**A measure for heterogeneity in spatial language variation**

An entropy-like measurement method for spatial distribution of language items

*Andreas Schönberg & Alfred Lameli*

Abstract

Dialectometric studies usually ask about the internally consistent groups of dialects within a  
language area (see Goebl 1984). However, when dealing with larger sets of geographically specified language data, the problem arises of identifying those regions that are particularly prone to variation or particularly sensitive to language change. In our project, we follow an approach based on the concept of entropy (e.g., Prokić & Nerbonne 2008) that, in contrast to other studies (Prokić et al. 2009), is not applied to strings of tokens, but geographic distributions.

Our study deals with data from a historical language survey of German dialects at 2500 sites in the regions of Baden (Germany) and Elsass (France). These data are interesting from the perspective that they contain information on different age groups and thus enable analyses on language change. In order to identify areas which are more sensitive to language change than others we use an entropy-like measure for the identification of heterogeneity/uniformity in spatial language distributions and calculate a normalized global index. Further a transformation into a local measure of spatial variation makes it possible to automatically identify individual regions with particularly high language variation (typically the transition zones between areas of linguistic variants). This is used, for example, to predict language change or to test the correlation of spatial variation that occurs for different linguistic phenomena.

Keywords: Spatial language variation, entropy, dialectrometry.

**1. Introduction**

Dialectometric studies usually ask about the internally consistent groups of dialects within a language area (see Goebl 1984). However, when dealing with larger sets of geographically specified language data, the problem arises of identifying those regions that are particularly prone to variation or particularly sensitive to language change. The question then is not so much about stability in an area (typically indicated by the definition of clusters), but about instability. More recent dialectometric studies have introduced a number of solutions to this problem, for example, based on resampling techniques (see, e.g., Wieling & Nerbonne 2015). In our project, we follow an approach based on the concept of entropy (e.g., Prokić & Nerbonne 2008) that, in contrast to other studies (Prokić et al. 2009), is not applied to strings of tokens, but geographic distributions.

The goal of our study is the development of an entropy-based method to highlight and statistically calculate the spatial variation of speech.

**3. Data and Methods**

**3.1 Data**

The Maurer data were collected by the German linguist Friedrich Maurer during the years XYZ and XYZ focusing on the Upper German dialects within the boundaries of the national territory at that 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 (see Figure 1). In contrast to the earlier survey by Wenker (**Lit**), Maurer has differentiated more strongly by social category. Thus, in addition to the age of the informants, their gender as well as the origin of their parents and more are documented.

