Dialect Feature Detection

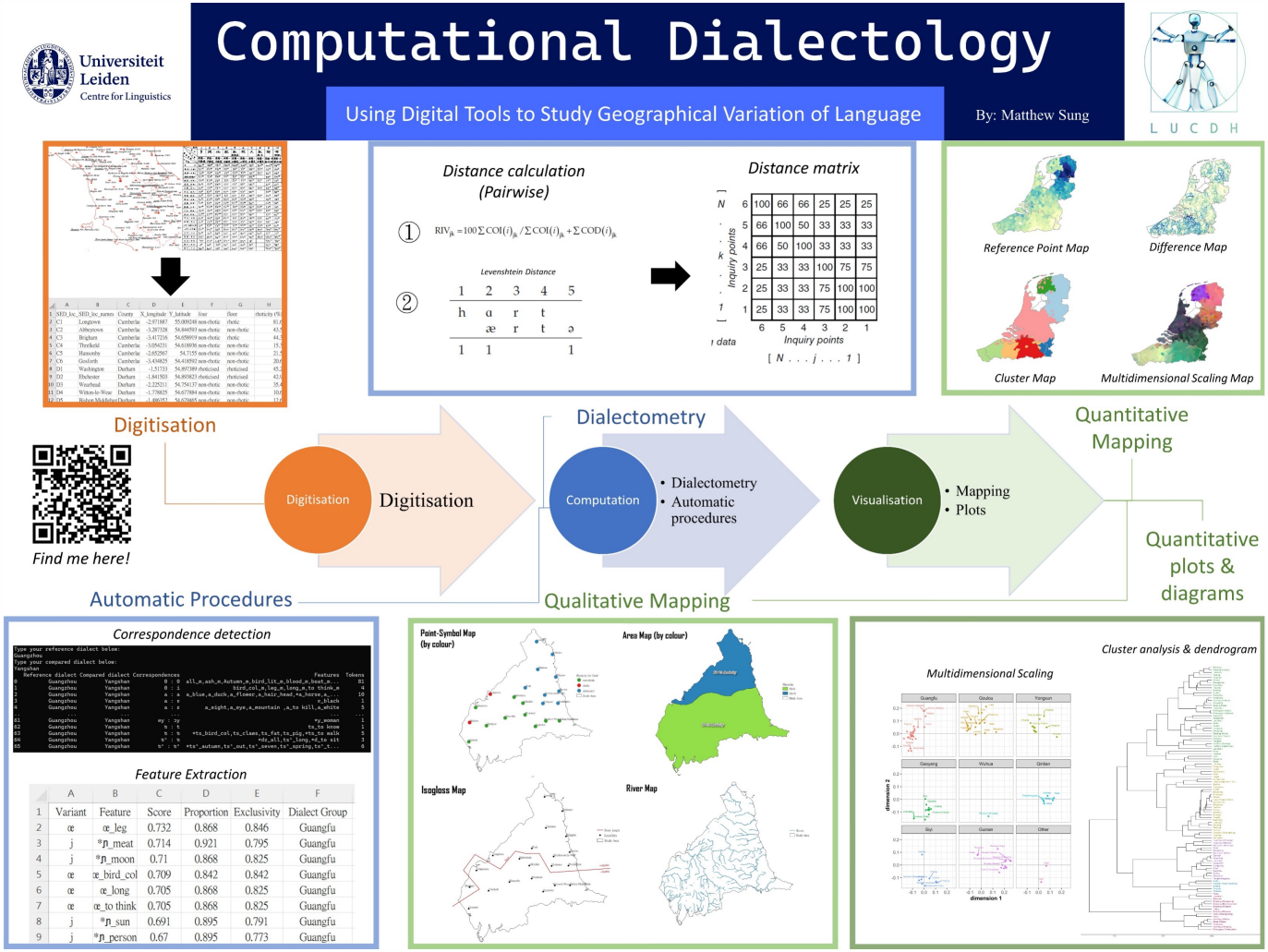
Jelena Prokic & Matthew Sung, LUCDH/LUCL

Universiteit Leiden
The Netherlands

Dialectometry

- Quantitative dialectology
- Uses large amounts of material
- Reduces the subjectivity
- Offers statistical analysis of differences

Dialectometry Workflow



Dialect Feature Detection

Workflow Summary

Workflow Summary

- Measuring distances (number of shared features)
  - phonetic/phonological level
  - morphological level
  - lexical level
  - syntactic level

Workflow Summary

- Measuring distances (number of shared features)
  - phonetic/phonological level
  - morphological level
  - lexical level
  - syntactic level

- Detection of dialect groups

Workflow Summary

- Measuring distances (number of shared features)
  - phonetic/phonological level
  - morphological level
  - lexical level
  - syntactic level

