and will be publicly available in the future.

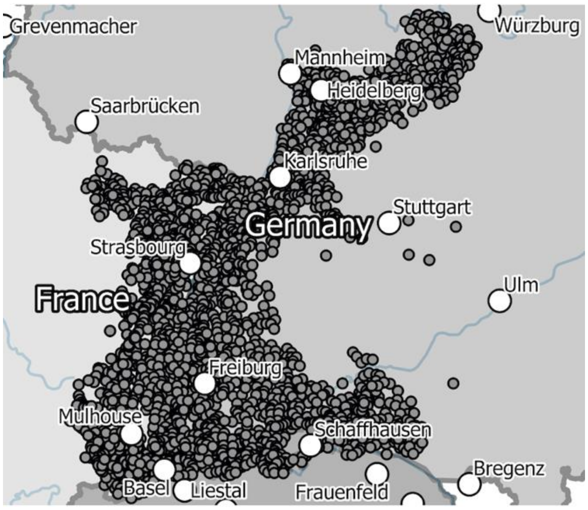

Figure 1: Study area.

3 Method

3.1 Local measure

In order to analyze the spatial variation of the area under discussion we compare the linguistic realizations of one site with the realizations of its geographic neighbors. From a technical point of view, for every site *r* we compare the linguistic realization of an individual item *i* of the questionnaire (e.g., a word) with its geographic neighbor *s*. $\mathrm{Coh}_{rs|i}$ is then the number of identities between *r* and *s* with $\mathrm{Coh}_{rs|i} = 1$ in case of identity and $\mathrm{Coh}_{rs|i} = 0$ otherwise.

To obtain a better insight into how the individual sites fit into the language region, the number of compared sites should be $S > 1$. We consider up to 10 neighbors ($0 \leq S \leq 10$), where 0 is used for the rendering of the original data. Coh_{rS} is then the average overlap between *r* and its set of neighbors *S* with $0 \leq \mathrm{Coh}_{rS} \leq 1$ and $\mathrm{Coh}_{rS} = 1$ indicating identity between *r* and *S* and $\mathrm{Coh}_{rS} = 0$ indicating no identity between *r* and *S*. In case a location has several variants for a linguistic variable (e.g., because of several participants or multiple responses), the number of matches between *r* and *s* is related to the number of local variants.

An example is provided by Figure 2. The centrally located site is opposed by a total of 5 nearest neighbors, which have a total of 2.5 matches with the central site, resulting in Coh = 2.5/5 = 0.5. The number of variants is irrelevant for this approach but is relevant for the global measure (cf. 3.2)

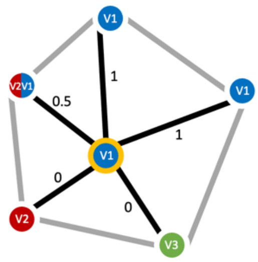

Figure 2: Model of variant distribution.

Inverting the scale results in a measure of diversity instead of coherence which we refer to as Div = 1-Coh. We use this Div measure in order to identify moments of particular dynamics on language maps.

Another point is worth mentioning. The nearest neighbor approach heavily relies on the definition of geographic coordinates and distances. In our approach, the geometric information of the spatial position for each survey site is thus originally stored in the WGS 84 format (longitude and latitude). Due to the ellipsoidal coordinate system, the distances are heavily distorted which directly effects the selection of the nearest neighbors. To use the quasi-exact distances a cartesian coordinate system is required. Therefore, we projected our data to the UTM system related to the ETRS89 ellipsoid.

3.2 Global measure

While the local measure indicates the integration of individual sites into its nearest spatial neighborhood, it says nothing about the coherence or heterogeneity of an overall map. Various options are available for this purpose. For example, the mean of all local Coh values could be taken as a global measure of coherence (CohG). However, as Figure 3 demonstrates, this measure is dependent on the number of linguistic variants in a data distribution, making it difficult to compare CohG across maps with different numbers of variants. For example, if a map shows two linguistic variants a complete random distribution results in 0,5 ≤ CohG ≤ 1 and 0.33 ≤ CohG ≤ 1 for three variants etc.