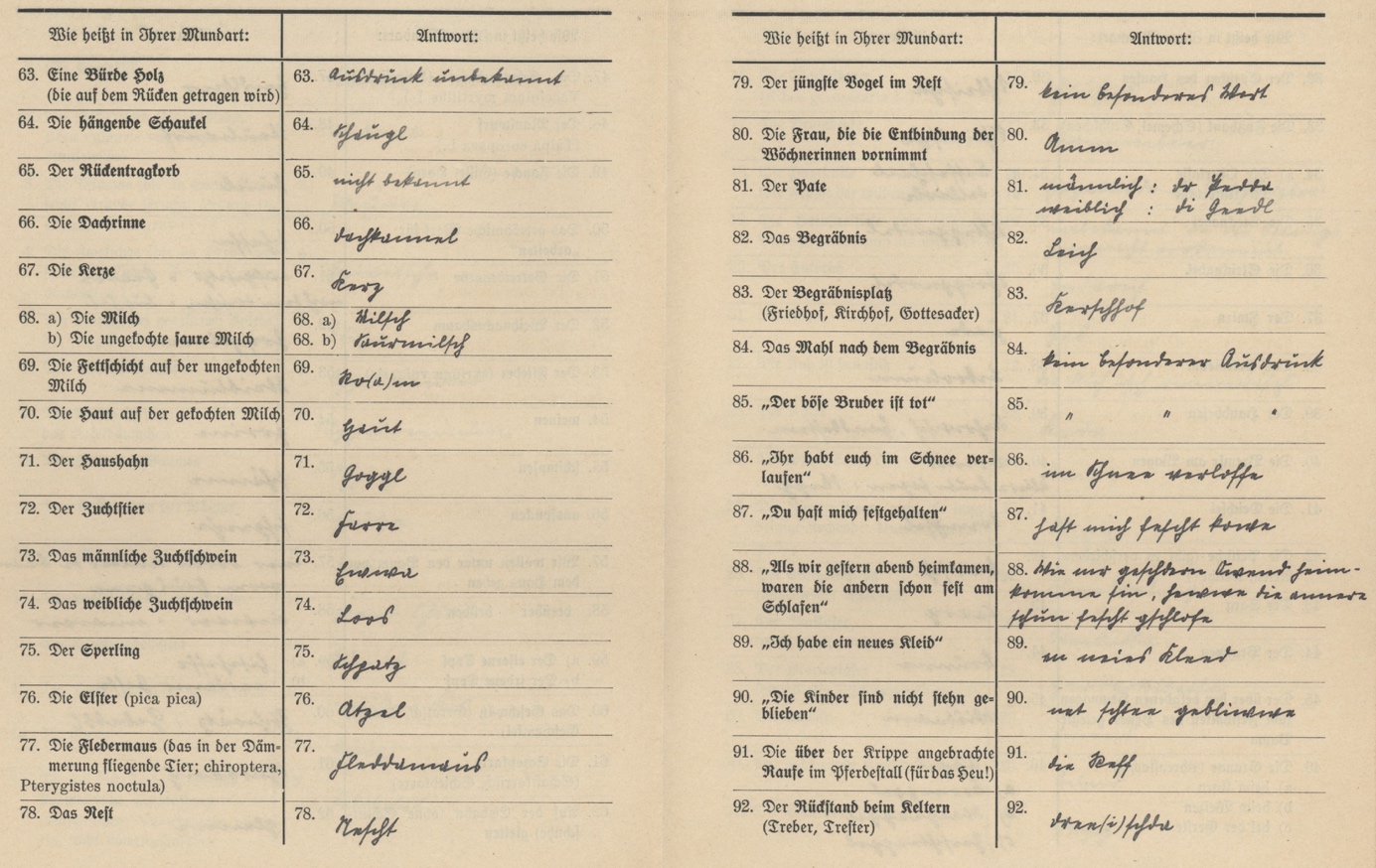


Figure 1. Example of Maurer’s questionnaire (extract)

This study focusses 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. The region is interesting from the perspective that here two cultural areas, today spread over two states, meet along a natural divide (the Upper Rhine Plain). From a natural environmental perspective, the region includes the Vosges Mountains to the East (nowadays France) and the Black Forest to the West (Germany). Linguistically, High Alemannic merges into Low Alemannic and Low Alemannic into Rhine Franconian from South to North. From West to East there is a transition zone into the Alemannic dialect group around the Lake Bodensee which is why the whole study region is of significant language variation as, for example, documented by the maps of the *Südwestdeutscher Sprachatlas* (**Lit**). In total, the data document 2500 locations, providing a quasi-total coverage of the region (Figure 2).