- Detection of dialect groups
  - clustering
  - multidimensional scaling

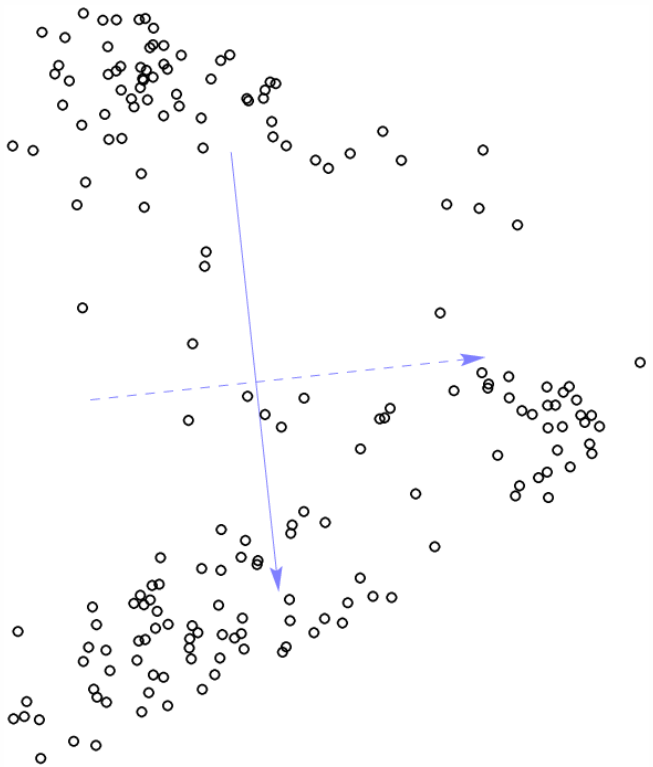
Workflow Summary

- Measuring distances (number of shared features)
  - phonetic/phonological level
  - morphological level
  - lexical level
  - syntactic level
- Detection of dialect groups
  - clustering
  - multidimensional scaling
- Linguistic interpretation*

Calculate Distances

- String edit distance (SED)
  - align pronunciations of the same word in two locations
  - calculate number of different sounds
- Repeat for each word and each pair of locations
- Example:

