Detecting Dialect Features Using Normalised Pointwise Information

H. W. Matthew Sung[*] H.W.M.SUNG@HUM.LEIDENUNIV.NL
Jelena Prokić[*] J.PROKIC@HUM.LEIDENUNIV.NL

[*] *Leiden University The Netherlands*

Abstract

Feature extraction refers to the identification of important features which differentiate one dialect group from another. It is an important step in understanding the dialectal variation, a step which has traditionally been done manually. However, manual extraction of important features is susceptible to the following problems, namely it is a time-consuming task; there is a risk of overlooking certain features and lastly, every analyst can come up with a different set of features. In this paper we compare two earlier automatic approaches to dialect feature extraction, namely Factor Analysis (Pickl 2016) and Prokić et al.'s (2012) method based on Fisher's Linear Discriminant. We also introduce a new method based on Normalised Pointwise Mutual Information (nPMI), which outperforms other methods on the tested data set.

1. Introduction

Feature extraction refers to the identification of important features which differentiate one dialect group from another. It is an important step in understanding the dialectal variation, a step which has traditionally been done manually. However, manual extraction of important features is susceptible to the following problems, namely it is a time-consuming task; there is a risk of overlooking certain features and lastly, every analyst can come up with a different set of features. This paper aims to fill this gap by proposing a novel application of Normalised Pointwise Mutual Information on a set of German dialect data, as well as comparing the results of this method with two previous approaches, namely Factor Analysis (Pickl 2016) and the method introduced in Prokić et al. (2012).

Dialectology, the study of geographical variation of language, is a century-long discipline in Linguistics, pioneered in Germany and France, later spread all over Europe, and eventually to various corners of the world, including Asia (e.g. China, Japan) and South America (e.g. Brazil, Mexico). Traditionally, to study geographical variation, dialectologists used to rely on plotting carefully selected set of dialect features on physical maps, drawing isoglosses and identifying dialect areas. In the 1970s, there was a change in the discipline. Instead of plotting isoglosses to study dialect variation in Gascogne, Séguy (1971, 1973) calculated distances between neighbouring localities by counting number of dissimilar dialect features between each neighbouring pair of localities, in order to investigate patterns of regional linguistic variation. Séguy's methodology was innovative, since he "conjuncted the existence underlying primary order which could be explored and expressed in quantitative terms" from the "apparent chaos of the [linguistic atlas] data" (Goebl 2018, 128). This methodology pioneered by Séguy was coined *dialectometry.*

In the past 40 years, computers have been introduced into dialectometry, and thus the field could do much more than what Séguy initially did. Goebl (1982, 1984) developed a way to calculate dialect similarity through *Relative Identity Value*, and he also introduced cluster analysis into dialectometry in order to find dialect groups automatically. Following Kessler's (1995) application of Levenshtein distance on Irish dialect data, Nerbonne and colleagues further expanded this methodology together with new analytic and visualization techniques, namely Multidimensional Scaling (MDS) and MDS map. These methods became very important tools for dialectologists to understand the grouping of dialects on the aggregate level.

The results generated by the methods mentioned above requires the analysts' own interpretation and expertise in order to understand the underlying linguistic factors responsible for clusters automatically detected by the algorithms. Clustering and multidimensional scaling rely on distance matrices, which do not offer any details or explanations of the identified dialect partitions. Namely, during the conversion from the qualitative dialect data to dialect distances the information on linguistic features is lost.

To overcome the problem, several approaches have been proposed to extract features responsible for the variation present in the dialect classification. One of them being Prokić et al.'s (2012) method based on Fisher's Linear Discriminant, which tries to seek features which have the biggest difference between the average distance within the cluster and the average distance between the clusters. Another approach proposed by Pickl (2016) relies on a dimensionality reduction technique which proceeds from features rather than distance matrices, namely Factor Analysis. Up until now there is no systematic comparison of various feature extraction methods. This paper aims to fill this gap by comparing two mentioned feature extraction methods on the same dialect data. Additionally, we will propose a new feature extraction method based on *Normalised Pointwise Mutual Information* (*nPMI*).

This paper is structured as follows. In Section 2 we present a literature review of the previous feature extraction methods. Sections 3 and 4 contain the description of the data and our research questions respectively. General procedures for all three methods are described in Section 5. In addition, in this section nPMI is introduced in more detail. The results of the comparison of three methods and evaluation results are presented in Section 6. Lastly, differences between three methods and further uses of the feature extraction methods are discussed in Section 7.

2. Previous approaches

Previous approaches in dialectometry which attempt to identify features characteristic for dialect groups include Pickl (2016) and Prokić et al. (2012). These approaches can be divided into two categories: bottom-up approaches (e.g. Pickl 2016) and top-down (e.g. Prokić et al. 2012). Bottom-up approaches seek simultaneously the dialect groups and distinctive features, whereas top-down approaches require a pre-defined dialect classification before features can be extracted.

2.1 Bottom-up approaches

One of the more common bottom-down approaches in dialectometry is Factor Analysis (FA), which has been used in exploring dialect areas and their characteristic features (Pickl 2013, Pröll 2015, Pickl 2016). FA is a dimensionality reduction technique, like Multidimensional Scaling (MDS, Borg and Groenen 2005), used in dialectometry (Embleton 1993, Heeringa 2004) to identify dialect groups. It condenses the variation of the categories in the data into a smaller number of patterns, or underlying factors, by grouping variants that "co-occur with a high frequency" (Pickl 2016, 82). Within the dialectal context, FA detects (gradient) membership of dialect areas (factors) and at the same time finds out which features contribute to the make-up of these groups and their respective strength of association to the group. The features can be lexical, phonetic, morphological or combined, as long as they are categorical (see Pröll 2015).

Unlike top-down approaches, FA does not require predefined groups. Pickl (2016) has argued that dialect areas are 'fuzzy', and FA can capture this fuzziness by identifying condensations of co-occurring variants instead of hard clusters. The *Factor Loading* is one of the biproducts when using FA, which indicates the relationship between the Factor (dialect group) and the location. Unlike traditional assumptions or representations of dialect areas, where within the group dialects are rather homogenous, each location has a different degree of factor loading for each dialect group.

Furthermore, Pickl (2016) illustrates the use of *Combined Factor Maps*, which represents the dominant factors, or strongest dialect group that each locality is associated with, yielding the highest concentrations of each dialect area (the 'surface dialect landscape'), like the map below in Figure 1.

Another biproducts of FA are *Factor Scores*. The dialect features most associated with a particular factor can be extracted based on the factor scores of the variant. The higher the factor score, the more it is associated with the respective factor.

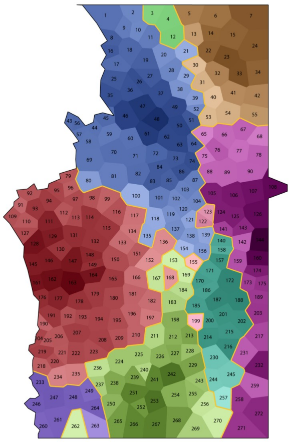


Figure 1: Combined Factor Map of the Sprachatlas von Bayerisch-Schwaben (SBS) survey sites (from Pickl 2016)

Other uses of FA include Nerbonne's (2006) feature identification of American English dialects recorded in the *Linguistic Atlas of the Middle and South Atlantic States (LAMSAS)*, which uses vowel features and Grieve's (2014) analysis of American English vowels using vowel formants[1].

A similar approach to FA is Principal Component Analysis (PCA). PCA is a set of mathematical procedures that seeks and groups sets of variables that strongly (positively or negatively) correlate to each other (Shackleton, Jr. 2005, 141). The analysis returns 'principal components', axes which group variables on the two poles, one being large positive values and the other being large negative values. The first principal component accounts for the most variance from the dataset, and the second principal component accounts for less variance than the first principal component. For each principal component, PCA could also reveal clusters of dialects and isolate sets of features that tend to co-occur. Shackleton, Jr. (2005) has illustrated the use of PCA for identifying dialect features between British English and American English speakers, and Leinonen (2010) has applied PCA on the Swedish vowel formants in order to investigate dialect levelling.

Lastly, the use of bipartite spectral graph partitioning by Wieling and Nerbonne (2011) can determine dialect groups and their sound correspondences simultaneously. The result returns clusters which include both the dialect groups and their respective sound correspondences together. The importance of the sound correspondences is then calculated post-hoc by taking the average of the Representativeness and Distinctiveness indices given in Wieling and Nerbonne (2011, 707).

[1] The vowel formants were pre-processed by a conversion into the Getis-Ord Gi z-score, an index for local spatial autocorrelation (Ord and Getis 1995). Local spatial autocorrelation calculates the degree a location is part of a high or low (formant 1 or formant 2) value cluster. The use of the Getis-Ord Gi z-score acts as a smoothing technique so that "the values of the smoothed variables only represent the underlying regional signals in the raw values of these variables" (Grieve 2014, 74).

2.2 Top-down approaches

There have been far fewer works done with the top-down approach in dialectometry. Prokić et al.'s (2012) method is a representation of this approach. Prokić et al. (2012) seeks features that differ little within a pre-defined group but differ enormously outside the group. This method was inspired by Fisher's Linear Discriminant (FLD), and since the authors did not give a name for this approach, we will address the method as FLD throughout the paper. The way it is done is by calculating the mean distance of a particular feature among all the pairs of dialects within the pre-defined group and outside the group with Levenshtein distance (Levenshtein 1966, Heeringa 2004). The pre-defined groups can be obtained by using cluster analysis (Prokić and Nerbonne 2008). Next, characteristic features are identified by seeking features with the largest differences between the within-group and outside-group differences.

The calculation of the within-group and between-group distance for one locality is illustrated in Figure 2. S represents a locality within the pre-defined group and the arrows represent the distances measured, which includes the pre-defined group (in blue) and the rest of the dialects outside the group (in yellow). This procedure is iterated for all the sites.

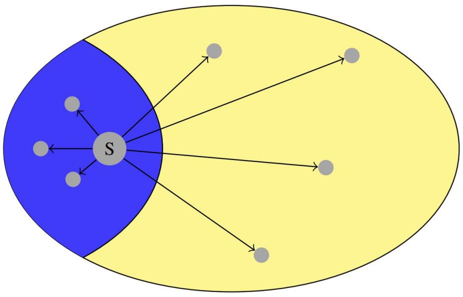


Figure 2: Illustration of distance calculation for the FLD method (from Prokić et al. 2012)

FLD has previously been tested on Dutch and German (Prokić et al. 2012). Both analyses have identified features which are rather homogenous within the pre-defined group. The authors have also found that the same word can show up as distinctive for more than one pre-defined group (with different variants). Lastly, this method is applicable to any feature type which can be defined with a numerical distance metric between elements. This includes words (as analyzed in Prokić et al. 2012), categorical data or vowels formants.

