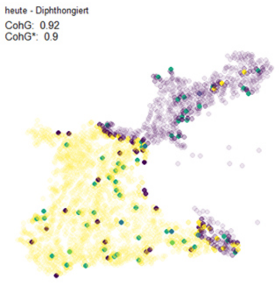

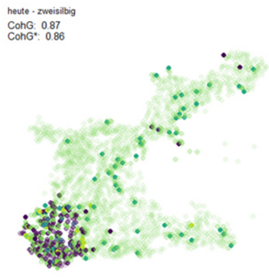

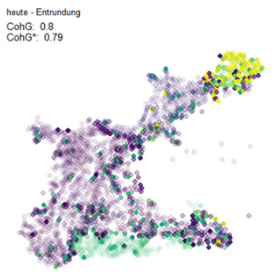

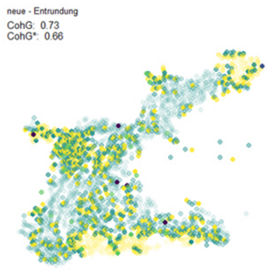

Research Center Deutscher Sprachatlas | Philipps-Universität Marburg

Alfred Lameli | Andreas Schönberg

A Measure for Linguistic Coherence in Spatial Language Variation

INTRODUCTION | Regions of linguistic similarity provide clues to the aggregated structuring of higher-level linguistic areas. 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. We perform an approach based on nearest neighbor comparison used as 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.

DATA | The study focuses on the Alemannic part of the Maurer data collected by the German linguist Friedrich Maurer during the year 1941. The survey was based on a questionnaire with 113 individual words and 10 sentences together with biographic information of the participants. In total, the data document 2344 locations in the Alemannic region, providing a quasi-total coverage of the region under discussion (Figure 1).

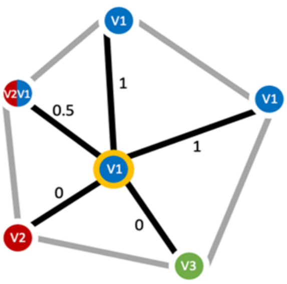

Figure 2: Model of distribution of variants.

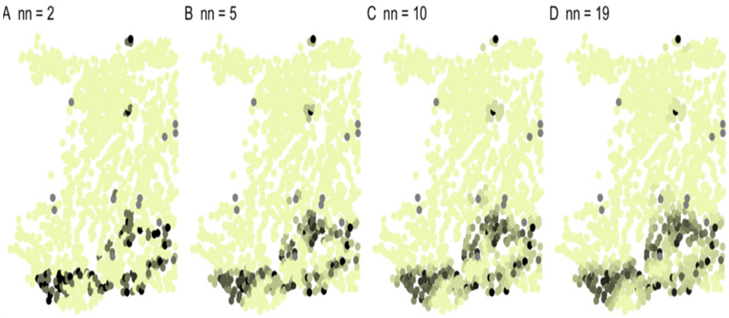

Figure 3: Local measure of linguistic coherence with number of nearest neighbors.

METHOD – 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. We use the mean of all local *Coh* values as a global measure of coherence (*CohG*). However, due to the dependency of this measure on the number of linguistic variants we perform a *CohG** correction in which *CohG* is divided by the number of variants and scaled 0 < *CohG** ≤ 1.

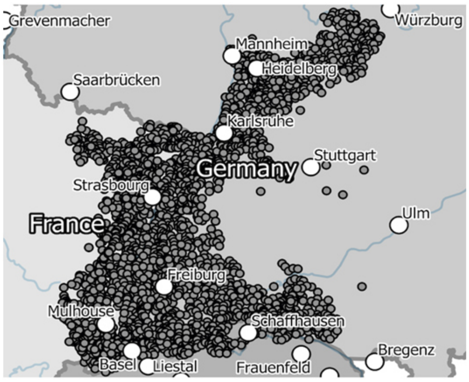

Figure 1: Study area.

METHOD – 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. 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*. *Cohrs|i* is then the number of identities between *r* and *s* with *Cohrs|i* = 1 in case of identity and *Cohrs|i* = 0 otherwise. For local variations the amount of matching language items is used (see Figure 2). The intensity of this gradient-like effect depends on the number of nearest neighbors. An increasing amount of nearest neighbors leads to a smoothing effect (see Figure 3) which can be used to detect sharp borders as well as island phenomena.