MDS Plot



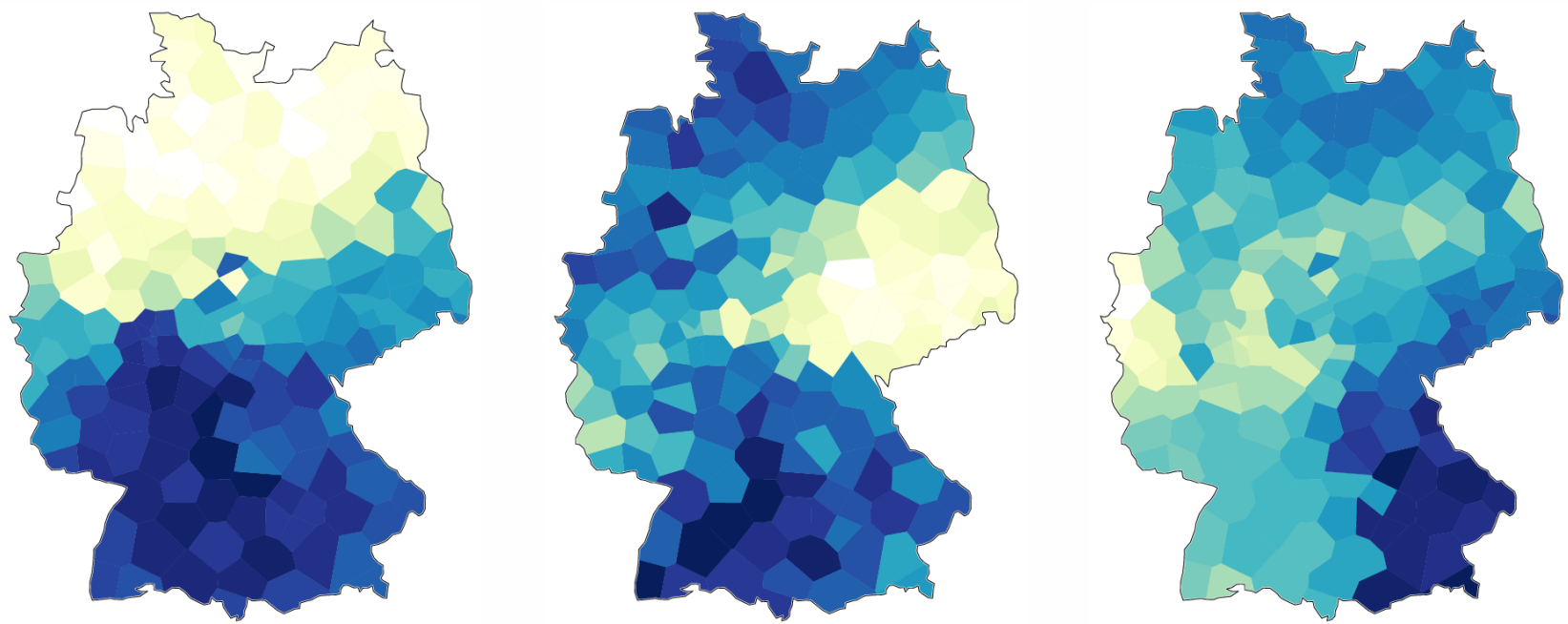
2 dimensions: r=0.74

Dialect Feature Detection

MDS Map

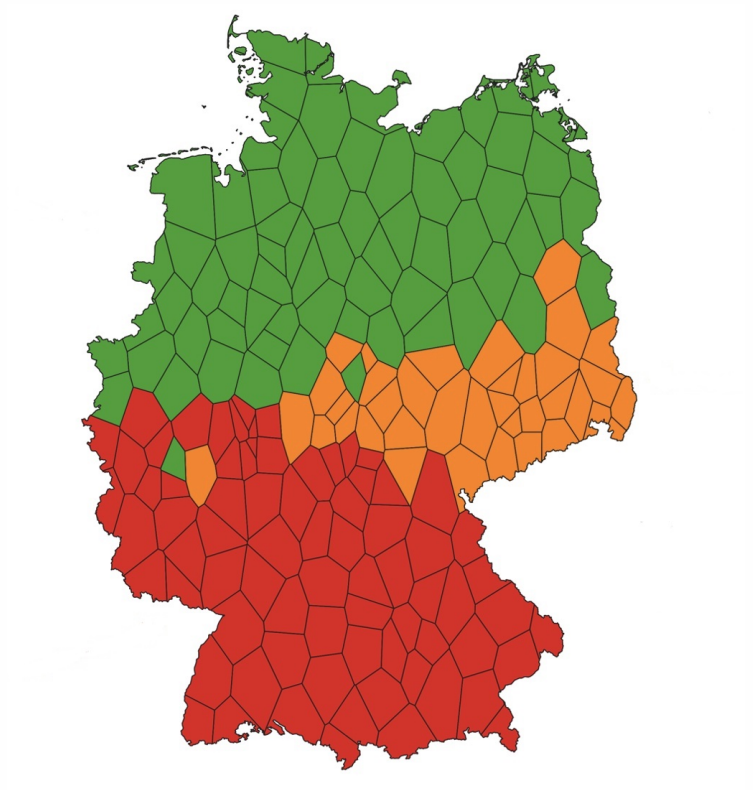
3 dimensions: r=0.80

MDS Dimensions



Dialect Feature Detection

Clustering



3 groups identified by Ward's clustering method

Characteristic Features

- What is typical for the northern dialects?

Characteristic Features

- What is typical for the northern dialects?
- What is typical for the southern dialects?

Characteristic Features

- What is typical for the northern dialects?
- What is typical for the southern dialects?
- Actually, hard to say

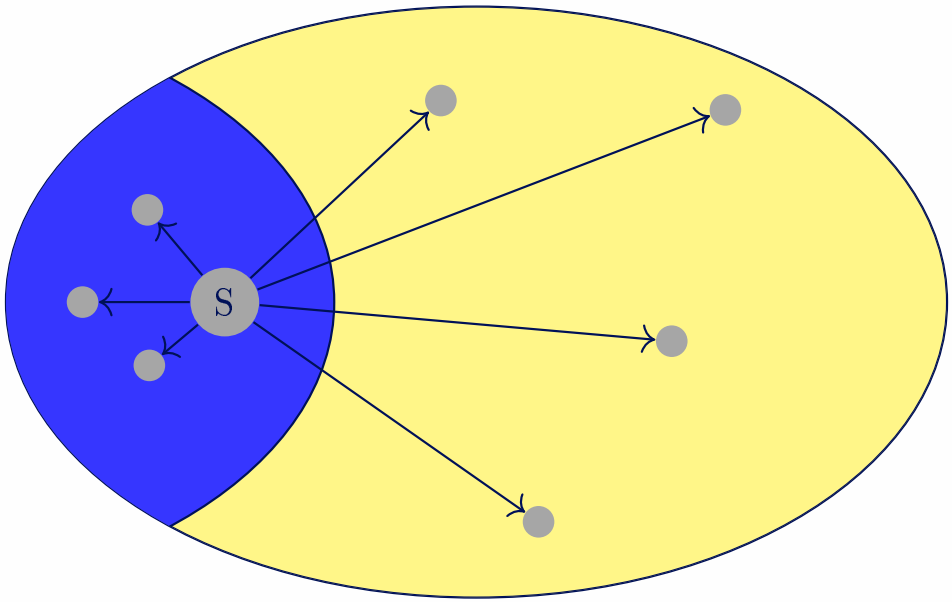
Related Work

- Various approaches
  - Shackleton (2005); Nerbonne (2006); Grieve (2009)
  - Wieling & Nerbonne (2011); Prokic et al. (2012)
    - Distinctiveness
    - Representativeness
  - Factor Analysis (Pickl 2016)