In order to solve this problem, we perform a CohG* correction in which CohG is divided by the number of variants and scaled 0 < CohG* ≤ 1. As becomes evident by Figure 3, CohG* is robust against the number of variants, while CohG, in contrast, is sensitive to it and converges to CohG* as the number of variants increases. Similar holds for the number of neighbors against which CohG* is robust while CohG is sensitive to it (not reported).

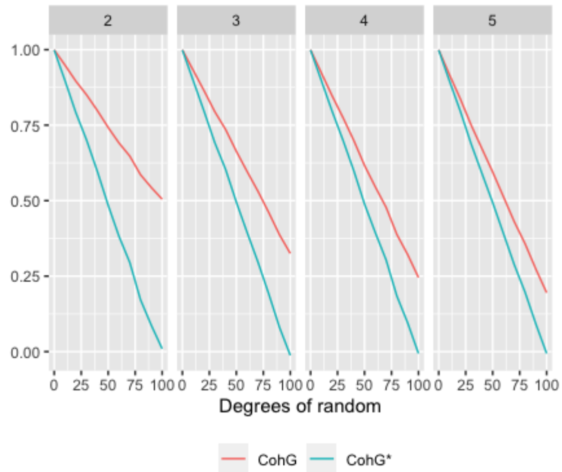

Figure 3: Comparison of CohG and CohG* based on simulated degrees of spatial coherence (0-100 %) for a data distribution with 2 to 5 linguistic variants.

Another view on CohG* is provided in Figure 4 and Figure 5. In these figures, data simulations are performed for the locations of the corpus, generating different degrees of random data distributions. Starting from a uniform distribution 20 % of the data of each map are successively overwritten with a random distribution.

While Figure 4 illustrates data simulation with two linguistic variants, Figure 5 illustrates the same procedure based on three linguistic variants. The figures show that while the CohG is related to the amount of variants, the CohG* values describe the same amount of coherence/homogeneity unattached to the number of variants.

Against this background, the Coh measure, and so is the CohG* measure, yields plausible results as far as different degrees of coherence or heterogeneity are concerned. However, it is still an open question how the values turn out in concrete use cases and what more detailed conclusions can be drawn from them.

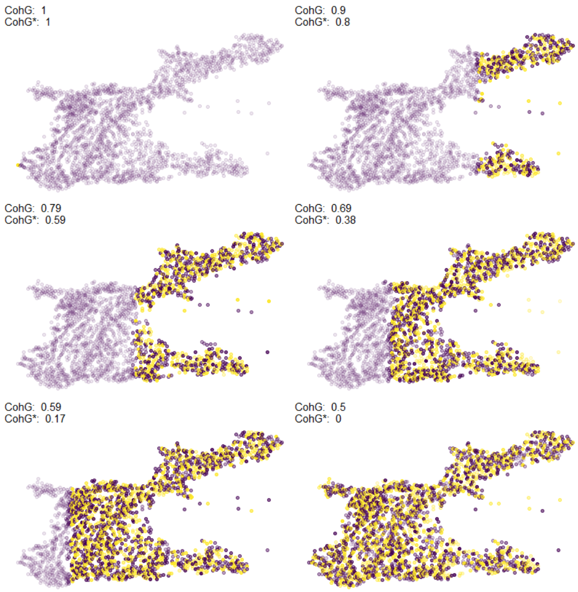

Figure 4: Simulation of different degrees of spatial heterogeneity (0 %, 20 %, 40 %, 60 %, 80 %, 100 %) for a map with two linguistic variables.