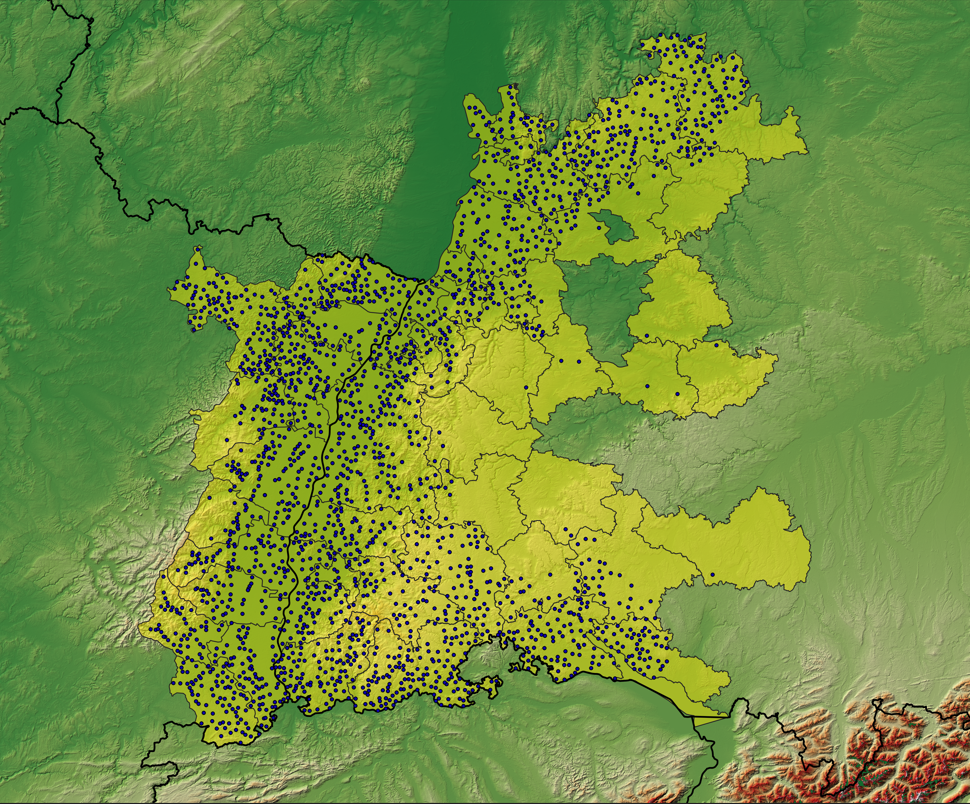


Figure 2. Study area

**3.2 Methods**

**3.2.1 Measure of linguistic coherence**

In order to analyze the spatial variation of the region under discussion we compare the linguistic realizations of one site with the realizations of its geographic neighbors. More technically, 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*. *Coh*rs|i is then the number of identities between *r* and *s* with *Coh*rs|i = 1 in case of identity and *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. In our analysis we consider up to 10 neighbors (2 ≤ *S* ≤ 10). *Coh*rS is then the average overlap between *r* and its set of neighbors *S* with 0 ≤ *CohrS* ≤ 1 and *Coh*rS = 1 indicating identity between *r* and *S* and *Coh*rS = 0 indicating no identity between *r* and *S*. The mean of all local *Coh* values results in a global measure of coherence referred to as *CohG* ranging between 0 ≤ *CohG* ≤ 1. 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 for cartographic purposes to identify moments of particular dynamics on language maps.

As an example, take a distribution of variant V1 which occurs 1008 times in the corpus, V2 occurring 222 times. Hence, 81.95% of the sites in the study area show V1. 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 indicating that, on average, 94% of the neighboring 5 sites share the same variant V1. As *CohG* tends to 1, it is clear that we are rather dealing with the spatial clustering of V1 and V2 and rather not with a random distribution of those variants. Testing the distribution of local *Coh* values against a normal distribution using a Wilcoxon rank sum test reveals *z* = -4.21, *p* <. 001, *r* = .94 indicating a statistical difference between the expected value *EV* and the empirically found *CohG* measure.

The nearest neighbor approach heavily relies on the definition of geographic coordinates and distances. In our study, the geometric information of the spatial position for each survey places is originally stored in the WGS 84 format (Longitude and Latitude). We project the survey data onto the UTM system using the ERTS98 ellipsoid. For our study area of southwestern Germany the UTM Zone 32 is appropriate for the calculation of spatial distances.[[1]](#footnote-1)

Among the existing dialectometric literature, our coherence measure is most 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. More technically, 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 directly process the variation given in the data set.