3. Data

The data used in this study comes from the *Phonetischer Atlas der Bundesrepublik Deutschland* (*PAD*), which was collected as a part of the project "Kleiner Deutscher Lautatlas – Phonetik" [Small German Sound Atlas – Phonetics] (Göschel 1992). The data were collected between the 1960s and 1970s. It consists of phonetic transcription for 201 words (based on the Wenker Sentences), all transcribed in International Phonetic Alphabet (IPA), from 182 survey locations across the Federal Republic of Germany. The map of localities can be found in Figure 3. In addition, the IPA transcriptions were further multi-aligned (Prokić et al., 2009; see more in Section 5.1) for the current study.

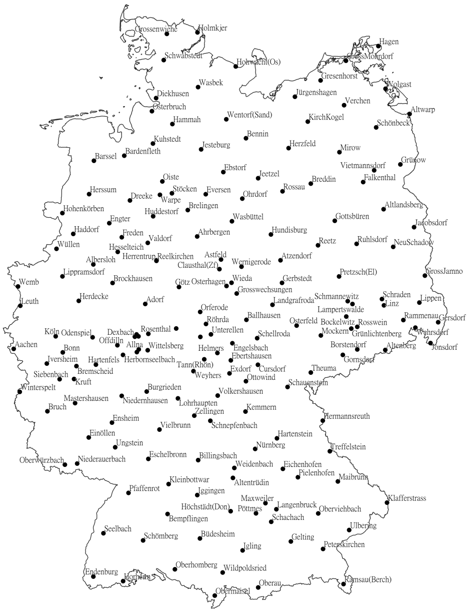


Figure 3: Locality map of the PAD

4. Research questions

Dialectologists study regional differences in a language and are interested in finding features distinctive to different dialect areas. While the aggregate approach in dialectometry gives us a partition of dialect groups based on the large amounts of data, it gives us little information about the linguistic structural characteristic for each group. This kind of information is very important for theoretical linguistics, but also for more applied tasks of dialect and speaker identification. Although there have been some approaches to this problem, there has not been a systematic comparison of these approaches.

Our goal in this paper is to compare the three feature extraction methods, namely FA, FLD and Normalised Point-wise Mutual Information (nPMI, a new top-down method which will be introduced in Section 5.3.3 below) and find out which method can find the most representative and distinctive features for each dialect group most reliably. The first two methods were chosen based on their accessibility and the fact that they can be used for the analysis of categorical data. FA and FLD are implemented in *GeoLing* and *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016) respectively, and are freely available.

5. General Procedures

Although general procedures for applying these three methods differ, they are all suitable for working with categorical data, i.e. dialect features at different linguistic levels.

In our study, we have chosen to use multi-aligned word transcriptions (Section 5.1) of dialect data for the analysis. On the one hand, this is a data format which all methods can process. On the other hand, multi-aligned data can retain all phonetic features in the data without having to manually pick a subset of the features. In previous dialectometric studies with FA, only chosen features instead of all phonetic segments from a list of words were used (e.g. Pickl 2016). Combining FA and multi-aligned phonetic data is another innovation introduced in this paper. Multiple aligned pronunciation data has previously been combined with FLD in Prokić and Nerbonne (2013).

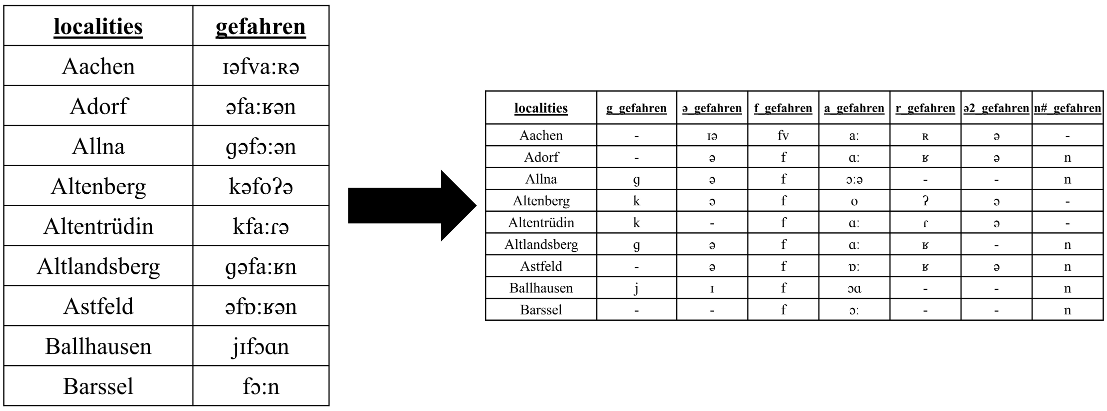
Regarding dialect classification step (Section 5.2), FA is a bottom-up approach, and does not require pre-classified dialect groups in order to extract features. FLD and nPMI are top-down approaches which start from already identified dialect groups in order to extract features.

Dialect features identified by each of the three methods are evaluated using two measures of quantity, namely *Exclusivity* and *Representativeness* (see Section 5.4). We additionally make use of distribution maps in our evaluation procedure (Section 5.3).

5.1 Multiple Sequence Alignment

In this study we use phonetic transcriptions of 201 words collected from 182 locations. We multi-align all pronunciations of each word in the data set using the LingPy library (List and Forkel 2021). As a result, each column in the Multiple Sequence Alignment (MSA hereafter, Prokić et al. 2009) comprises of individual segments, i.e. consonants and vowels (monophthongs and diphthongs are counted as single vowel).

In MSA, each column contains different variants (different phonetic realisations) of the same variable (consonant or vowel phonemes or individual segments of a word), which we call a dialect 'feature'. This step has a lot in common with the way one establishes correspondences with the comparative method. An illustration can be found in Figure 4. The multi-aligned data is then manually checked for potential misaligned segments. The MSA data is used as an input for all three feature extraction methods.



localities	gefahren
Aachen	ɪəfvaːʀə
Adorf	əfaːʁən
Allna	gəfɔːən
Altenberg	kəfoʔə
Altentrüdin	kfaːɾə
Altlandsberg	gəfaːʁn
Astfeld	əfɒːʁən
Ballhausen	jɪfɔɑn
Barssel	fɔːn

localities	g_gefahren	ə_gefahren	f_gefahren	a_gefahren	r_gefahren	ə2_gefahren	n#_gefahren
Aachen	-	ɪə	fv	aː	ʀ	ə	-
Adorf	-	ə	f	ɑː	ʁ	ə	n
Allna	g	ə	f	ɔːə	-	-	n
Altenberg	k	ə	f	o	ʔ	ə	-
Altentrüdin	k	-	f	ɑː	ɾ	ə	-
Altlandsberg	g	ə	f	ɑː	ʁ	-	n
Astfeld	-	ə	f	ɒː	ʁ	ə	n
Ballhausen	j	ɪ	f	ɔɑ	-	-	n
Barssel	-	-	f	ɔː	-	-	n

Figure 4: Illustration of Multiple Sequence Alignment

5.2 Dialect classification

The following procedures only apply to FLD and nPMI. The classification of dialects on the dialects in PAD requires two steps: distance calculation and cluster analysis.

5.2.1 Distance calculation

This procedure is a step which transforms qualitative data, i.e. columns of phonetic realisations of different segments of words in the multi-aligned PAD data, into quantitative data, i.e. dialect distances.

The calculation is based on the inverse value of *Relative Identity Value* (*RIV*, Goebl, 1982, 1984), also known as the *Relative Distance Value* (*RDV*, Goebl 2018). The formula for RDV is provided in (1) below. To calculate RIV in a pairwise comparison (between two dialects), the number of matching features/ Co-identity (COI in (1)) is divided by the total number of features compared, i.e. number of matching columns (COI) plus number of unmatching columns/ Co-difference (COD in (1)). The resulting value is the distance of the dialect pair ranging from 0 to 1. The RDV is 1 − RIV, which gives the pairwise distance instead of similarity.

$$RDV_{jk} = 1 - \frac{\sum COI_{jk}}{\sum COI_{jk} + \sum COD_{jk}} \quad or \quad 1 - \frac{no.\ of\ shared\ features\ in\ both\ dialects}{total\ number\ of\ features\ compared} \tag{1}$$

The calculation of RDV was applied to all the dialect pairs in the PAD data, yielding 16471 pairwise distances. These distances are stored in a distance matrix and analysed by means of cluster analysis.

5.2.2 Cluster analysis

Cluster analysis refers to the partition of objects (dialects in our case) into groups (Manning and Schütze 1999). The application of cluster analysis helps us to identify dialects which are similar enough to be considered as the same group. There are many algorithms used in cluster analysis, but the most commonly used algorithms in dialectometry are the so-called agglomerative hierarchical clustering algorithms. These algorithms find successive clusters based on previously established clusters, creating a hierarchical representation of the clusters in a dataset.

For our study, we have chosen Ward's method (Ward 1963), which is also known as the minimal variance method. This cluster algorithm merges clusters which will yield the smallest increase in the sum of the square distances of each element from its cluster's mean. The choice of this cluster algorithm is based on Prokić and Nerbonne's (2008) evaluation of several cluster algorithms, which shows that Ward's method yields relatively good performance in external validation using *Modified Rand Index* (Hubert and Arabie 1985) and *Entropy* (Zhao and Karypis 2001).