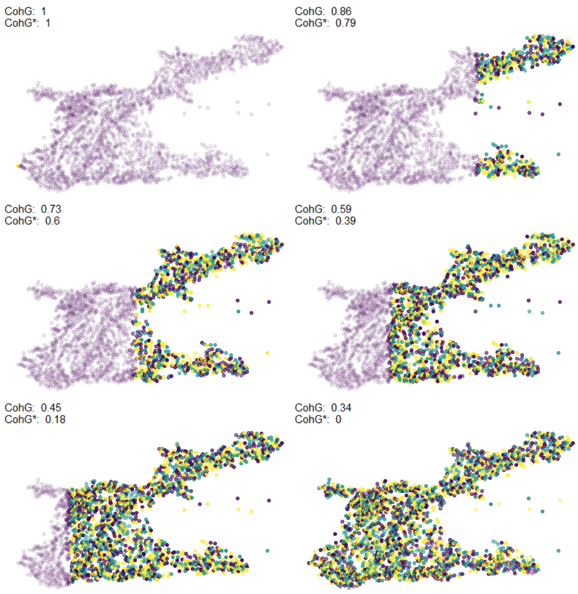

Figure 5: Simulation of different degrees of spatial heterogeneity (0 %, 20 %, 40 %, 60 %, 80 %, 100 %) for a map with three linguistic variables.

4 Use cases

4.1 Lambdacism in *Kirche* 'church'

As a first example we focus on a rather simple spatial pattern provided by the distribution of *-r-* and *-l-* sounds in the word *Kirche* 'church' (*Kirche* vs. *Kilche*) in the southern part of our study area (Figure 6). This is a so-called lambdacism, which is typical for some regions of the German-speaking area (cf. Lameli 2015).

Figure 6 illustrates the distribution of these variants in the southern part of the study area. *Kirche* (blue) occurs 1008 times, *Kilche* (red) 222 times. Hence, 81.94% of the sites in the study area show *-r-*. In a random distribution the expected probability that a particular site's neighbor shares the same variant is EV = (1008-1) / (1230-1) = 81.94%. For the same distribution we reveal under the consideration of 5 nearest neighbors CohG* = .94 (Coh = .9) indicating that, on average, 94% of the neighboring 5 sites share the same variant *-r-*.

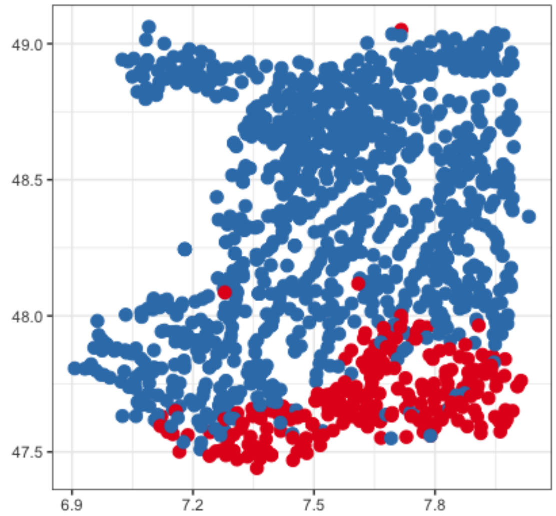

Figure 6: Example of a spatial distribution of linguistic variants *-r-* (blue) and *-l-* (red) in the word *Kirche* 'church'.

As CohG* tends to 1 (with zero as random distribution), a clear separation of the variants is evident which, at the same time, indicates the spatial clustering of *-r-* and *-l-*. Indeed, very few locations aside, all variants cluster in contiguous areas.

Testing the distribution of local Coh values against a normal distribution using a Wilcoxon rank sum test reveals a statistical difference between the expected value EV and the empirically found Coh measure (z = -4.21, p < .001, r = .94). What these measures refer to becomes evident when plotting 1-Coh (= Div) on a map (Figure 7).

As expected, the highest Div values are at the border zone between the variants. Most interestingly, there are differences depending on the spatial alternation of the variants. For example, on the left, where we find a mix of variants, Div values are high. In contrast, in the center, where we find a separation of *Kirche* and *Kilche*, Div values are low. The spots illustrated by Figure 7 thus allow conclusions to be drawn about zones of increased linguistic dynamics: around the sites with high values (intense colors) there is a high degree of variation, around the sites with low values (pale colors) there is a lower degree of variation. While the former can be expected to be more sensitive to language change regarding the variable under discussion, the latter can be expected to be more robust to language change.