**3.2.2 other methods???**

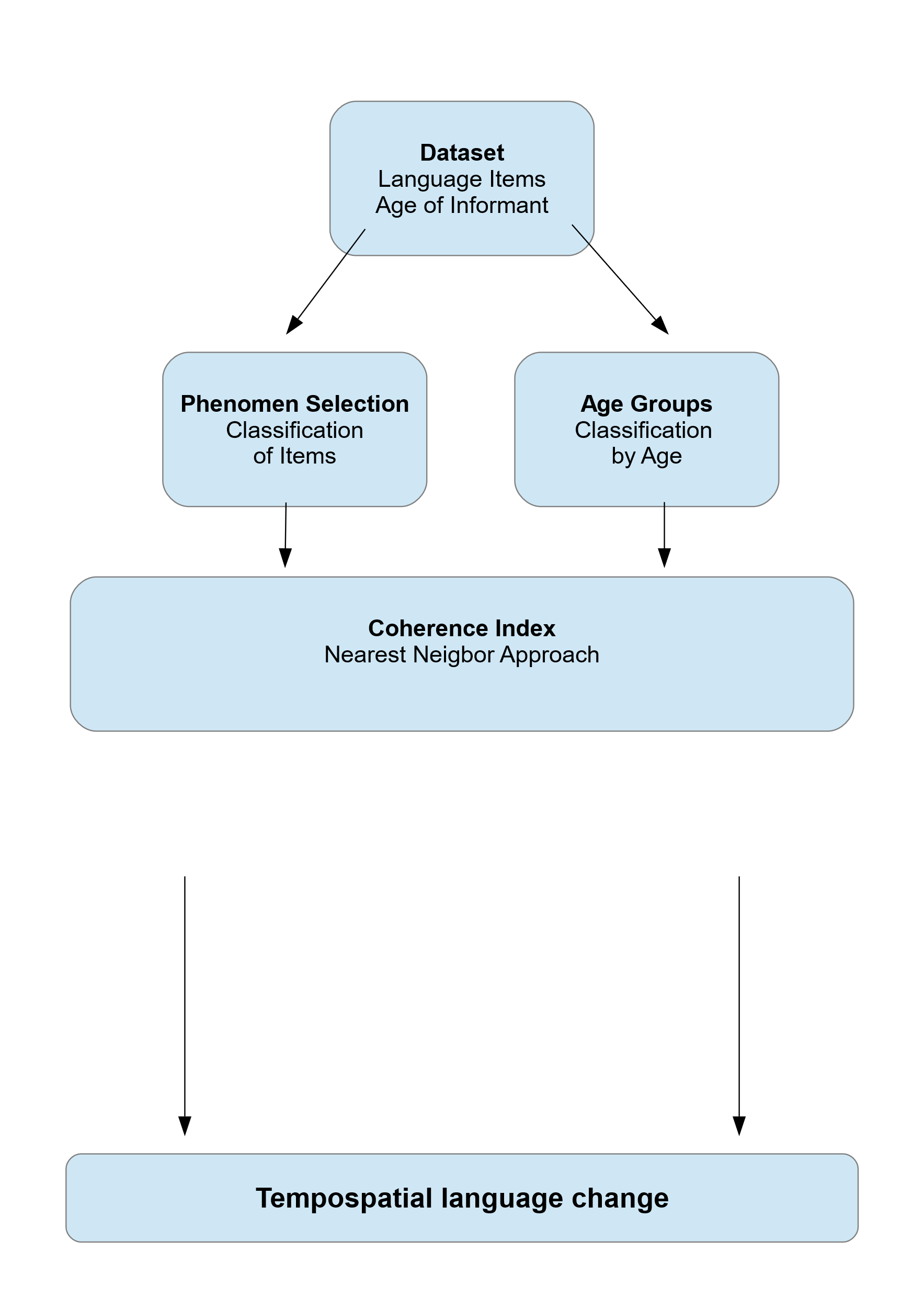
****

Fig. 2: Simplified schematic workflow for data analysis.

**3.2 Data preprocessing**

**Classification of language items**

The language items in the maurer survey data typically consist of dozens of unique strings. Using the original strings to visualize spatial distribution of variation would lead to little meaningful results and relatively useless for our cause. Despite the normal variation of language items the dataset has much more unique strings due to leading or tailing additives like article, genus or even different use of Unicode/special symbols. To process those strings we will use regular expressions to classify desired phenomenon in a semi-automated approach. Therefore the distribution and amount of unique strings for the selected language items are investigated for any phenomenon of interest. In the following the phenomenon of interest is classified by assigning the class value based on the matches to the regular expression which describe the respective classes for the strings. For our study we will use some selected language items and respectively phenomens (see table X). By the use of this classification we can generate maps for the spatial distribution of langue items.

|  |  |  |
| --- | --- | --- |
| Language item | Pattern (regEx) | Classes |
| „hunde“ | "nd|nt","ng|n.g","nn|n$" | "nd" , "ng" , "nn" |
| „heute“ | „x“,“y“,“z“ | „x“,“y“,“z“ |
| „brunnen“ | „x“,“y“,“z“ | „x“,“y“,“z“ |
| „weitere“ | „x“,“y“,“z“ | „x“,“y“,“z“ |

Table X: selected language items and phenomenon classification

Eventuell eigene unterkapitel mit den Phänomenen und Begründung warum interessant?