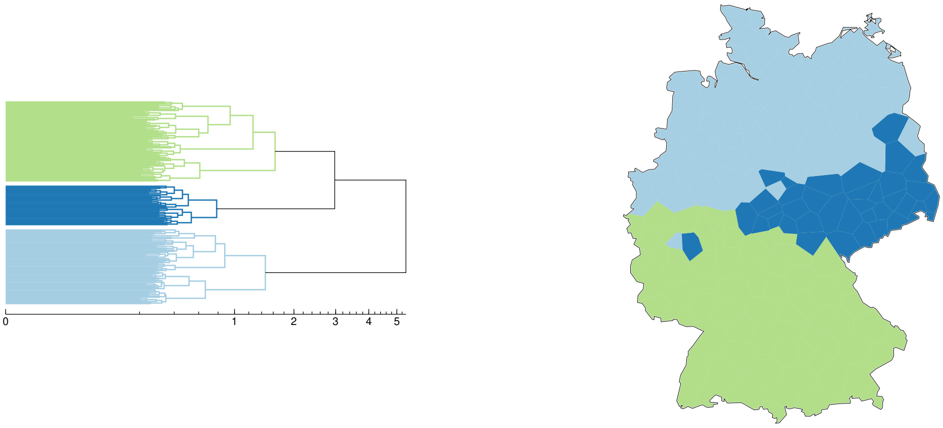


Figure 5: On the left: Dendrogram of German Dialect Distances in PAD. On the right: Cluster map of German dialects in PAD

We have chosen the 3-cluster solution for our data, as illustrated in Figure 5. These three clusters overlap largely with some of the traditional German dialect areas (Wiesinger 1983). For instance, the light blue cluster in Northern Germany largely overlaps with the Low German area. We refer to this cluster as 'Low German' in our analysis. The dark blue cluster covers the Upper Saxon and the Thuringian area in Wiesinger's (1983) classification. Since there are more Upper Saxon dialects in this region, we label this cluster as 'Upper Saxon' in our analysis. Lastly, the light green cluster covers the western Central German dialect as well as Upper German dialect area. We decided it is appropriate to call this cluster 'Southern German' based on the geographical proximity of the cluster. These cluster groups will be used as the labels for FLD and nPMI.

5.3 Feature extraction

5.3.1 Factor analysis

As previously mentioned, FA identifies condensations of co-occurring variants and returns factor loadings for each location, which indicate the relationship between the locality and the factor, as well as the factor scores, which indicate the amount of association each variant of a feature has to the factors. The relationship between factors, factor loadings, factor scores and each locality (local dialect) under factor analysis (based on Pickl 2016) is illustrated in Figure 6 below.

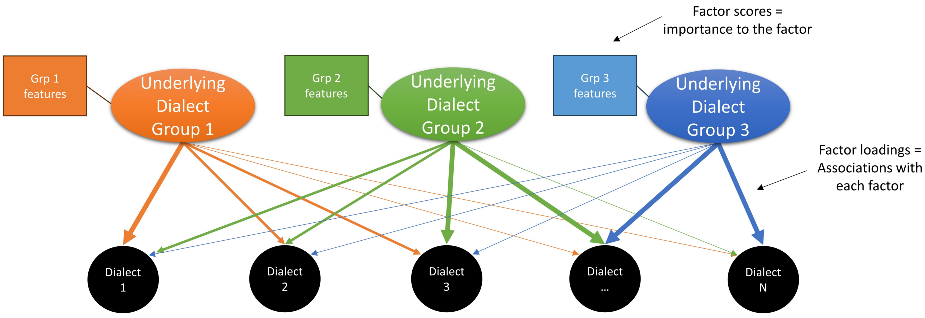


Figure 6: Illustration of Factor Analysis in Dialectometry

The number of factors has to be determined by the analyst(s), and we have decided to set it as 3 factors for FA to extract so that all three methods are comparable with each other. The factor maps for the top 3 factors are presented in Figure 7. In addition, we have identified the dominant factors for all the localities, and thus created a combined factor map, following Pickl (2016). The evaluation of the feature extraction methods for FA will be based on the combined factor map, which can be found in Figure 8. Since the geographical distribution of the dominant factors in Figure 8 is very similar to the cluster analysis, we will refer to each factor (dialect group) with the same labels as the cluster analysis (Section 5.2.2).

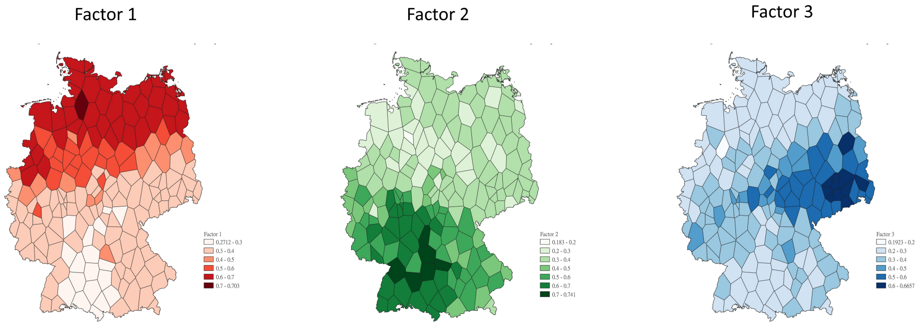


Figure 7: Maps of Factor Loadings per Factor

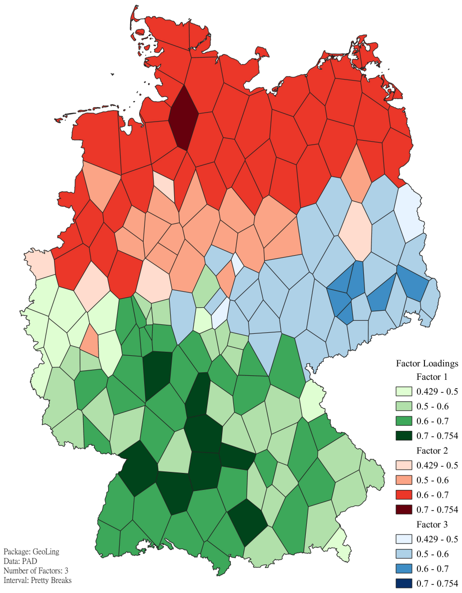


Figure 8: Combined Factor Map of the PAD localities

5.3.2 Fischer's Linear Discriminant

With respect to the variances, Fischer's Linear Discriminant maximizes the differences in the means between two datasets (Schalkoff 1992, 90). Prokić et al's (2012) version of FLD for feature extraction works in a similar fashion.

Let a group of dialect be g, and the number of members be $|g|$, while the area that is not part of group g is G, and the members of G is $|G|$. Site s include sites both within and outside g. For each feature *f*, distance d is calculated between a site s within the cluster and all other sites *s'* within the cluster. We iterate the calculation for all the sites within the cluster and take the mean difference in order to get a within group difference d_f^g for a feature. This process is iterated for all the features (columns) in the dataset. The formula for the mean within-group difference is given in (2).

$$\bar{d}_f^g = \frac{2}{|d|^2 - |d|} \sum_{s,s' \in g} d_f(s, s') \tag{2}$$

The mean difference between a site s and sites outside the cluster d_f^g is calculated with the formula in (3).

$$\bar{d}_f^G = \frac{1}{|g|(|G| - |g|)} \sum_{s \in g, s' \in G} d_f(s, s') \tag{3}$$

The calculated with- and between-group differences are then standardized into z-scores. We use *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016) to calculate the ranking of the features by *between-score* minus *within-score*, where the largest difference is ranked the highest, and the smallest difference is ranked lowest.

5.3.3 Normalised Pointwise Mutual Information

Pointwise Mutual Information, or PMI, is an association measure based on probabilities and co-occurrence (Church and Hanks 1990). Originally used in detecting word association based on their co-occurrences, it has also gained popularity in dialectometry for the use of automatic detection of segmental distances (Wieling et al. 2009). The idea behind PMI is comparing the probability of observing two categories, x and y, together (joint probability) and independently (by chance). The assumption is that if there is genuine association between x and y, then the joint probability would be much greater than their probability together by chance (Church and Hanks 1990, 23).

The required probabilities (how often certain element occurs within the respective column) include probability of variant x or *p(x)* found within one column in the MSA data; the probability of dialect group y or *p(y)* within the classification column based on the cluster analysis in Section 5.2.2. Lastly, the final probability required is the co-occurrence of variant x given dialect group y, or *p(x,y)*.

$$pmi(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

The PMI scores are calculated using formula given in (4) (Church and Hanks 1990), which is the log base 2 of the probability of the co-occurrence of a given variant and a given group, out of all the possible instances that they could co-occur in the data. We normalise all scores based on Bouma (2009), presented in (5).

$$npmi(x,y) = \frac{pmi(x,y)}{-log_2 p(x,y)} \tag{5}$$

The steps described above are iterated for all the variants found in the same column, and for each dialect group in the classification label column. When the nPMI score for all the variants and dialect groups have been processed for the first column, the same procedures are iterated until the last column of the MSA data has been processed.

5.4 Exclusivity and Representativeness

In order to evaluate how characteristic extracted features are for each dialect group, we have utilised two measures, namely *Exclusivity* and *Representativeness*, for our evaluation[2].

Exclusivity concerns the extent to which a specific variant is only found within the given cluster. To formalise this idea, we have used formula (6) to calculate *Exclusivity* for our extracted features, using the notations from the FLD formulae in (2) and (3).

$$\frac{|g_f|}{|g_f| + |G_f|} \quad or \quad \frac{no.\ of\ dialects\ with\ the\ variant\ in\ the\ cluster}{total\ no.\ of\ dialects\ with\ the\ feature\ in\ the\ data} \tag{6}$$

Representativeness on the other hand calculates the number of dialects within the cluster which has the specific variant. The formalisation of *Representativeness* is presented in formula (7). This formula was adopted from Wieling and Nerbonne (2011).

$$\frac{|g_f|}{|g|} \quad or \quad \frac{no.\ of\ dialects\ with\ the\ variant\ in\ the\ cluster}{total\ no\ of\ dialects\ in\ the\ data} \tag{7}$$

6. Results

Each of the three methods will be evaluated by the *Exclusivity* and *Representativeness* of the top 10 features[3], with the aid of distribution maps of these features in their respective dialect groups. The distribution maps depict the dialect area which the features were extracted from (in gray, see

[2] *Exclusivity* is different from *Distinctiveness* found in Wieling and Nerbonne (2011). We have compared the two indices and they are highly correlated, although the calculation for *Distinctiveness* is more computational heavy, hence we use *Exclusivity* instead.

the Base Map), as well as the areas which a specific feature is found, indicated by both colour and hatching patterns. The features for each colour/ pattern are explained in the legend. Furthermore, the maps also indicate each feature's *Exclusivity* and *Representativeness.* Because of space, only the first 6 maps will be shown in the appendix.