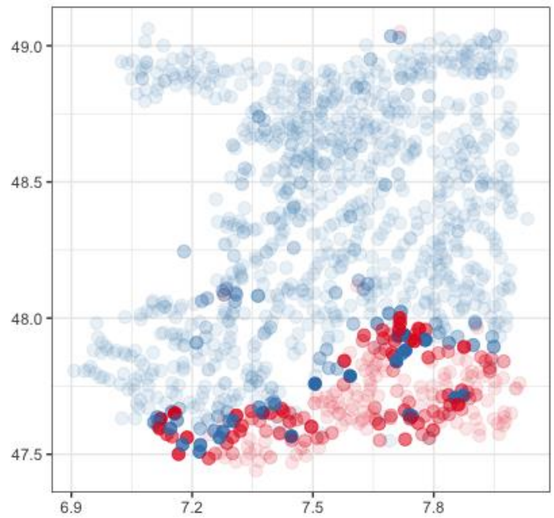

Figure 7: Local measure of linguistic coherence (Div = 1-Coh) applied to the data of Figure 6

Methodologically, it should be emphasized that, due to the nearest neighbor approach, the described procedure always computes a gradient-like result. Even if there is a sharp separation between variants (Figure 6) a gradient would be computed (Figure 7). The intensity of this gradient-like effect depends on the number of nearest neighbors. Using the minimum of two nearest neighbors will result in exact three index values and the resulting map would set a focus on areas which differ from their surroundings. This may be useful to detect islands of variation in rather coherent areas. With increasing numbers of nearest neighbors, the amount of possible index values will increase and return much more smoother transitions. This is helpful for the detection of areas with variation in a cluster-like way. Areas with variation in close distances would be smoothed to clusters which would be differentiated from surrounding homogeneous areas.

4.2 Subtractive Plural in *Hunde* 'dog-PL'

Another example is provided by Figure 8, which focuses on the whole language area of the Maurer data. The map illustrates the variation of the word ending in *Hunde* (‘dog-PL’; CohG* = .87) considering three variants (<nd>, <ng>, <nn>), of which <nn> (phonologically /n/) and <ng> (phonologically /ŋ/) has been considered as subtractive plurals (Birkenes 2014). While the *Kirche* example considers only two linguistic variants, Figure 8 refers to three variants. Figure 8 combines three different views. On the left side is the distribution of variants without any preparation, in the middle the representation of the coherence measure (expressed in Div) including information on the variants and on the right side the representation of coherence (Div) without information on the linguistic variants.

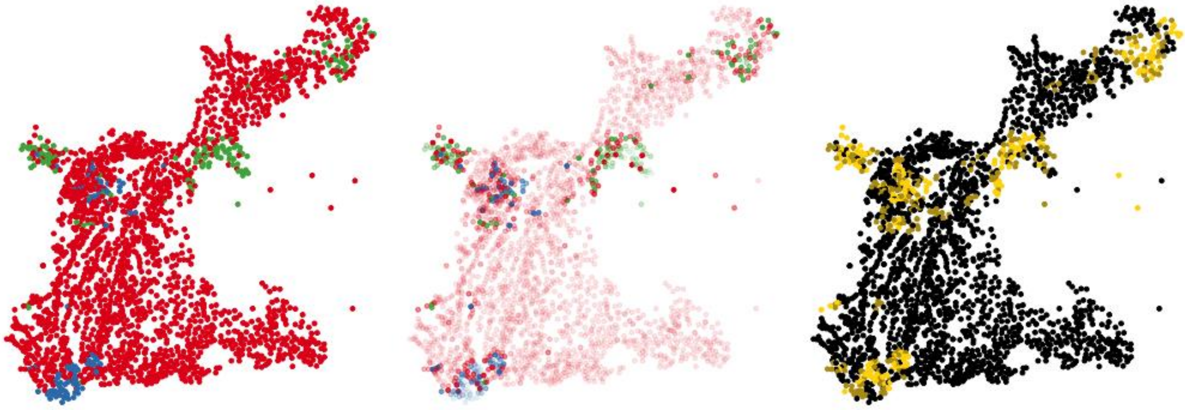

Figure 8: Local measure of linguistic coherence (Div = 1-Coh) for a linguistic variable with three variants (*Hunde* ‘dog-PL’); green = <ng>, red = <nd>; blue = <nn>; left: distribution of variants; middle: Div measure with information on linguistic variants; left: distribution of variants; middle: Div measure without information on linguistic variants.