USE CASES | In order to perform our measure we choose two examples: First, the distribution of *-r-* and *-l-* sounds in the word ‘Kirche’ (church) (e.g. Kirche vs. Kilche) the so-called lambdacism (Figure 5). A more complex example is given by the subtractive plural in ‘Hunde’ (Dogs) where 3 variants occur (Figure 4).

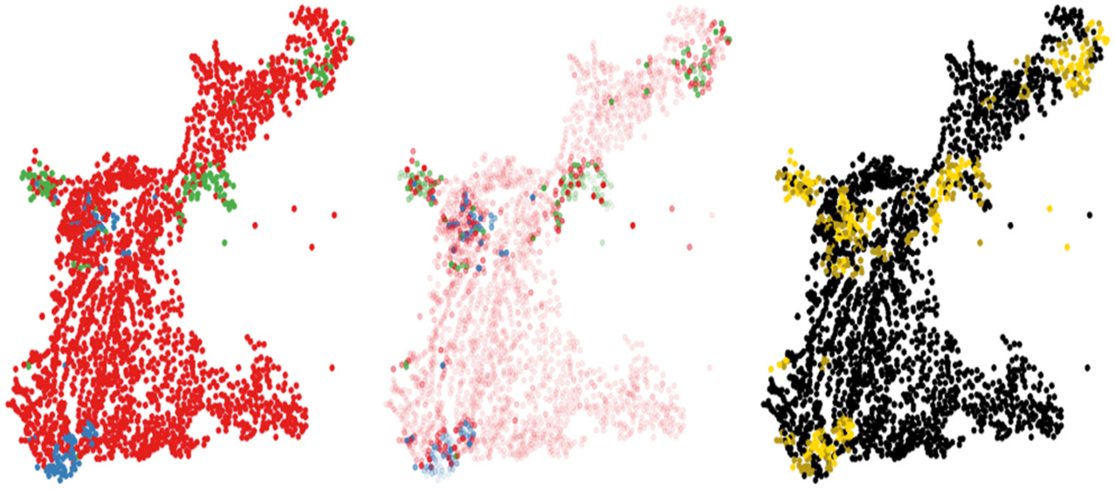

Figure 4: Local measure of linguistic coherence (Div = 1-Coh) for a linguistic variable with three variants (Hunde ‘dog-PL’); green = <ng>, red = <nd>; blue = <nn>; left: distribution of variants; middle: Div measure with information on linguistic variants; right: Div measure without information on linguistic variants.

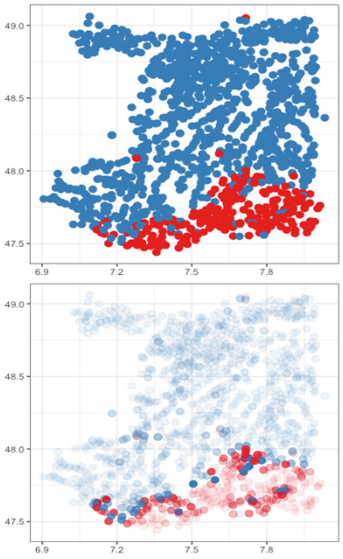

Figure 5: Above: spatial distribution of linguistic variants *-r-* (blue) and *-l-* (red) in the word ‘Kirche’ (church).
Below: Local measure of linguistic coherence (Div=1-Coh).

DISCUSSION | The Coh measure 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.

Obviously, analysis using Coh does not specify how long the variative phase will persist. Even without mapping, a first impression is given by the *CohG** whether the lemmas in question show a strong spatial clustering or not which is useful for huge datasets.

ACKNOWLEDGMENTS | We are grateful to Peter Auer, Michael Cysouw, Alexandra Lieb and Maj-Brit Strobel for discussion. This research is funded by the German Research Foundation under the project “Alemannisch variativ” (DFG, grant number 452440801).

References:
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.

Philipps Universität Marburg

Forschungszentrum Deutscher Sprachatlas