6.1 Factor Analysis

The top 10 features identified for the Southern German area are almost all related to the phenomenon of dropping final -n in verbs and plural nouns (see Appendix Map A). On the other hand, we mostly see the realization of a schwa in the ge- prefix, retention of final -n and short <e> and <ä> realized as [ɛ] in the Upper Saxon dialects (Appendix Map B). Lastly, for Low German, the most abundant feature identified is the loss of initial g- in the ge- prefix. Other features include the Ingvaeonic Nasal Spirant Law, assimilation of Germanic *-hs and lenition of Germanic *p were identified (See Appendix Map C), as well as non-palatalisation of the -s- before -t (in the word Wurst 'sausage'). Although some sound changes were known to be typical for their regions, to what extent are they distinctive to and representative to their respective regions?

Dialect Group	Exclusivity	Representativeness
Southern German	0.836	0.648
Upper Saxon	0.368	0.787
Low German	0.766	0.800

Table 1: Average Exclusivity and Representativeness scores for the Top 10 features (Factor Analysis)

The top 10 features identified for Southern German and Low German dialects show high exclusivity scores with an average of just over 0.75, as shown in Table 1. As illustrated in Figure 9, a representative map taken from the Southern German analysis, we can see that most of the dialects with the feature (i.e. dropping of word-final -n in Figure 9) are found within the gray area, which is the dominant Southern German area. There are, however, still a handful of dialects with the same feature found outside the gray area.

This is not the case for the Upper Saxon area, though. In Table 1, the average exclusivity score for Upper Saxon is much lower than the other two dialect groups, with only 0.368. To give an idea of what a low exclusivity score implies, we turn to Figure 10, which shows the retention of word-final -n as a representative map for the top 10 features for the Upper Saxon area. Unlike Figure 9, we can see that the dialects with the retention of -n can be found within the gray area, but also outside the gray area to a large extent, all over Germany. This means that the extracted feature is not very exclusive to the Upper Saxon region, hence a low exclusivity score. The same can be said for the rest of the top features for Upper Saxon.

In terms of representativeness, we can see that in the Southern German (Appendix Map A) and Low German area (Appendix Map C), the top identified variants have pretty much covered the whole gray region. The same can be said in the Upper Saxon area (Appendix Map B), the variants can pretty much be found within the gray region. These can be captured by the average score of representativeness of over 0.64 in Table 1.

The representativeness score for the Upper Saxon area seems indicate that FA performed quite well at extracting features that are representative for this area. However, this requires some more attention. As we have mentioned, the exclusivity performs poorly for the Upper Saxon area. This

[3] Dialect areas do not necessarily require 10 or another set number of features for us to make dialects distinguishable. When applying the elbow method on the nPMI values by the feature ranking, we do not see a clear cluster of top features, although we are able to identified the elbow with 6 to 10 features, depending on the dialect group. For our analysis, we have chosen 10 features so that we can consistently compare the methods with our metrics exclusivity and representativeness.

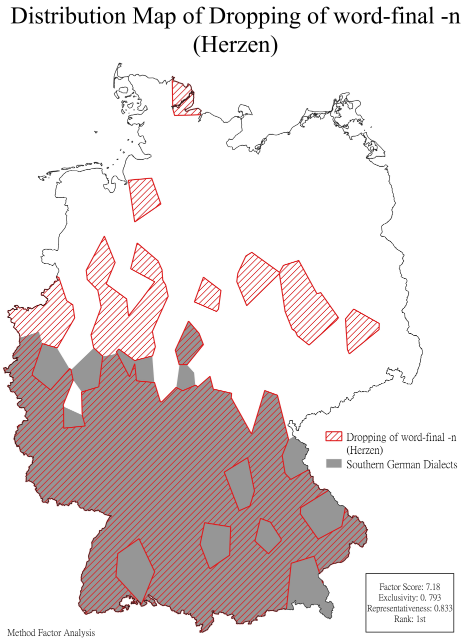


Figure 9: Distribution Map of Dropping of word-final -n (Herzen)

suggests that a high representativeness of dialects with the identified feature found in a dialect group does not imply the feature is being exclusive to the area.

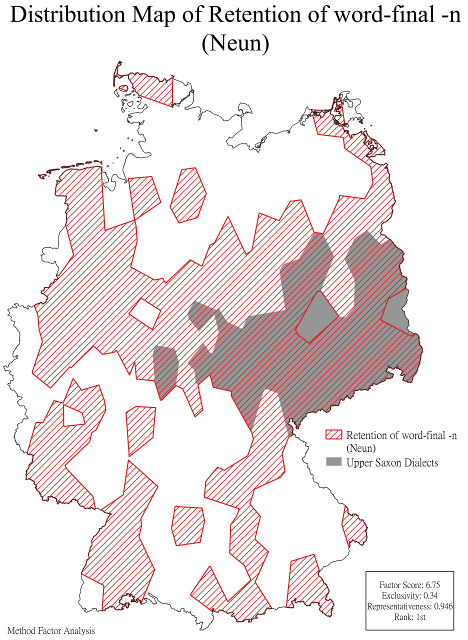


Figure 10: Distribution Map of Dropping of word-final -n (Neun)

6.2 Fisher's Linear Discriminant

The top 10 features identified for the Southern German area are mostly different from the ones extracted using Factor Analysis[4]. Although there are four features related to the loss of final -n, there are more features related to the High German Consonant Shift being detected, as well as pre-t palatalization (see Appendix Map D). Other than the retention of final -n and the pronunciation of short <e> and <ä> as in *besser* and *Gänsen*, FLD also identified some different features from FA for the Upper Saxon dialects. These features include the realization of a long u in *Bruder* (monophthongisation of MHG *uə) and short <ü> realized as [ʏ] (see Appendix Map E). For Low German, the vast majority of the top 10 features are loss of initial g- in the ge- prefix (see Appendix Map F). Other features include the assimilation of word-final -n > -m following labials, the Ingvaeonic Nasal Spirant Law, assimilation of Germanic *-hs and lenition of Germanic *p, like those found in Factor Analysis.

Dialect Group	**Exclusivity**	**Representativeness**
Southern German	0.713	0.809
Upper Saxon	0.423	0.764
Low German	0.781	0.777

Table 2: Average Exclusivity and Representativeness scores for the Top 10 features (Fischer's Linear Discriminant)

Comparing to the top 10 features from the Factor Analysis, the average exclusivity score of the Southern German area is lower. There is a slight increase of coverage of the non-gray area where Upper Saxon dialects are located. This is illustrated in a representative map in Figure 11. The average exclusivity score for the top 10 features of Upper Saxon is again low, like in the Factor Analysis. Also similar to those in FA, the distribution maps show that other than the first feature, the rest of the top 6 features (in Appendix Map E) are also found to a large extent in different parts of Germany, with the 4th feature (retention of -n in Neun) covering almost 2/3 of Germany.

For the Low German dialect area, Fisher's Linear Discriminant performs a lot better than the other two dialect groups at extracting exclusive features. This is the only group which has an average of over 0.78 in exclusivity. This is supported by the distribution maps of the top 6 features (Appendix Map F), where most dialects with the feature are found within the gray area. An example is given in Figure 12, which is a representative map from the Low German area.

The problem FA faced when dealing with the Upper Saxon area also surfaced for FLD. Again, although the Representativeness is relatively high in this region, the distribution maps (Appendix Map E) have illustrated again that a high Representativeness does not equate high Exclusivity, since these are all found in a much wider geographical area than the Upper Saxon region.

Lastly, FLD has the most success in identifying the most exclusive features in the Low German area, since the dialects with the top 6 extracted features (shown in Appendix Map F) are mostly found in the Low German area, with only limited number of exceptions compare to the other two dialect areas.

6.3 Normalised Pointwise Mutual Information

The top 10 features of the Southern German area extracted using nPMI almost overlap completely with FA, since these seven features are related to the loss of final -n in plural nouns and verbs. Furthermore, the retention of schwa in the word *Garten* also implies the loss of final -n, since these two features are often found together in the -en suffix (in plural nouns and infinitive of verbs),

[4] It should be noted that FLD does not return a precise value of the feature. This requires manual inspection of the distribution map in order to determine what the feature value is.

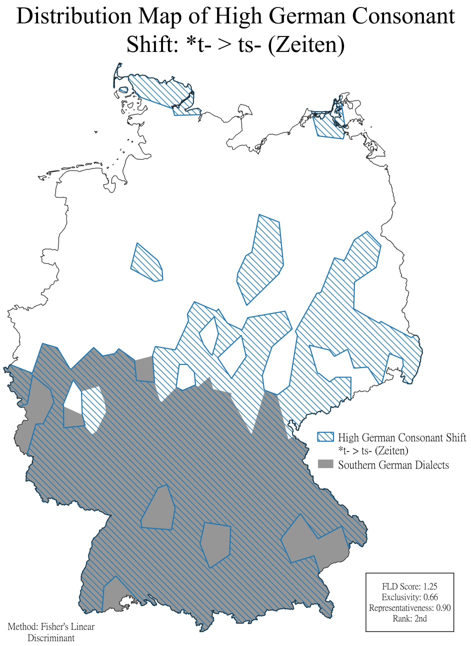

Distribution Map of High German Consonant Shift: *t- > ts- (Zeiten)
High German Consonant Shift *t- > ts- (Zeiten)
Southern German Dialects
FLD Score: 1.25
Exclusivity: 0.66
Representativeness: 0.90
Rank: 2nd
Method: Fisher's Linear Discriminant

Figure 11: Distribution Map of High German Consonant Shift (Zeiten)

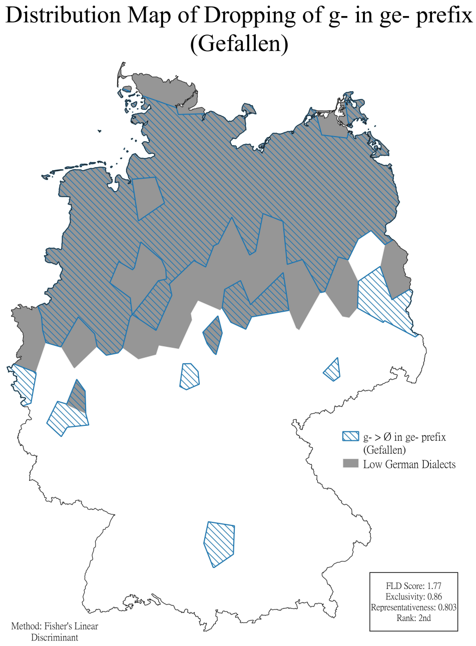

Distribution Map of Dropping of g- in ge- prefix (Gefallen)
g- > Ø in ge- prefix (Gefallen)
Low German Dialects
FLD Score: 1.77
Exclusivity: 0.86
Representativeness: 0.803
Rank: 2nd
Method: Fisher's Linear Discriminant

Figure 12: Distribution Map of High German Consonant Shift (Zeiten)

see Appendix Map G. Lastly, one more feature, the consonant in the ge- prefix, is realized with a voiceless k-.