Obviously, the coherence map in the middle clearly highlights the spots of linguistic variation. Among them are areas where only two variants interact (e.g., <nd> and <nn> in the South, <nd> and <ng> in the North), but also areas where all three variants meet (in the center). Similar to the previous example the coverage of individual variants is mapped.

The map on the right, on the other hand, emphasizes where generally such patterns of variation are encountered. This map consequently emphasizes the contrast between homogeneous and heterogeneous moments of the spatial data distribution. In this case, too, conclusions can be drawn (as in the previous example) about the extent of regional variation and possible language change events.

From a methodological perspective, the following is worth mentioning. By integrating the nearest neighbors, a smoothing effect is created, which shows linguistic variation in places where actually no variation is documented by data collection. The idea behind this is that variation is probably more widespread than what is captured by data collection. For example, if only one person is asked about a particular linguistic variant at each of two surveyed locations (which is very often the case in dialectological studies), it would possibly be wrong to take different answers per se as evidence of strict linguistic differences between those locations. Instead, it must be expected that both variants would be encountered in both localities and would be appropriately documented with other participants if data were repeatedly collected. However, the probability of this decreases with increasing geographical distance. The measure thus provides a prediction for language variation that is not visible in the data.

5 Discussion

The Coh measure, as does the Div measure respectively, reveals spots of local variation, which indicate horizontal (i.e. geographical) or vertical (i.e. social, pragmatic) heterogeneity. As Labov (2004) points out, these spots of increased language variation might be possible starting points of language change. In this regard, Bellmann (1983) considers the model in Figure 8.

Starting from a situation where variant A is the only available realization of a particular linguistic variable, at a certain time variant B becomes an alternative. This is the situation illustrated by Figure 6 for both scenarios (above and below). However, the Coh measure goes beyond local variation by modeling the closest relative area of influence of that alternative.

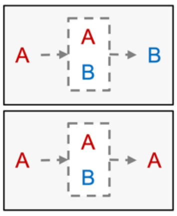

Figure 9: Possible stages in the formation of language variation and change on the example of two variants A and B; above: scenario 1; below: scenario 2.

Obviously, analysis using Coh (like Figure 7) does not specify how long the variative phase will persist. Furthermore, it could be that variant B disappears again, and it could just as well be that variant B prevails (scenario 1, Figure 9 above) while A disappears (scenario 2, Figure 9 below). Consequently, Coh does not allow for a clear prediction of the process of language change, but it does illustrate that, if language change does occur, it is likely to occur at the spots with high Div (= 1-Coh). Against this background, the relevance of the Coh measure is to indicate spots of particular linguistic dynamics. Identifying these spots enables both prediction and explanation of ongoing and/or completed language change.
Applying the coherence measure to a collection of multiple linguistic phenomena, as shown in Figure 10, leads to a new perspective on the structuring of linguistic space. Instead of highlighting the clusters of linguistic similarity, rather the zones of particular linguistic dynamics are identified. From looking at the coherence values, even without mapping, a first impression is given whether the lemmas in question show a strong spatial clustering or not. This is useful for huge datasets with dozens of linguistic variables. At the same time, it becomes evident that the measure is sensitive for outliers (i.e., isolated sites), which are evident by individual points.

Among the existing dialectometric literature, our coherence measure is comparable to the technique introduced by Rumpf et al. (2009) using Kernel Density Estimation (KDE). Our measure explicitly considers geographical neighborhood, but, in contrast to the KDE approach, it is more focused on local variation. Instead of calculating an adequate bandwith, we choose a certain number of neighbors in order to test for the integration of an individual site into the linguistic area. In this respect, the underlying concept is that linguistic space develops in small-scale communication zones, not in large-scale continua. From a technical perspective, a difference to the KDE approach is that we do not rely on the definition of individual variant-occurrence maps as an intermediate step of analysis, but process the variation given in the data set directly.