Top-down Approaches

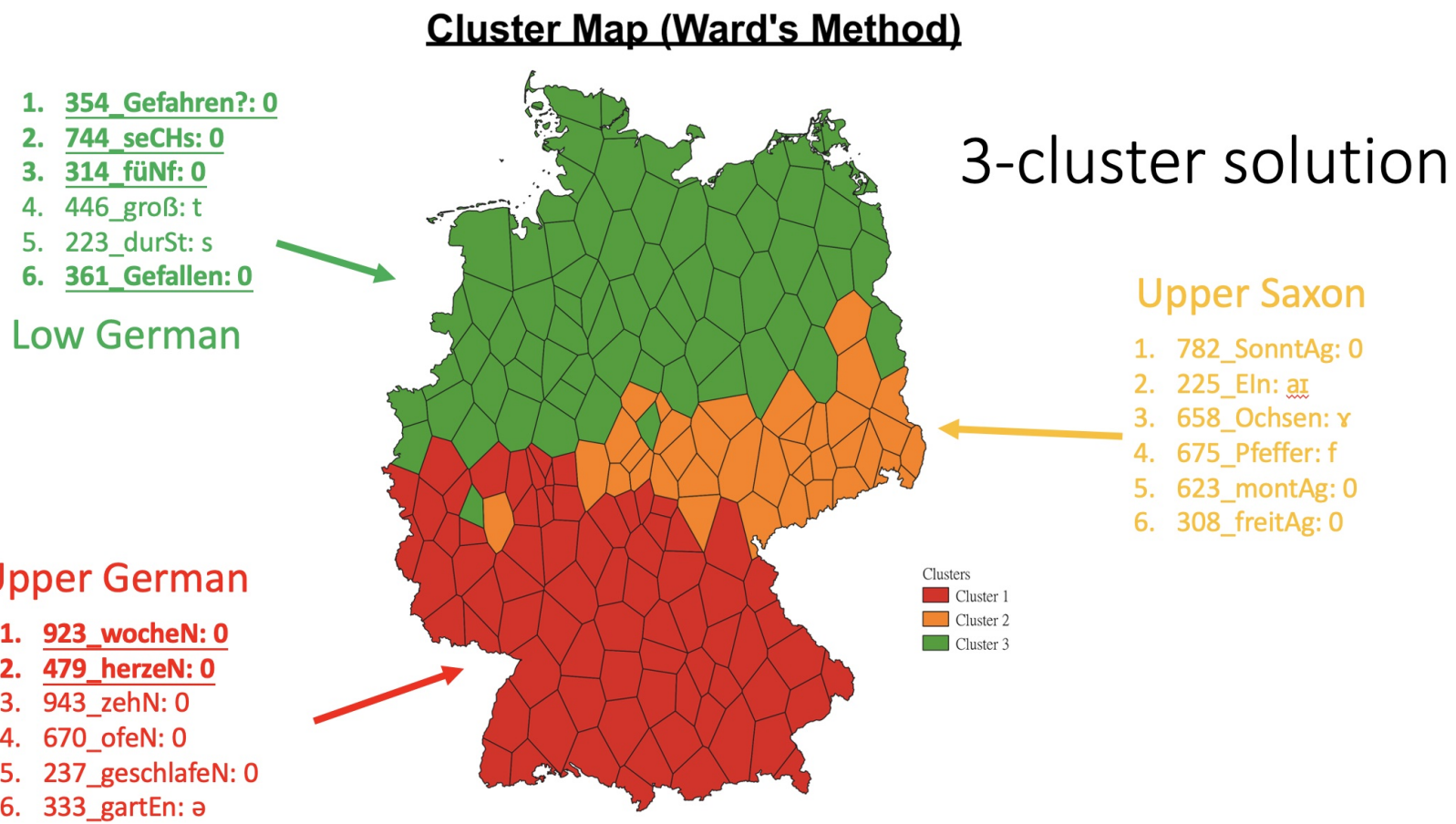
- Prokic et al. 2012; Normalized Pointwise Mutual Information (Sung & Prokic, in progress)
- Detect clusters
- Compare the mean distance of all pairs of sites within a group, to the mean distance of the pairs

Top-down Approaches



Top-down approach: compare all sites within and outside a cluster

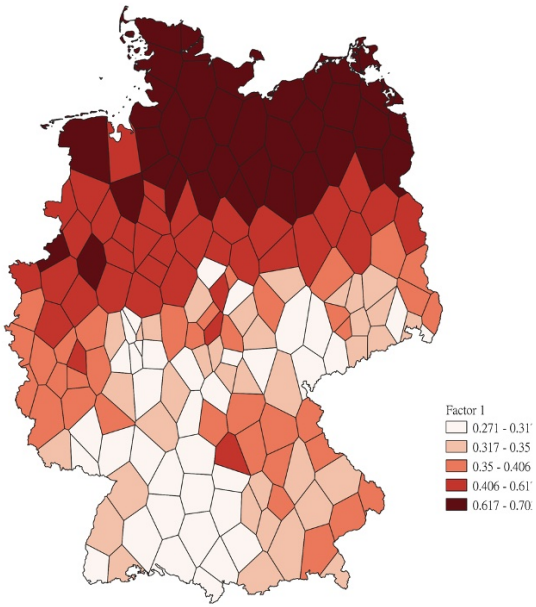
Identified Features



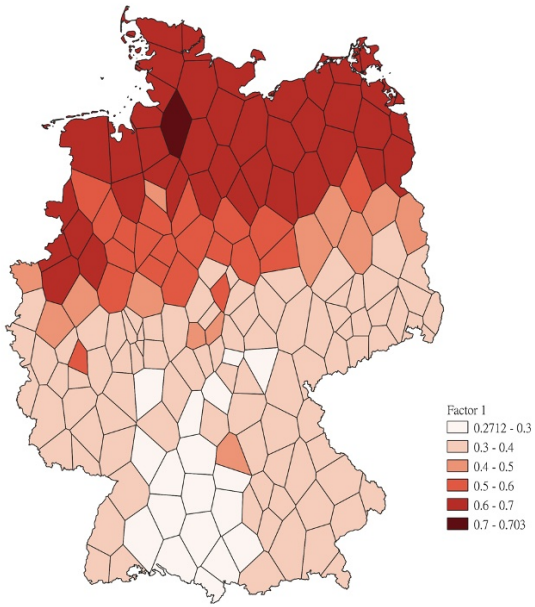
Bottom-up Approach

- Pickl 2016
- Bottom-up approach
- Seeks simultaniously
  - distinctive features
  - clusters