For the Upper Saxon area, half of the features extracted are related to the contraction for the days of the week. Other features include features which resemble Standard German, including raising from Middle High German *ei to ai, retention of final -n and the umlaut of au <äu> is realized as [ɔɪ]. Some features, however, do not conform to the Standard. These features are spiratisation of pf- (which came from the High German Consonant Shift), as well as the Saxon back unrounding. These features are not detected by the previous two methods. Lastly, for Low German, there are overlaps with the features identified from FA, including loss of g- in the ge- suffix, Ingvaeonic Nasal Spirant Law, Assimilation of Germanic *-hs, and the unaffected segments from the High German Consonant Shift.

Dialect Group	Exclusivity	Representativeness
Southern German	0.848	0.647
Upper Saxon	0.840	0.365
Low German	0.831	0.769

Table 3: Average Exclusivity and Representativeness scores for the Top 10 features (Normalised Pointwise Mutual Information)

The average Exclusivity for the top 10 features extracted from all three groups are the highest among all three methods. Although there are still instances of the identified variants found outside the gray areas, all the top features have an exclusivity higher than 0.83. The distribution maps for nPMI can be found in Appendix Map G, H and I.

The previous two methods did not identify highly exclusive features for the Upper Saxon area. Unlike FA and FLD, we do not see a huge number of the identified variant outside the gray area. Instead, we see that most of the identified variants fall within the dialect area for Upper Saxon (Appendix Map H), as well as the Southern (Appendix Map G) and Low German dialects (Appendix Map I). A representative map (Figure 13) from the Upper Saxon area illustrates the difference between the nPMI result and results from FA and FLD.

The Representativeness scores of the features extracted with nPMI are not as high as FA and FLD for all dialect areas. The biggest difference in the average Representativeness score between the three methods is found in the Upper Saxon area, where with nPMI, it has a score of 0.365, whereas with other methods, it is above 0.76. However, the features extracted are indeed much more exclusive than the ones extracted with FA and FLD. nPMI seems to favour exclusivity more than FA and FLD, especially in cases like Upper Saxon. It appears to seek features with higher Exclusivity in the expense of Representativeness.

6.4 Evaluation with more features

Sections 6.1-6.3 presented the analysis of Exclusivity and Representativeness based on the top 10 features. When the number of features increases, we start to see changes with the average Exclusivity and Representativeness. Thus, it is useful to assess the characteristics of the features extracted with a varying number of features extracted with all these methods. In this case, we are comparing the average across all three dialect groups with top 10 features against all features with positive scores, by finding the mean across the three dialect groups.

Method	Exclusivity	Representativeness
FA	0.657	0.745
FLD	0.639	0.783
nPMI	0.839	0.594

Table 4: Average Exclusivity and Representativeness scores for each method (Top 10 features)

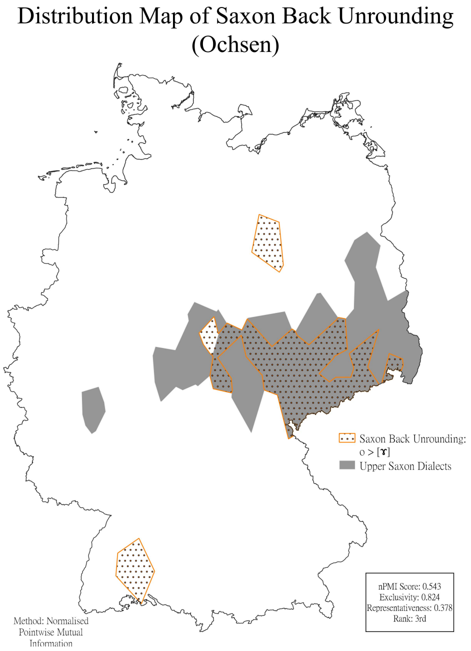


Figure 13: Distribution Map of Saxon Back Unrounding (Ochsen)

Method	**Exclusivity**	**Representativeness**
FA	0.556	0.175
FLD	n/a[5]	n/a[6]
nPMI	0.703	0.167

Table 5: Average Exclusivity and Representativeness scores for each method (All Features with Positive Values)

In terms of Exclusivity, as the number of features increases, it drops accordingly. However, despite the drop, nPMI still consistently remains the method which extracts the most exclusive features. On the other hand, FA has the highest Representativeness, (but we do not know whether FLD would yield a higher or lower Representativeness score than FA). Contrastively, nPMI has the lower Representativeness no matter what the number of extracted features is.

7. Discussion

7.1 Explanation

Different methods have shown their abilities to extract features with different degrees of Exclusivity and Representativeness. These parameters are dependent on the mathematical procedures used in each method.

[5] We could not extract values for FLD for all features with a positive FLD score because the result extracted from Gabmap does not indicate which variant is responsible given dialect group. To identify the variant, it requires manual inspection of the map as a following step. Without automation, we have only extracted top 30 variants per dialect group in our study manually.

[6] Same as previous footnote.

Factor Analysis tends to identify features with a higher Representativeness (see Tables 1 to 3), but not necessarily exclusive, as we have seen in the maps in Section 6.1. This could potentially be explained by the fact that FA's factors has a much larger geographical coverage than the 'dominant' factors as shown in the combined factor map in Figure 8, which only represent the surface dialect groups[7]. As we have seen from Figure 7, each factor can be found across the whole Germany, but the intensity of each factor (Factor Loading) varies. The combined factor map only shows the factor loading which is dominant for the particular locality, but in theory, the other factors are still present in the same dialect, only weaker, i.e. 'latent' (Pröll et al. 2014).

This is a possible explanation to why some variants identified to be related to a dialect group is found outside the 'dominant' area. Factor analysis might be a better model to explain dialect variation in terms of the overlay of different dialect areas (co-existing factors in one place), which shows the gradual change from one dialect area to another. On the other hand, when extracting features for a surface dialect group, FA may not be a suitable method.

Fischer's Linear Discriminant has been extracting features with an Exclusivity and a Representativeness score in between FA and nPMI. This could be explained by FLD's algorithm. FLD uses average distances within and between clusters, which leads to the extraction of features that hit the middle point of both distinctiveness and representativeness (represented by Exclusivity and Representativeness scores), i.e. the harmonic mean of the between-group and within-group distances.

Lastly, the algorithm of normalised Pointwise Mutual Information was designed to find association between categories, based on the probabilities and co-occurrence of the categories (variants and dialect groups) of the data. As our results have shown, nPMI consistently seeks the most exclusive features, yet the features are not overly localised (we do not get features with only 1 instance as the highest ranked feature). This does not mean that it fails to identify highly-localised variants, but they are just ranked lower. Furthermore, we have noticed that when a decision has to be made between Exclusivity and Representativeness, nPMI seeks exclusive features in the expense of Representativeness, like in the case for Upper Saxon. This property of nPMI is something which becomes very useful in dialect feature extraction.

We argue that for identifying distinctive features, nPMI would be the best method out of the three methods, because it extracts features with higher exclusivity, which is key to the task. Representativeness on the other hand is not as important, since a high Representativeness does not imply high exclusivity, as we have seen from the analysis of Upper Saxon with FA and FLD.

7.2 Caveats

The nPMI method which we have proposed in this paper has demonstrated its ability to seek exclusive features with data from linguistic atlases. However, it should be noted that linguistic atlas data usually contain the best curated features, meaning the words that were included in the atlases were often chosen and filtered (Francis 1983: 52), and not based on random sampling (like in an interview setting). Nonetheless, this method still builds on top of the existing dialectometric methods and allows dialectologists to understand better their data. It also opens doors to more possible venues for further research, such as the ones mentioned in Section 7.3.

7.3 Outlook for future studies

In addition to finding the features responsible for dialect clusters, the features extracted with the nPMI method can be applied to other studies in a number of ways. Firstly, nPMI can be used as an index for assessing the features proposed for previous dialect classifications. Since nPMI is a top-down method, users can calculate the nPMI scores using the classification from a previous study, then assess whether the proposed features were ranked as high as the nPMI scores. The application

[7] Pröll et al. (2014) regarded this as "a 'dialect classification' obtained through the dominance of individual factors on their corresponding region".

of nPMI in this context may shed lights on the reasoning behind the proposed classification by previous dialectologists. For instance, some features that are exclusive to a dialect group might not have been used as a criterion in a previous classification, like 1) *ŋj- > (*ɲ- >) j- and 2) possession of /œ/ in Guangfu Yue dialects (Sung and Prokic 2023).

Supplementary Material

The Python script for our feature extraction method using nPMI is available here: `https://github.com/dialmatt123/feature_extraction_nPMI`.

Acknowledgement

We would like to thank our anonymous reviewers for their valuable comments and suggestions to our paper. We would also like to thank Professor Alfred Lameli at Marburg University for sharing his expert knowledge on the German dialect features.

References

Borg, I. and P. J. F. Groenen (2005), *Modern multidimensional scaling: Theory and applications*, Springer Science and Business Media.

Bouma, G. (2009), Normalized (pointwise) mutual information in collocation extraction, *Proceedings of the Biennial GSCL (German Society for Computational Linguistics and Language Technology) Conference 2009*, Vol. 30, pp. 31–40.

Church, K. and P. Hanks (1990), Word association norms, mutual information, and lexicography, *Computational linguistics* **16** (1), pp. 22–29.

Embleton, S. (1993), Multidimensional scaling as a dialectometrical technique: Outline of a research project, *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, Springer Netherlands, pp. 267–276.

Francis, W. N. (1983), *Dialectology: an introduction.*

Goebl, H. (1982), *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Verlag der Österreichischen Akademie der Wissenschaften.