6 Conclusion

This paper introduces a nearest neighbor approach as a diagnostic tool in order to find regions which are more sensitive to language variation and change than others. For this purpose, a local measure of coherence is used (Coh). In addition, a global coherence measure (CohG) as well as a corrected global measure (CohG*) was used to quantitatively assess the spatial coherence of more comprehensive data distributions (e.g., on maps) and to automatically identify linguistic items with higher/lesser language variation. Two case studies illustrate the application of the method and the informative quality of the measures.

Limitations

The method works reliably, even if a map contains multiple variants. However, if there are more than, say, 10 or 15 variants, it can happen that no clear spots can be identified on the maps. For this matter, a more probabilistic approach would be desirable, which is currently not implemented.

Another limitation is the distance measure used for the identification of nearest neighbors. Currently, nearest neighbors are defined using Euclidean distance. This is not a problem if the analysis takes place in flat terrain (e.g., the Upper Rhine Plain). In mountainous terrain, however, this can lead to slight biases. To solve this problem, we will implement more realistic distance measures such as travel time in the future.

From a linguistic perspective, a limitation of the method is that even if it informs about the variation spots, it does not provide any information about the direction in which a possible language change could develop. However, such a statement is difficult to make without concrete comparative language data (e.g., diachronic data) or social interpretation. Since the Maurer data allow an analysis in apparent-time, further approaches for investigation will be possible in the future.