FA: Factor 1



Equal Quantile

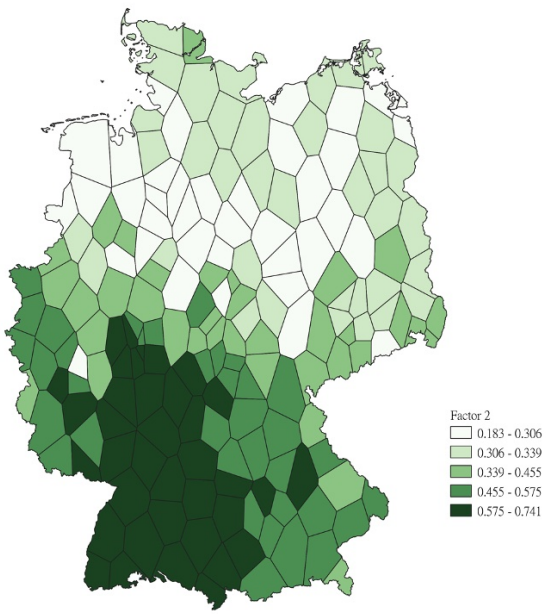


Pretty Breaks

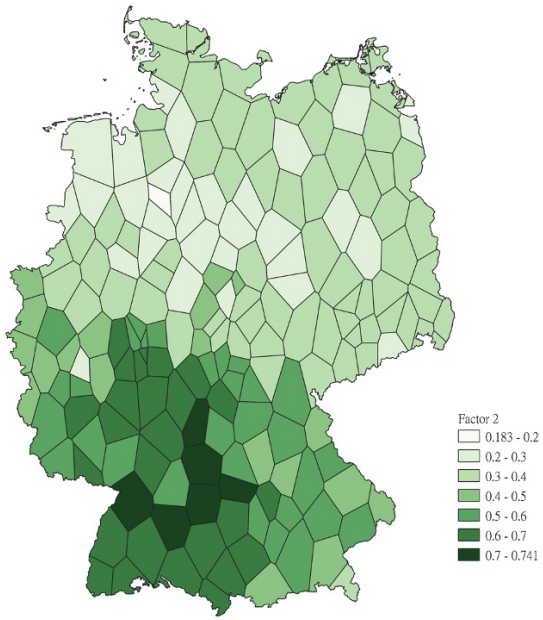
- Explained Variance: 21.13%
- Region: Low German, Low Franconian, North Frisian
- Top 6 features:
  - 354_Gefahren: 0
  - 744_seCHs: 0
  - 314_füNf: 0
  - 361_Gefallen: 0
  - 857_waCHsen: 0
  - 129_bleiB: f

Dialect Feature Detection

FA: Factor 2



Equal Quantile

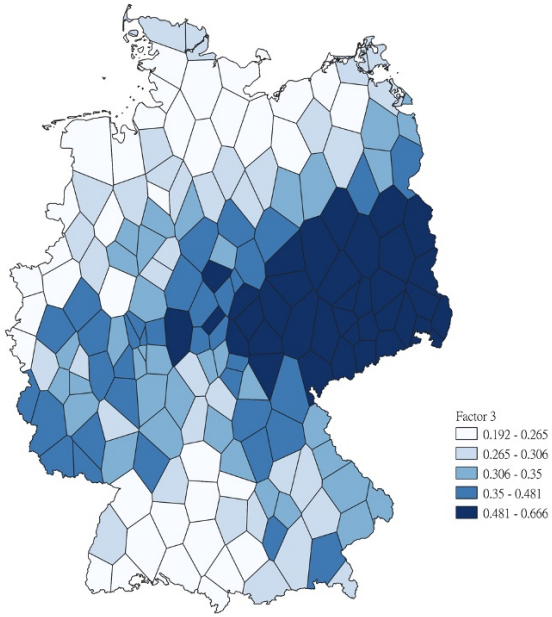


Pretty Breaks

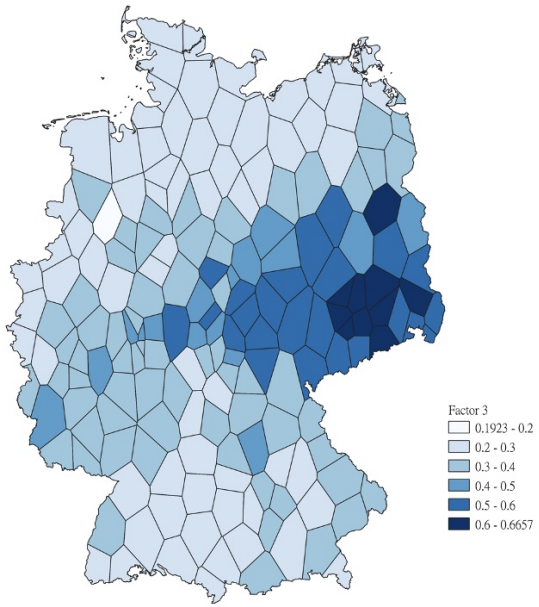
- Explained Variance: 20.55%
- Region: Swabian/ Alemannic, North Bavarian, Rhein Franconian and Hessian
- Top 6 features:
  - 479_herzeN: 0
  - 426_gestorbeN: 0
  - 952_zeiteN: 0
  - 923_wocheN?: 0
  - 343_gebleibeN: 0
  - 663_ochseN: 0