**Generation subsetting**

In addition to the general language item classification and distribution maps we are able to generate subsets for each langue item by using the age data for the informant. This allows the comparison of the distribution of language item classes and their respective differences over time. This allows the comparison of the distribution of language item classes not just in a spatial way but although over time (Satz alternative). To perform though a tempo spatial analysis we will group the places by the age of the respective informant in three classes.

|  |  |  |
| --- | --- | --- |
| Generation | Range Age from | Range Age to |
| „young“ | Min (XY) | 20 |
| „middle“ | 21 | 50 |
| „old“ | 51 | Max (XY) |

Table X: Generations and selected Ranges of age.

**3.3 Neares neigbor approach**

For our goal to compute an entropy-like index of homogeneity/uniformity we will use a nearest neighbor approach based on the “spatstat” R-package developed by Baddeley et al 2015). First an observation window is computed for the study area including all survey places in a two dimensional plane. By using this observation window a point pattern is computed which represents all survey places and allows the calculation of distances due to the planar space. In the next step a selected amount of nearest neighbors is calculated for each survey point within the point pattern which returns a table with the specific survey point IDs of the neighbors for each survey point.

Further the language item class for each neighbor is compared to the survey point the neighbor is assigned to. The neighbors are marked as equal (using a “1”) if the language item class matches with the central point and on the other hand marked as unequal (using a “0”) despite which class as long as it is unequal.

Next the amount of neighbors which are matching in terms of the language item class with the central point are summed up. At this state each survey place consist of a value representing the amount of equal language item class in the neighborhood depending on the selected amount of nearest neighbors.

This value is further normalized based on the amount of nearest neigbors which comput? What and idex.

**4. Results**

To investigate our hypothesis we classified XX langue items in total and computed the equivalent amount of language distribution maps as well as coherence maps.

**4.1 Influence of Nearest Neighbor amount**

Irgendwo muss noch erwähnt werden in Methods, dass wir die influrence testen wollen?

To estimate the influence of the selected amount of Nearest Neighbor in general the coherence index function was tested with a sequence of 2 to 10 Nearest neighbors. For a more objective testing this sequence was applied for every classified language item. In the result we can show that the number of nearest neighbors does not have a significant influence according to the general global coherence index value (see figure X).

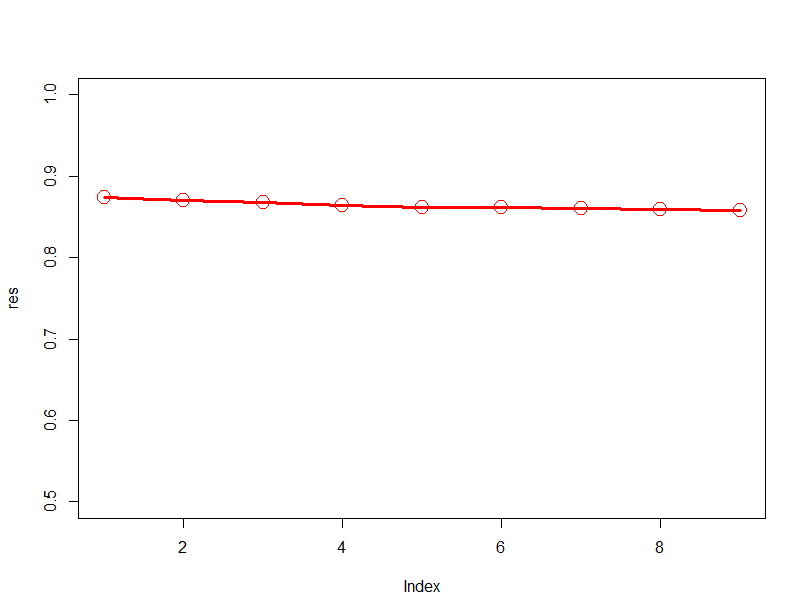
Eine Grafik je testlauf oder alle in einem (je nahc dem was aussieht).

Figure X: Resulting gobal coherence index value based on different nearest neighbor maounts used in the function.