Goebl, H. (1984), *Dialektometrische Studien: Anhand Italoromanischer, Rätoromanischer und Galloromanischer Sprachmaterialien aus AIS und ALF*, Vol. 3, Niemeyer, Tübingen.

Goebl, H. (2018), Dialectometry, *in* Boberg, C., J. Nerbonne, and D. Watt, editors, *The Handbook of Dialectology.*

Grieve, J. (2014), A comparison of statistical methods for the aggregation of regional linguistic variation, *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* **28**, pp. 53.

Göschel, J. (1992), *Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas"*, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Heeringa, W. (2004), *Measuring dialect pronunciation using Levenshtein distance*, PhD thesis, University of Groningen, Groningen.

Hubert, . and P. Arabie (1985), Comparing partitions, *Journal of classification* **2** (1), pp. 193–218.

Kessler, B. (1995), Computational dialectology in irish gaelic.

Leinonen, T. (2010), An acoustic analysis of vowel pronunciation in swedish dialects.

Leinonen, T., Ç. Çöltekin, and J. Nerbonne (2016), Using gabmap, *Lingua* **178**, pp. 71–83.

Levenshtein, V. I. (1966), Binary codes capable of correcting deletions, insertions, and reversals, *Soviet physics doklady* **10** (8), pp. 707–710.

List, J.-M. and R. Forkel (2021), Lingpy. a python library for historical linguistics. https://lingpy.org.

Manning, C. and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Nerbonne, J. (2006), Identifying linguistic structure in aggregate comparison, *Literary and Linguistic Computing* **21** (4), pp. 463–476.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011), Gabmap-a web application for dialectology.

Ord, J. K. and A. Getis (1995), Local spatial autocorrelation statistics: distributional issues and an application, *Geographical analysis* **27** (4), pp. 286–306.

Pickl, S. (2013), *Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*, Franz Steiner Verlag.

Pickl, S. (2016), Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation, *The future of dialects* pp. 75–98.

Prokić, J. and J. Nerbonne (2013), Analyzing dialects biologically, *Classification and evolution in biology, linguistics and the history of science* p. 147, Franz Steiner Suttgart.

Prokić, J. and J. Nerbonne (2008), Recognising groups among dialects, *International journal of humanities and arts computing* **2** (1-2), pp. 153–172.

Prokić, J., M. Wieling, and J. Nerbonne (2009), Multiple sequence alignments in linguistics, *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELTR 2009)*, pp. 18–25.

Prokić, J., Ç. Çöltekin, and J. Nerbonne (2012), Detecting shibboleths, *Proceedings of the EACL 2012 Joint Workshop of LINGVIS UNCLH*, pp. 72–80.

Pröll, S. (2015), Raumvariation zwischen muster und zufall.

Pröll, S., S. Pickl, and A. Spettl (2014), Latente strukturen in geolinguistischen korpora, *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder. Akten des*, Vol. 4, pp. 247–258.

Schalkoff, R. (1992), Pattern recognition: Statistical, structural and neural approaches.

Séguy, J. (1971), La relation entre la distance spatiale et la distance lexicale, *Revue de linguistique romane* **35**, pp. 335–357.

Shackleton, Jr., R. G. (2005), English-american speech relationships: A quantitative approach, *Journal of English Linguistics* **33** (2), pp. 99–160.

Sung, H. W. M. and J. Prokic (2023), What are guangfu dialects?, *27th International Conference on Yue Dialects*, Ohio State University, Online Presentation. https://u.osu.edu/yue2023/.

Séguy, J. (1973), La dialectométrie dans l'atlas linguistique de la gascogne, *Revue de linguistique romane* **37**, pp. 1–24.

Ward, Jr., J. H. (1963), Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**, pp. 236–244.

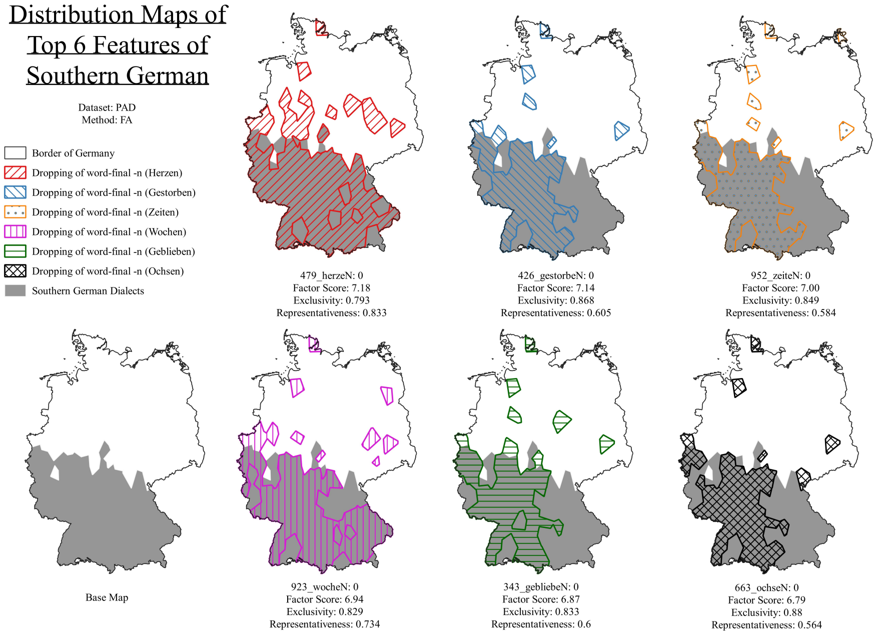
Wieling, M. and J. Nerbonne (2011), Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features, *Computer Speech Language* **25** (3), pp. 700–715.

Wieling, M., J. Prokić, and J. Nerbonne (2009), Evaluating the pairwise string alignment of pronunciations, *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELTR 2009)*, pp. 26–34.

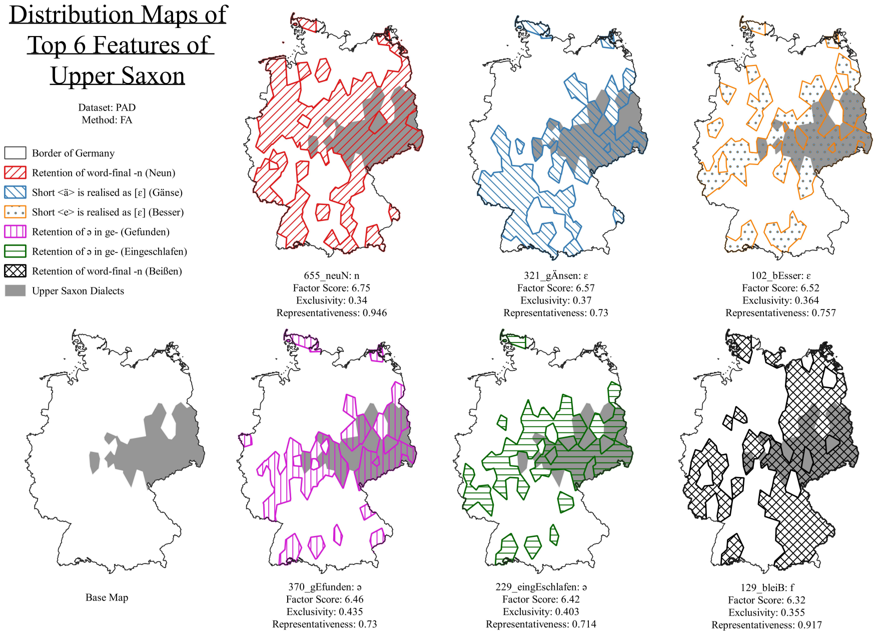
Wiesinger, P. (1983), Die einteilung der deutschen dialekte, *in* Besch, Werner, editor, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, Vol. 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, Berlin/New York: de Gruyter, Berlin, New York, pp. 807–900. http://www.degruyter.com/view/books/9783110203332/9783110203332.807/9783110203332.807.xml.

Zhao, Y. and G. Karypis (2001), Criterion functions for document clustering: Experiments and analysis, *Technical Report 01-40*, Department of Computer Science, University of Minnesota, Minneapolis, MN.