Dialect Feature Detection

FA: Factor 3



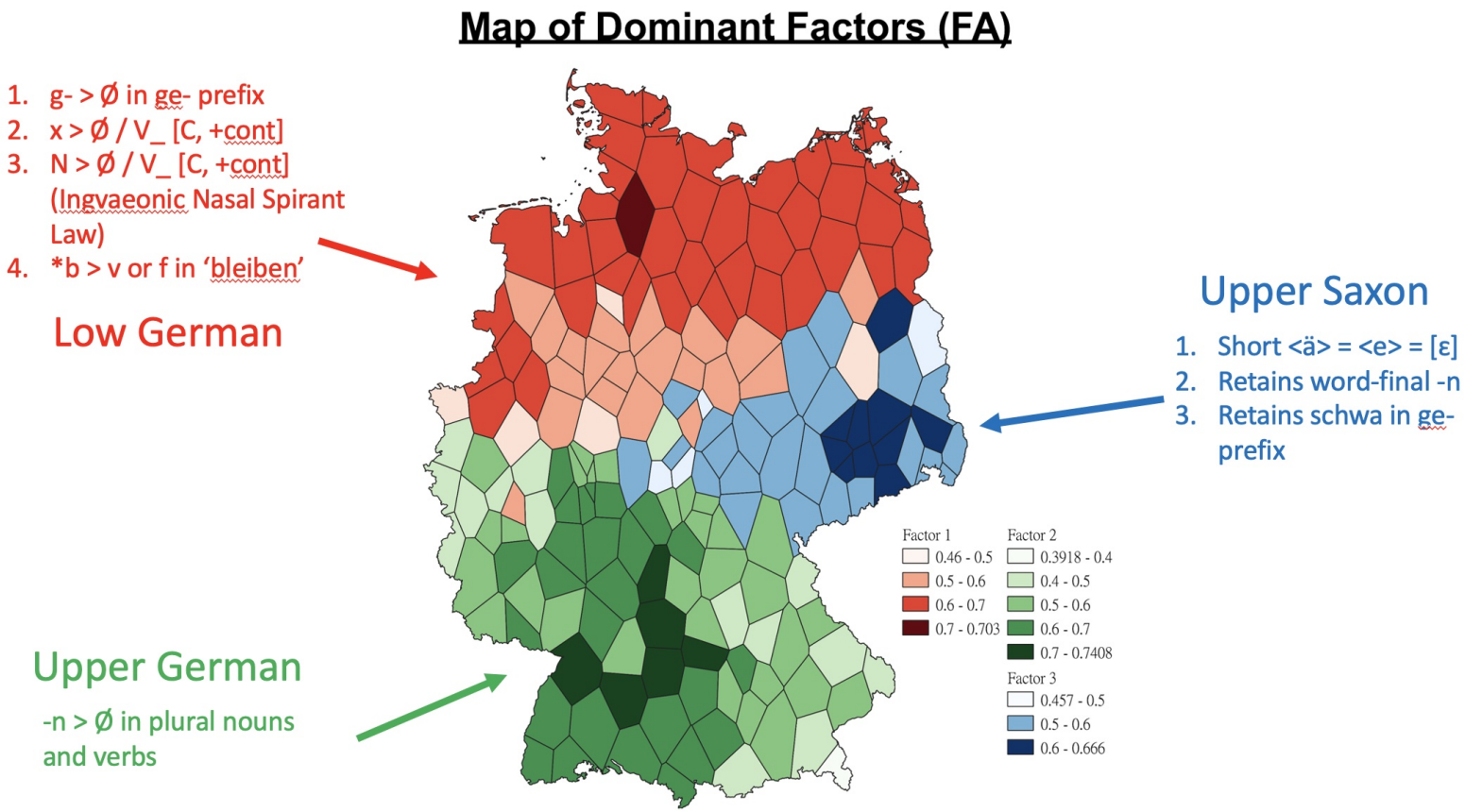
Equal Quantile



Pretty Breaks

- Explained Variance: 14.61%
- Region: Upper Saxon and Thüringian
- Top 6 features:
  - 655_neuN: n
  - 321_gÄnsen: ɛ
  - 102_bEsser: ɛ
  - 370_gEfunden: ə
  - 229_gEschlafen: ə
  - 92_beißeN: n

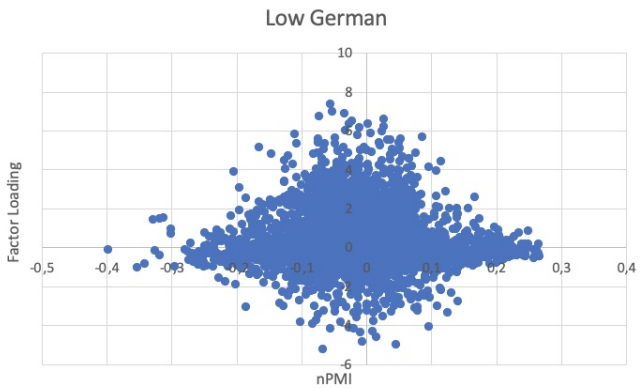
FA: Dialect groups

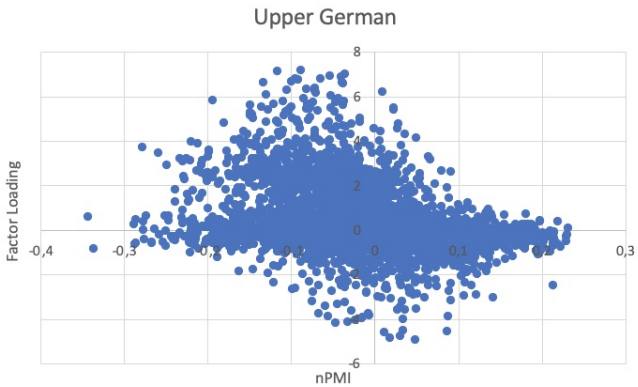


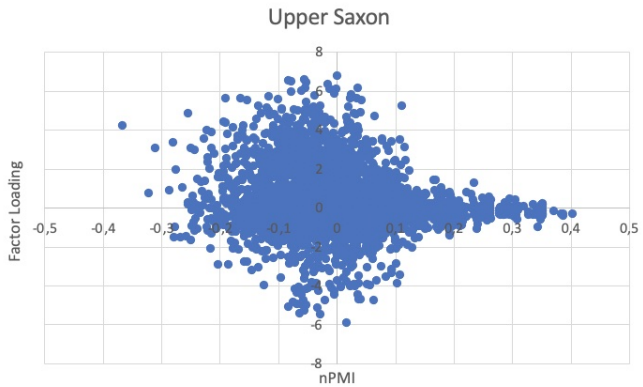
Which method is better?