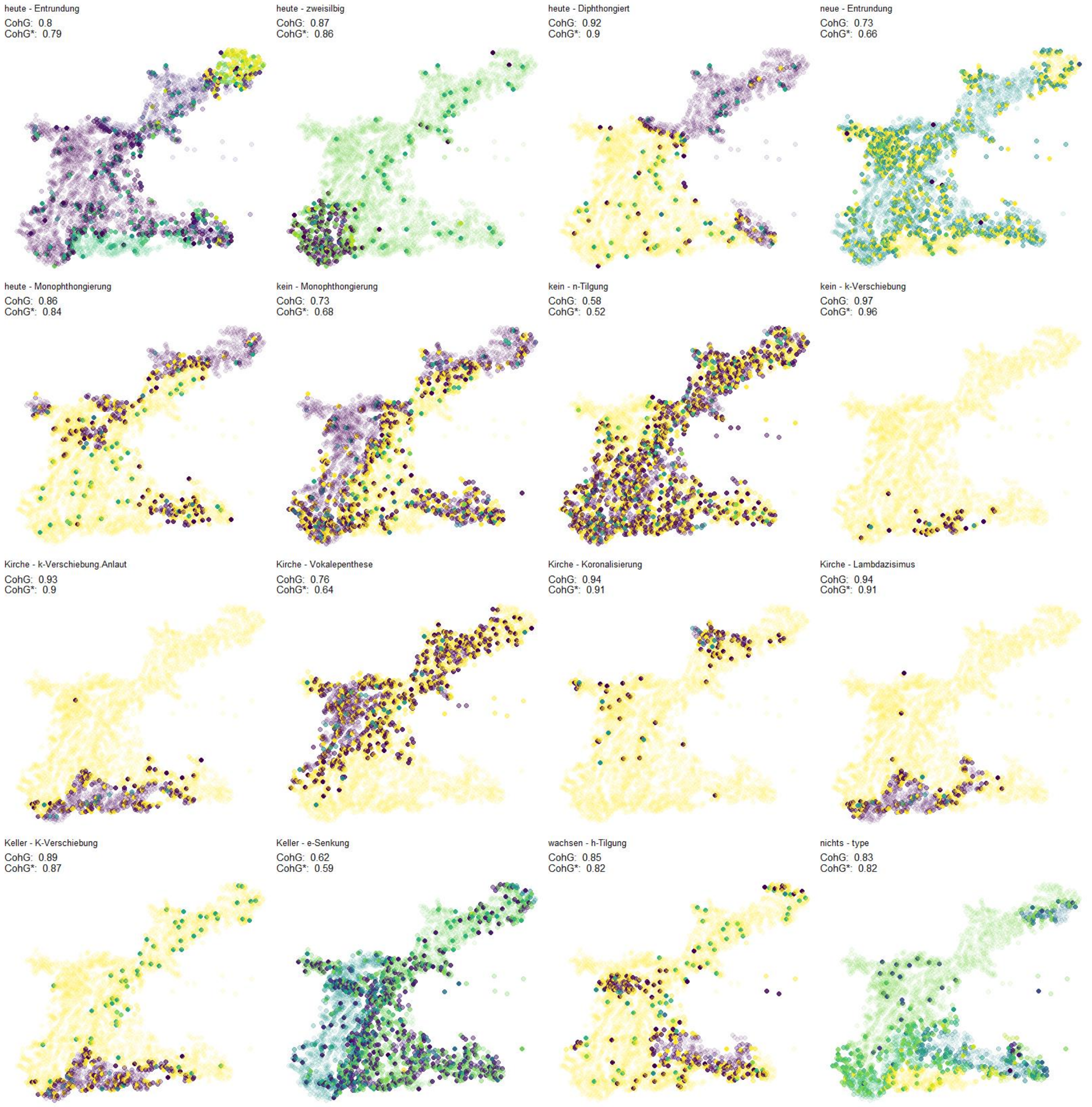

Figure 10: Local measure of linguistic coherence (Div = 1-Coh) for different linguistic variables

Ethics Statement

This work complies with the ACL Ethics Policy.

Acknowledgments

…to be written…

References

Adrian Baddeley, Rolf Turner. 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, 12(6):1-42. URL https://www.jstatsoft.org/v12/i06/.

Günter Bellmann. 1983. Probleme des Substandards im Deutschen. In Klaus J. Mattheier (ed.) *Aspekte der Dialekttheorie*. Niemeyer: Tübingen:105-130.

Magnus Breder Birkenes. 2014. *Subtraktive Nominalmorphologie in den Dialekten des Deutschen. Ein Beitrag zur Interaktion von Phonologie und Morphologie*. Steiner: Stuttgart.

Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, and Cédric Scherer. 2021. Rvision - Colorblind-Friendly Color Maps for R. R package version 0.6.2. URL https://sjmgarnier.github.io/viridis/.

Hans Goebl. 1984. *Dialektometrische Studien. Anhand italoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Niemeyer: Tübingen:191-193.

Wilbert Heeringa. 2003. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. University Press: Groningen.

William Labov. 1994. *Principles of linguistic change. Vol. 1: Internal factors*. Blackwell: Oxford.

Alfred Lameli. 2015. Zur Konzeptualisierung des Sprachraums als Handlungsraum. In Michael Elmentaler et al. (eds.) *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder*. Steiner: Stuttgart:59-83.

Jelena Prokić, John Nerbonne, Vladimir Zhobov, Petya Osenova, Kiril Simov, Thomas Zastrow and Erhard Hinrichs. 2009. The computational analysis of Bulgarian dialect pronunciation. *Serdica. Journal of Computing*, 3:269-298.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Maj-Brit, Strobel. 2021. Die Verschriftungen in der Dialekterhebung Friedrich Maurers in Baden und im Elsass als Evidenz für die Verbreitung der Standardlautung. *Zeitschrift für Germanistische Linguistik*, 49(1):155-188.

Georg Wenker. 2013. *Schriften zum „Sprachatlas des Deutschen Reichs“. Gesamtausgabe*. Olms: Hildesheim, New York, Zürich.

Martijn Wieling and John Nerbonne. 2015. Advances in Dialectometry. *Annual Review in Linguistics,* 1(1):243-264.

Ferdinand Wrede, Walther Mitzka and Bernhard Martin. 1926–1956. Deutscher Sprachatlas auf Grund des von Georg Wenker begründeten Sprachatlas des Deutschen Reichs und mit Einschluß. von Luxemburg in vereinfachter Form. Elwert: Marburg.