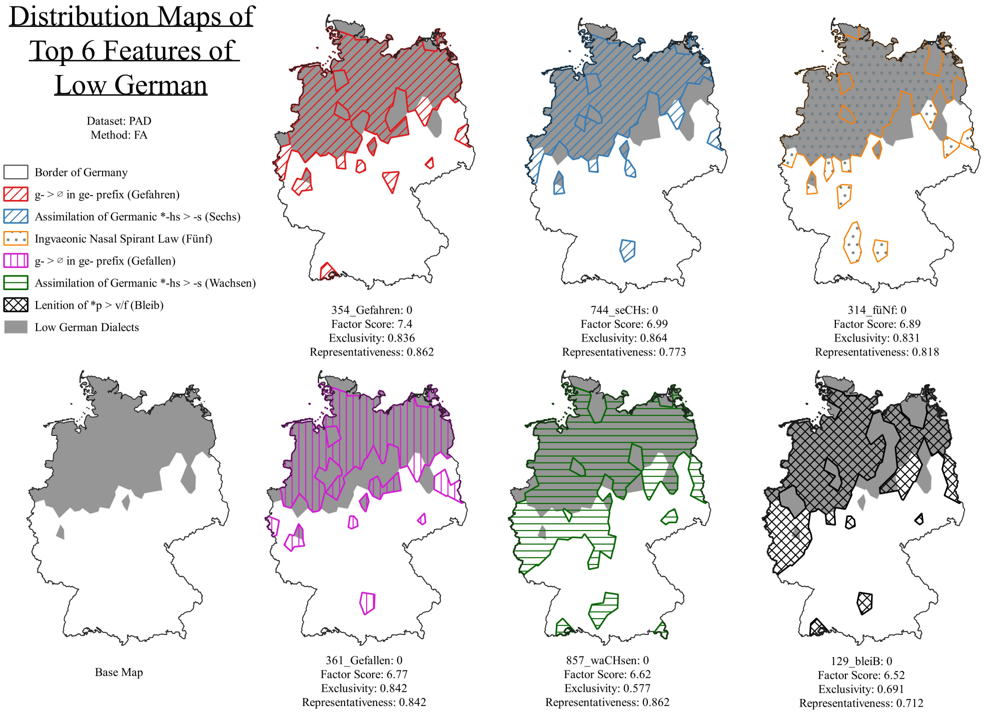
Appendix



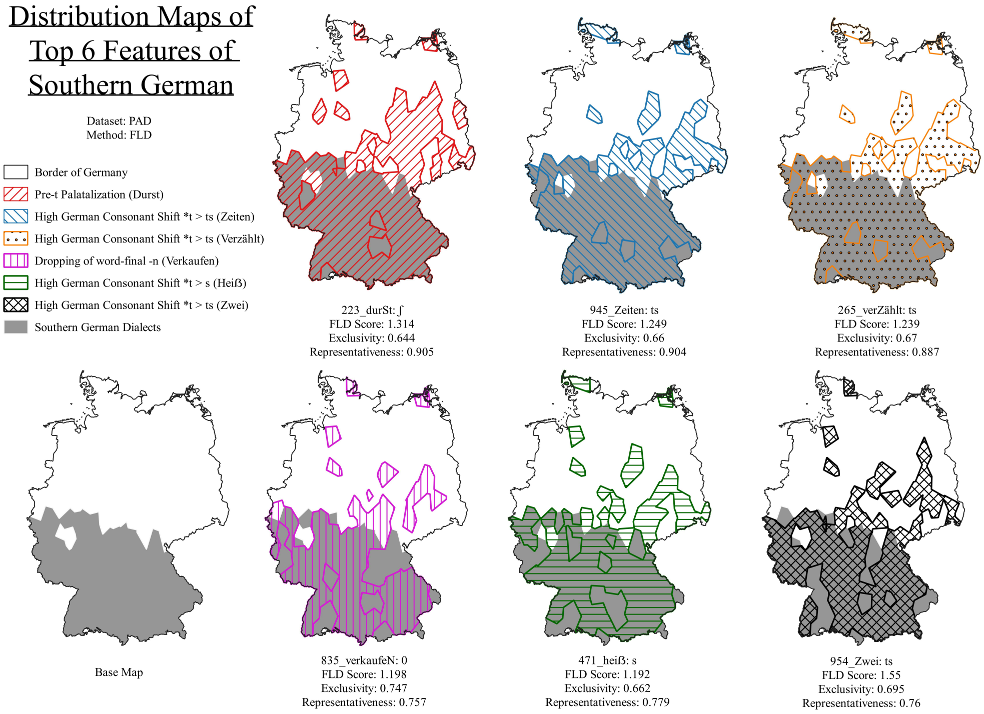
Appendix A



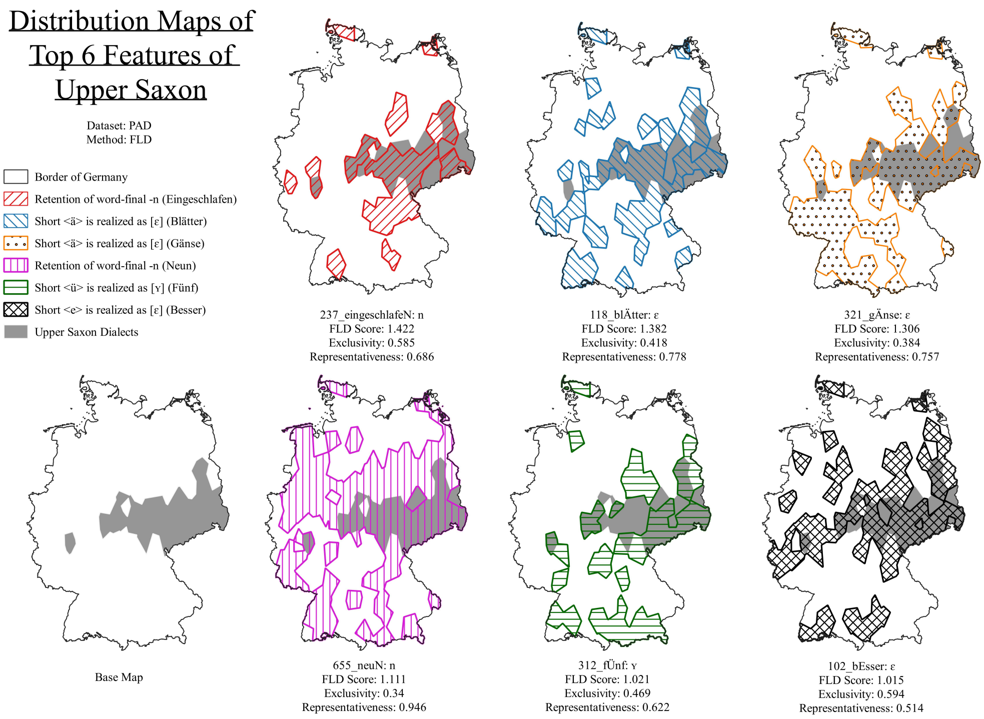
Appendix B


Distribution Maps of Top 6 Features of Low German
Dataset: PAD
Method: FA
Border of Germany
g- > ∅ in ge- prefix (Gefahren)
Assimilation of Germanic *-hs > -s (Sechs)
Ingvaeonic Nasal Spirant Law (Fünf)
g- > ∅ in ge- prefix (Gefallen)
Assimilation of Germanic *-hs > -s (Wachsen)
Lenition of *p > v/f (Bleib)
Low German Dialects
354_Gefahren: 0
Factor Score: 7.4
Exclusivity: 0.836
Representativeness: 0.862
744_seCHs: 0
Factor Score: 6.99
Exclusivity: 0.864
Representativeness: 0.773
314_füNf: 0
Factor Score: 6.89
Exclusivity: 0.831
Representativeness: 0.818
Base Map
361_Gefallen: 0
Factor Score: 6.77
Exclusivity: 0.842
Representativeness: 0.842
857_waCHsen: 0
Factor Score: 6.62
Exclusivity: 0.577
Representativeness: 0.862
129_bleiB: 0
Factor Score: 6.52
Exclusivity: 0.691
Representativeness: 0.712

Appendix C


Distribution Maps of Top 6 Features of Southern German
Dataset: PAD
Method: FLD
Border of Germany
Pre-t Palatalization (Durst)
High German Consonant Shift *t > ts (Zeiten)
High German Consonant Shift *t > ts (Verzählt)
Dropping of word-final -n (Verkaufen)
High German Consonant Shift *t > s (Heiß)
High German Consonant Shift *t > ts (Zwei)
Southern German Dialects
223_durSt: ʃ
FLD Score: 1.314
Exclusivity: 0.644
Representativeness: 0.905
945_Zeiten: ts
FLD Score: 1.249
Exclusivity: 0.66
Representativeness: 0.904
265_verZählt: ts
FLD Score: 1.239
Exclusivity: 0.67
Representativeness: 0.887
Base Map
835_verkaufeN: 0
FLD Score: 1.198
Exclusivity: 0.747
Representativeness: 0.757
471_heiß: s
FLD Score: 1.192
Exclusivity: 0.662
Representativeness: 0.779
954_Zwei: ts
FLD Score: 1.55
Exclusivity: 0.695
Representativeness: 0.76

Appendix D


Distribution Maps of
Top 6 Features of
Upper Saxon
Dataset: PAD
Method: FLD
Border of Germany
Retention of word-final -n (Eingeschlafen)
Short <ä> is realized as [ε] (Blätter)
Short <ä> is realized as [ε] (Gänse)
Retention of word-final -n (Neun)
Short <ü> is realized as [ʏ] (Fünf)
Short <e> is realized as [ε] (Besser)
Upper Saxon Dialects

237_eingeschlafeN: n
FLD Score: 1.422
Exclusivity: 0.585
Representativeness: 0.686

118_blÄtter: ε
FLD Score: 1.382
Exclusivity: 0.418
Representativeness: 0.778

321_gÄnse: ε
FLD Score: 1.306
Exclusivity: 0.384
Representativeness: 0.757

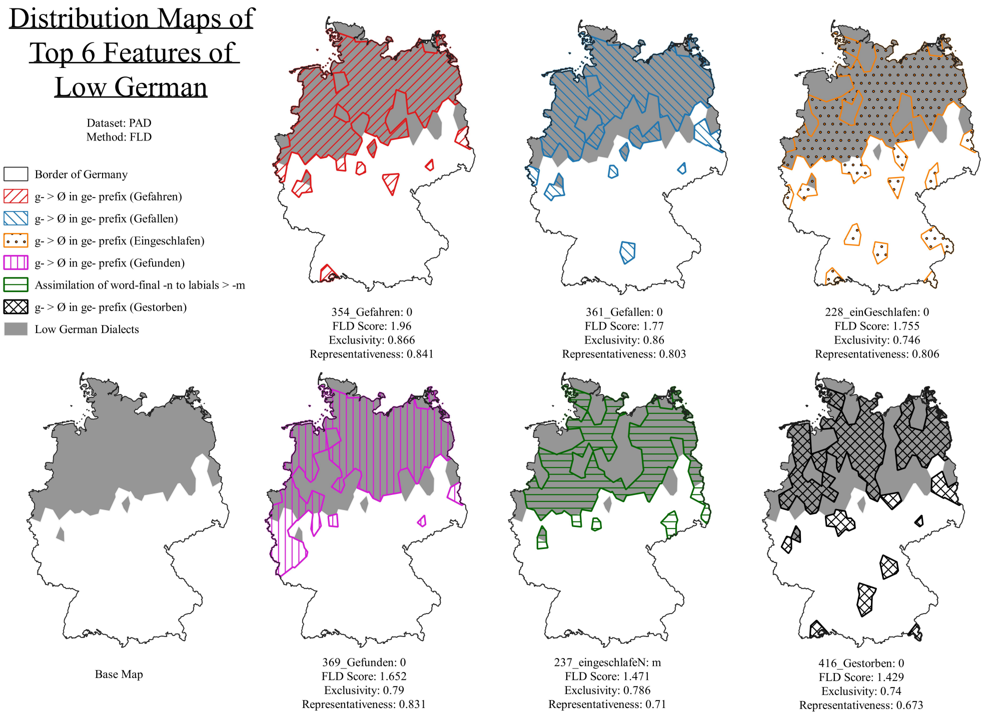
Base Map

655_neuN: n
FLD Score: 1.111
Exclusivity: 0.34
Representativeness: 0.946

312_fÜnf: ʏ
FLD Score: 1.021
Exclusivity: 0.469
Representativeness: 0.622

102_bEsser: ε
FLD Score: 1.015
Exclusivity: 0.594
Representativeness: 0.514

Appendix E


Distribution Maps of
Top 6 Features of
Low German
Dataset: PAD
Method: FLD
Border of Germany
g- > Ø in ge- prefix (Gefahren)
g- > Ø in ge- prefix (Gefallen)
g- > Ø in ge- prefix (Eingeschlafen)
g- > Ø in ge- prefix (Gefunden)
Assimilation of word-final -n to labials > -m
g- > Ø in ge- prefix (Gestorben)
Low German Dialects