- Representativeness (the proportion of the variant found within the cluster)
- Distinctiveness/Exclusivity (is the variant found only within the cluster)
- Pool of Variation (how many variants there are for this feature)

Correlation

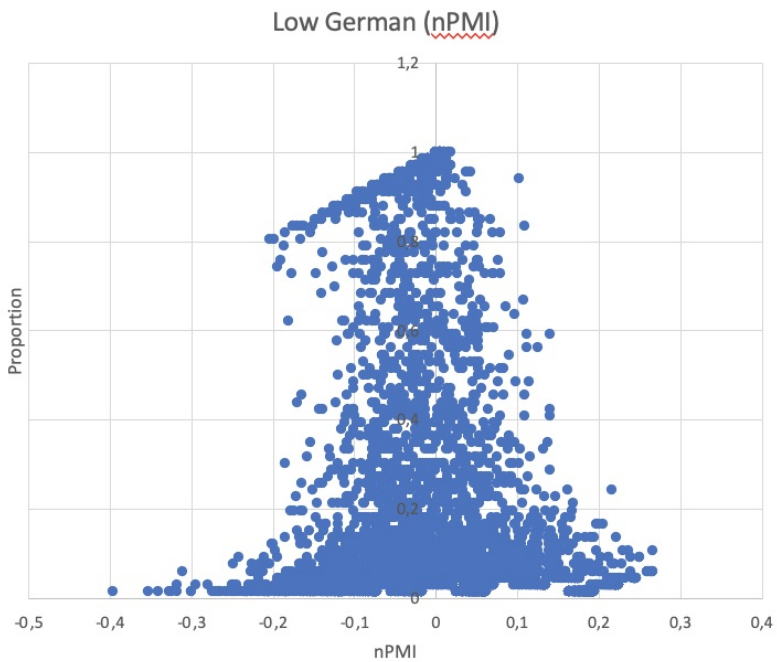


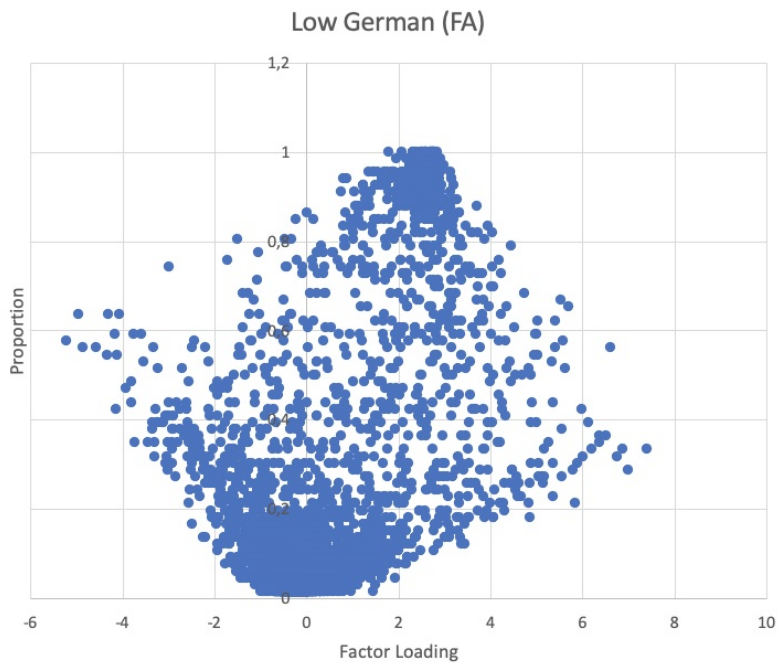




Dialect Feature Detection

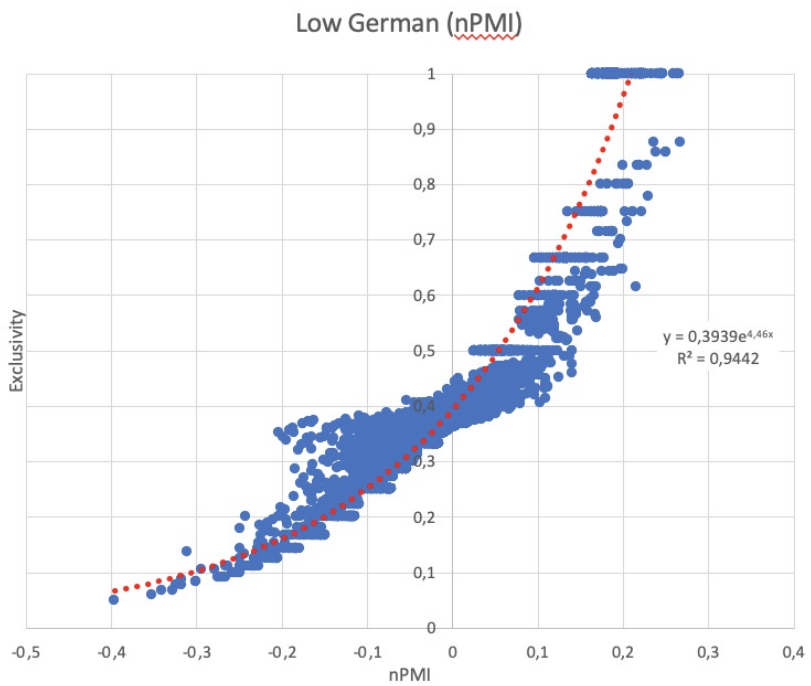
Representativeness

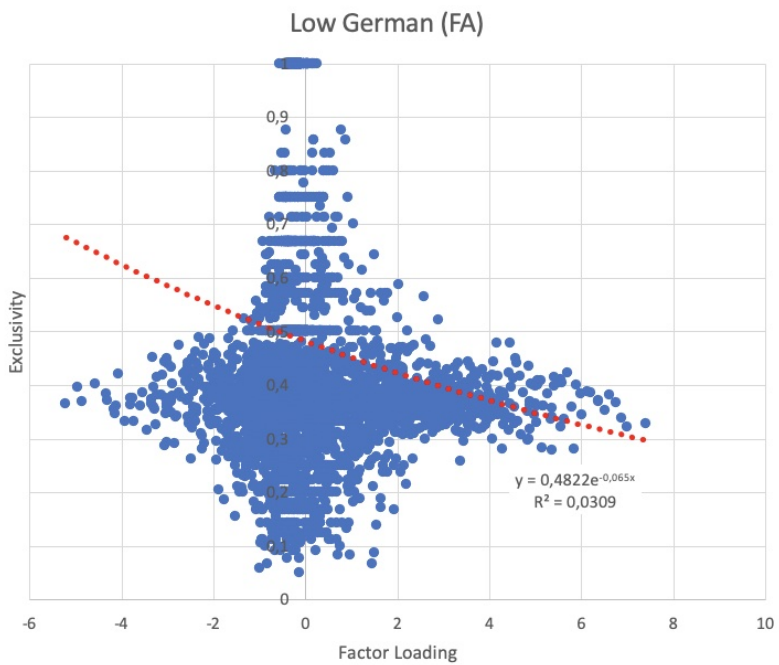




Dialect Feature Detection

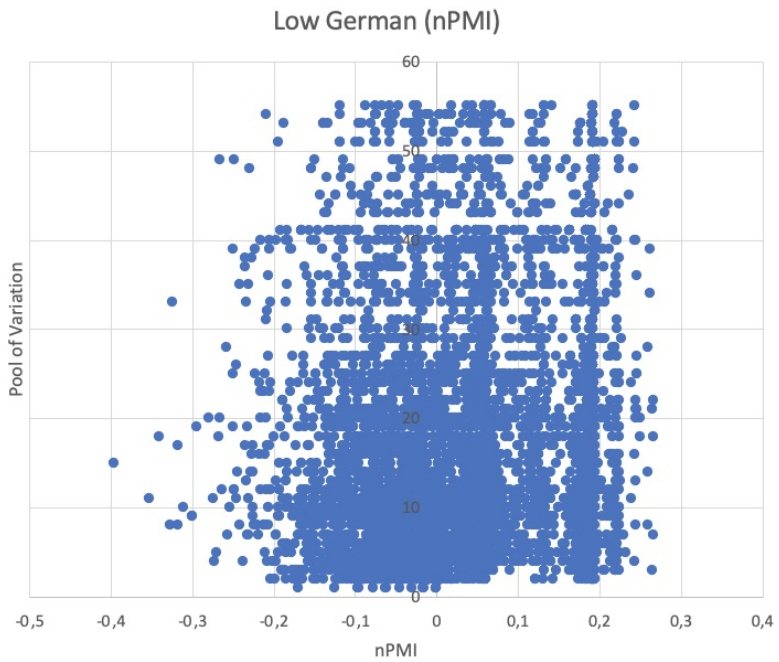
Exclusivity

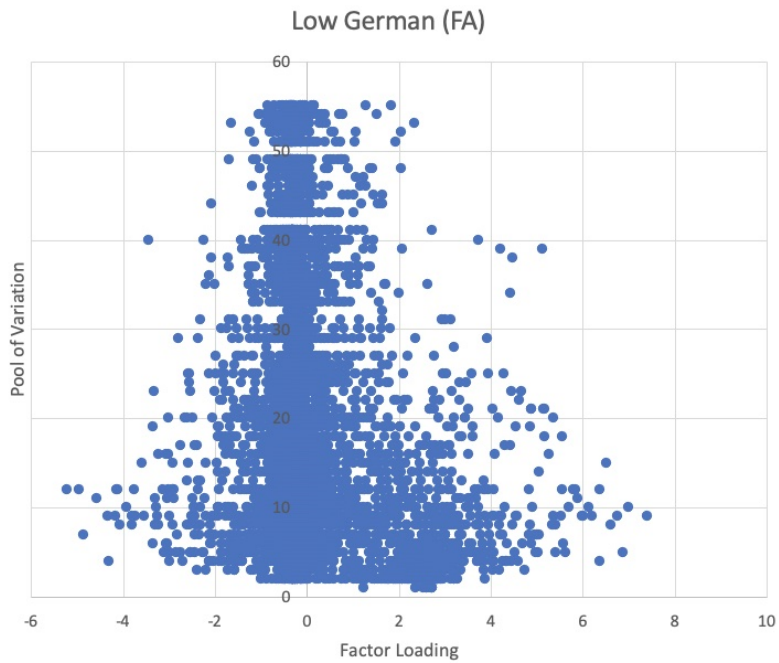




Dialect Feature Detection

Feature Variation





Observations

- There is no clear correlation between PMI and FA scores
- Representativeness
  - PMI: features with highest scores are only found in a small subset of dialects within the group, which suggests some very localized features being detected
  - FA: tends to detect features which are used by more dialects in the cluster, features which are more 'supra-'regional

Observations

- Exclusivity:
  - PMI: shows a clear recurring sub-linear curve; the higher the nPMI score, the more exclusive the feature is
  - FA: the most exclusive features are not found with features with a high factor loading, which suggests perhaps there is yet another parameter (not found yet) which FA relies on
- Pool of variation
  - PMI: there is no clear pattern here
  - FA: features with high factor loading tend to be features with a smaller pool of variation

Thank you!