Visual comparison of NN amount

While there is nearly no effect on the global coherence index value the amount of nearest neighbor used for the index computation has a decent effect on the spatial picture if visualized (see fig XX).

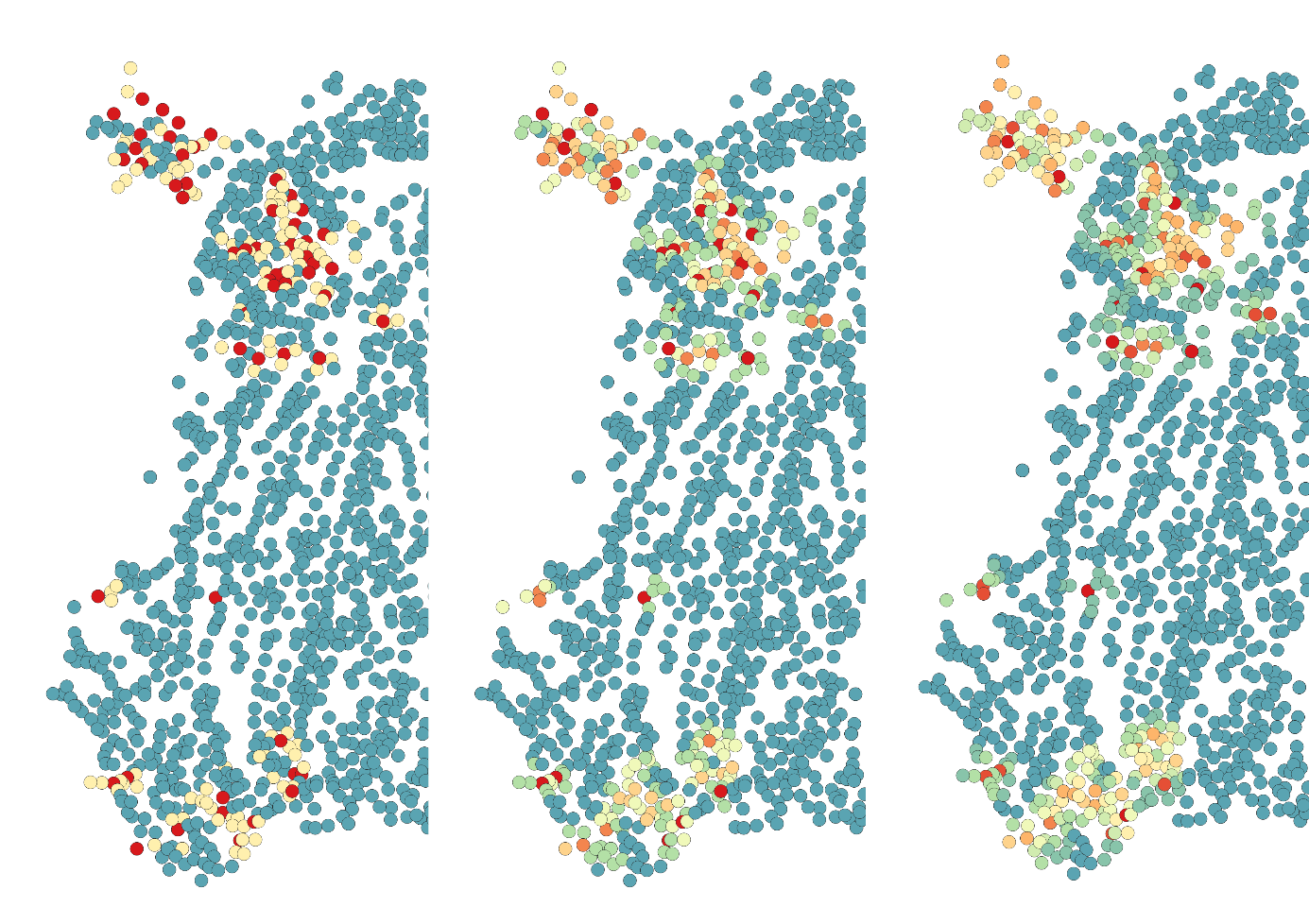
****

Figure X: Smoothing effect based on higher NN amount (exemplarischer text) NN 2,5,10

The use of higher amounts of nearest neigbors leads to a smoothing effect for the index values and the respective distribution. If using two nearest neighbors which represents the minimum value the possible resulting unique values are limited to three: all neighbors are equal, half neighbors are equal, no neighbor is equal. This leads to relatively sharp crossings. Using more nearest neighbors would constantly increase the amount of possible index values within the same range of values which leads to a more gradient-like smoothing.