354_Gefahren: 0
FLD Score: 1.96
Exclusivity: 0.866
Representativeness: 0.841

361_Gefallen: 0
FLD Score: 1.77
Exclusivity: 0.86
Representativeness: 0.803

228_einGeschlafen: 0
FLD Score: 1.755
Exclusivity: 0.746
Representativeness: 0.806

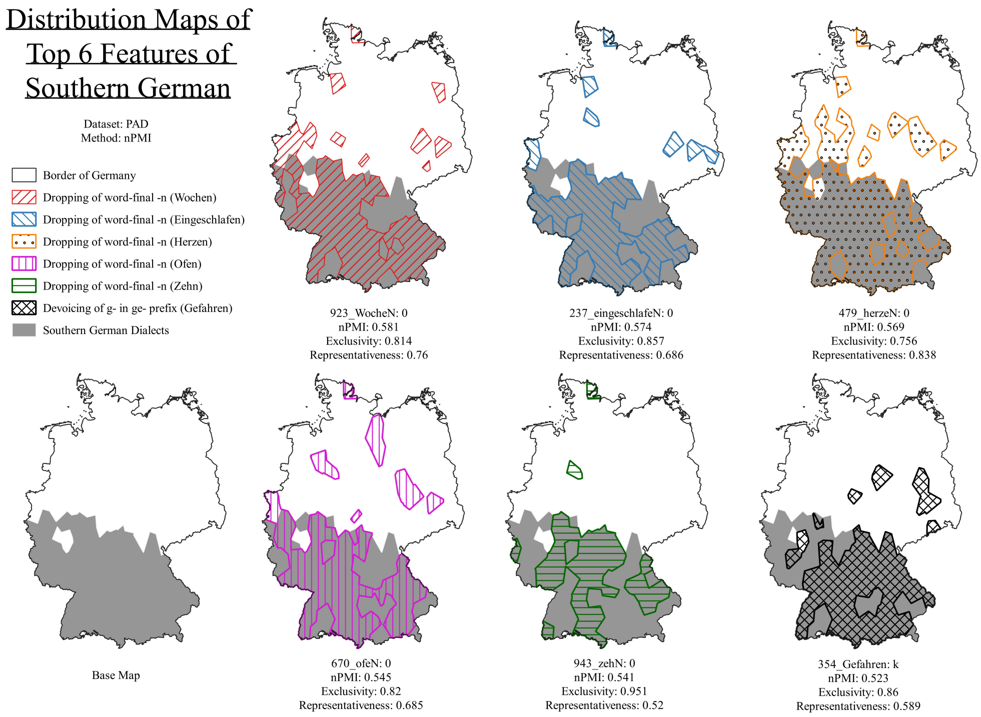
Base Map

369_Gefunden: 0
FLD Score: 1.652
Exclusivity: 0.79
Representativeness: 0.831

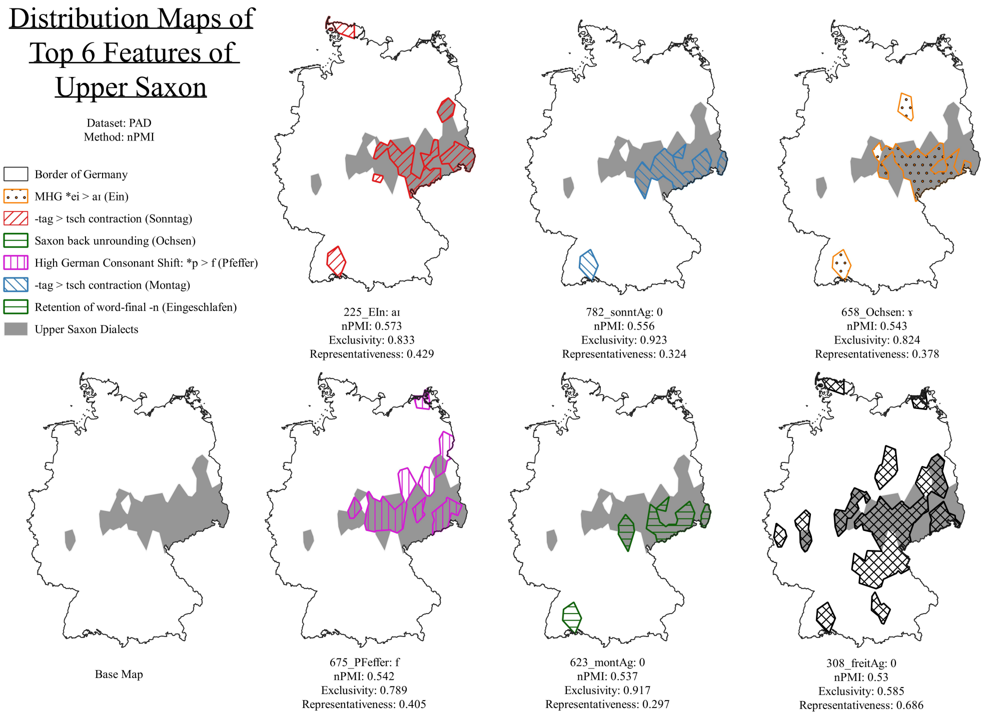
237_eingeschlafeN: m
FLD Score: 1.471
Exclusivity: 0.786
Representativeness: 0.71

416_Gestorben: 0
FLD Score: 1.429
Exclusivity: 0.74
Representativeness: 0.673

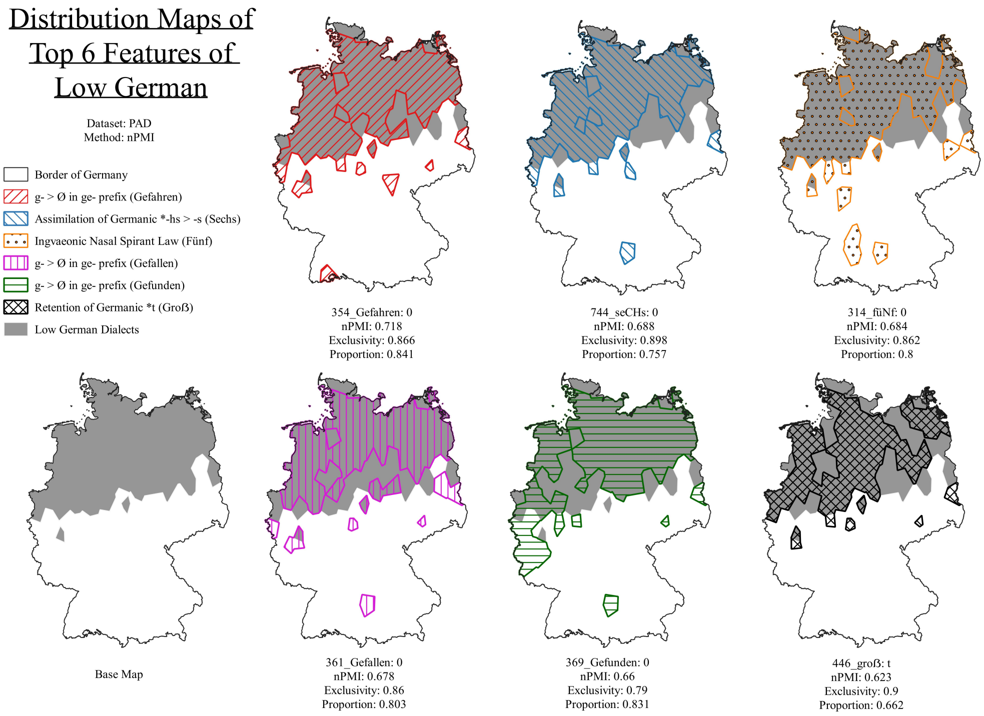
Appendix F



Appendix G


Distribution Maps of Top 6 Features of Upper Saxon
Dataset: PAD
Method: nPMI
Border of Germany
MHG *ei > aɪ (Ein)
-tag > tsch contraction (Sonntag)
Saxon back unrounding (Ochsen)
High German Consonant Shift: *p > f (Pfeffer)
-tag > tsch contraction (Montag)
Retention of word-final -n (Eingeschlafen)
Upper Saxon Dialects
225_EIn: aɪ
nPMI: 0.573
Exclusivity: 0.833
Representativeness: 0.429
782_sonntAg: 0
nPMI: 0.556
Exclusivity: 0.923
Representativeness: 0.324
658_Ochsen: ɤ
nPMI: 0.543
Exclusivity: 0.824
Representativeness: 0.378
Base Map
675_PFeffer: f
nPMI: 0.542
Exclusivity: 0.789
Representativeness: 0.405
623_montAg: 0
nPMI: 0.537
Exclusivity: 0.917
Representativeness: 0.297
308_freitAg: 0
nPMI: 0.53
Exclusivity: 0.585
Representativeness: 0.686

Appendix H


Distribution Maps of
Top 6 Features of
Low German
Dataset: PAD
Method: nPMI
□ Border of Germany
▨ g- > Ø in ge- prefix (Gefahren)
▨ Assimilation of Germanic *-hs > -s (Sechs)
⁘ Ingvaeonic Nasal Spirant Law (Fünf)
▥ g- > Ø in ge- prefix (Gefallen)
▤ g- > Ø in ge- prefix (Gefunden)
▩ Retention of Germanic *t (Groß)
■ Low German Dialects

354_Gefahren: 0
nPMI: 0.718
Exclusivity: 0.866
Proportion: 0.841

744_seCHs: 0
nPMI: 0.688
Exclusivity: 0.898
Proportion: 0.757

314_füNf: 0
nPMI: 0.684
Exclusivity: 0.862
Proportion: 0.8

Base Map

361_Gefallen: 0
nPMI: 0.678
Exclusivity: 0.86
Proportion: 0.803

369_Gefunden: 0
nPMI: 0.66
Exclusivity: 0.79
Proportion: 0.831

446_groß: t
nPMI: 0.623
Exclusivity: 0.9
Proportion: 0.662

Appendix I