Hier eine reihe von min zu max um den „smoothing effect zu zeigen“

Hier weiterer Text, sobald ergebnise vorliegen  
Reicht hier ein Beispiel zB Hunde mit 2,5,10 NN oder auch mehrere Language items für den Test?

Weitere ergebnisse wie aufbauen?

Entweder erst global nur werte um einschätzung ohne bild zu zeigen.

Dann Classifizerung und daneben COH und beschreiben.

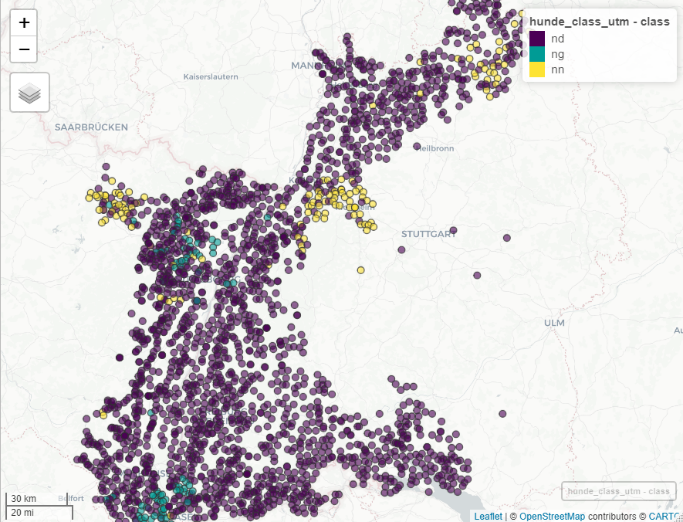
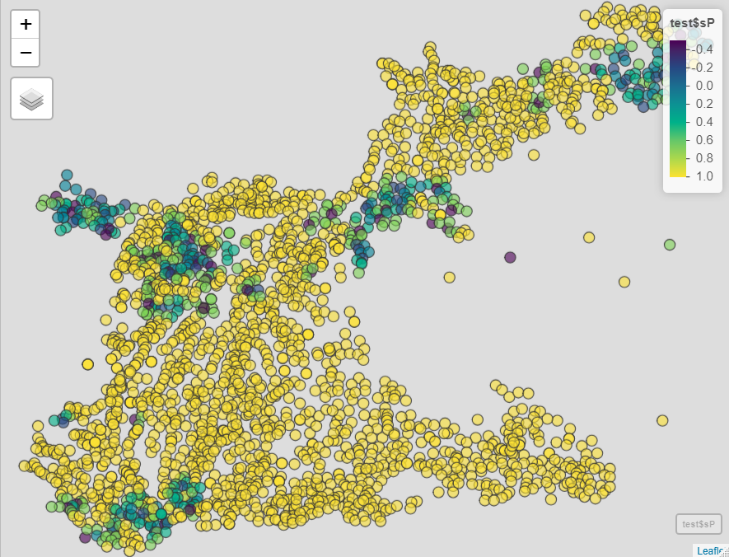
Dann Generationene anwenung

ODER

Langue item nach einander und jewails Classifizerung und COH zeigen und global, local beschreiben, ggf Generation nur ebei einem ausgesuchten? (das passt besser zur logik in methods auch die Zeile einzeln nacheinander)

**4.2 Coherence Index Global Value**

The global value for our coherence index can be used to detect language items which may have interesting result even without a visualization. This is probably helpful for huge datasets. For our selected language items the global index value varies from XX to XX. While XX   
erklären, was Index wert bedeutet. (Ggf Tabelle der ergebnisse?)

****

**5. Diskussion**

Auf Hypothese eingehen und über ergebnisse verifizieren H1 (oder eher in conclusion?)

Our first hypothesis was that we are able to develop an entropy-like spatial index value which highlights the spatial variation crossing for selected languge items. Based in the reusölts we can confirm this.

Zu Global Index

Wie ist der global wert zu interprätoieren? Die rechnung des Wertes und das ergbeiss ist ja auch im local maß gültig, daher hier eher generelle diskussion über das mass und die rnage an sich?

Funktioniert das auch?

The global index value should give a first and simple overview if an language item is of a higher interest. This is usefull for huge datasets with dozens of language items. Values around the maximum range (1) indicate a homogeneous distribution and may be of lesser interest because this indicates less to even no change in variation.

0 erwartungswert gleich verteilt

Was ist dann minus wert?

Ergbenis ist IMMER ein Gradient, slebts wenn die Grenze Messerscharf ist wird ein Gradient erzeugt, der ggf einen falsche Eindruck des räumlichen Bildes erzeugt.  
Gradient darf nicht interpretiert werden. (Micheal) (aber das liegt ja an den NN anzahlen)

The developed function always computes a gradient-like result due to the nearest neighbor approach. Even if there is a very sharp crossing between two variants of language item classes a gradient would be computed. Thus sharp crossings cannot be differentiated to more smoother crossing over several langue items for equal distances. It is recommended to avoid the interpretation of this gradient in terms of the sharpness or smoothness of the crossing. The intensity of this gradient-like effect depends on the amount of nearest neigbors due to the change of possible index values with the same value range. Using the minimum of two nearest neigbors will result in exact three index values and the resulting map would set a high focus on areas which differ from the sourroundings. This may be useful to detect islands of variation in generally homogeneous areas. With increasing numbers of nearest neigbors the amount of possible index values will increase and return much more smoother crossings. 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 souraounding homogenious areas.

Beispiel Grafik einbauen?

(ganz scharfe grenze, serh heterogene grenzen, danbene die coh ergebnisse)

Coherenz Karte zeigt NICHT um welche Variation es sich handelt -> Informationsverlust  
Man kann nicht unterscheiden zwischen Übergang a zu b zu c oder a zu b zu a  
Coherenz zeigt WO Variation stattfindet, nicht WELCHE Variation

The coherence index is a derivate of the originally spatial language distribution map. In the process of calculating the coherence index by the nearest neighbor approach the information about the class of language items is lost. Therefore the output map can no longer be used for interpretation of the distribution for the specific class of language item but the distribution of variation in general. Crossings from language item class a to b and c will no longer be differentiated to crossings of a to b and again to a.

Vorteil coh index map vs standart verteilungskarte

Zb bei generation, da fehlende werte durch NN gesmooth werden.

**6. Conclusion**

Conclusion Texfeld

**References:**

**Dummy, D.M. (2019):** Titel*.* In: Journal. Volume: X, Pages: X-X

**Dummy, D.M. (2019):** Titel*.* In: Journal. Volume: X, Pages: X-X

**(Baddeley et all 2015)**

**Baddeley, A., Rubak, E. & R. Turner (2015):** Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015. URL <https://www.routledge.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/>

**Dummy, D.M. & M.D. Dummy (2019):** Titel very long lonnnnnnnnnnnnnnnnnnnnnnnnnnnnng*.* In: Journal. Volume: X, Pages: X-X

**Dummy, D.M. & M.D. Dummy (2019):** Titel very long lonnnnnnnnnnnnnnnnnnnnnnnnnnnnng*.* In: Journal. Volume: X, Pages: X-X

**Dummy, D.M. & M.D. Dummy (2019):** Titel very long lonnnnnnnnnnnnnnnnnnnnnnnnnnnnng*.* In: Journal. Volume: X, Pages: X-X

**Dummy, D.M. & M.D. Dummy (2019):** Titel very long lonnnnnnnnnnnnnnnnnnnnnnnnnnnnng*.* In: Journal. Volume: X, Pages: X-X

**Dummy, D.M. & M.D. Dummy (2019):** Titel very long lonnnnnnnnnnnnnnnnnnnnnnnnnnnnng*.* In: Journal. Volume: X, Pages: X-X

Adrian Baddeley, Ege Rubak, Rolf Turner (2015). Spatial Point Patterns: Methodology and  
Applications with R. London: Chapman and Hall/CRC Press, 2015. URL

https://www.routledge.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/

1. Data and tools (package, scripts) are publicly available according to the FAIR principles under **<hyperlink github repo>**. Nearest neighbors are identified using the spatstat R-package (Baddeley et al. 2015). [↑](#footnote-ref-1)