Chapter 5

Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation

Simon Pickl
University of Salzburg

In this article, a threefold link is established between the concept of dialect areas as scientific constructs, prototype theory as a descriptive model and factor analysis as an operationalisation of the former two. While the idea of using prototype theory to model emic, folk concepts of dialect areas is not new, it is here for the first time used to establish a scholarly, etic model of dialect areas, which will make it easier to compare emically and etically defined dialect areas in the future. Dialect areas can be conceived of as being crisp or fuzzy, but in most cases, they are best conceptualised as being fuzzy. Following work by Gaetano Berruto, fuzzy dialect areas are defined on the basis of sets of similarly distributed variants. In a second, more practical step, an operationalisation of this model is presented that uses factor analysis to extract spatial patterns from geolinguistic data that satisfy the model's definition of dialect area. This methodology is illustrated by applying it to dialect data from Bavarian Swabia (Southern Germany). The geolinguistic structures revealed demonstrate the utility of factor analysis as a tool both for a detailed, in-depth differentiation of fuzzy dialect areas and for the detection of hitherto unknown, even very weak spatial patterns.

1 Conceptualising dialect areas as fuzzy categories

A priori, there is no such thing as a dialect area within a language space, i.e. the division of space in such areas is not a linguistic fact but an abstract concept that can differ depending on which definition is preferred and what criteria are chosen. Instead, language space can be conceived of as a dynamic arrangement of more or less mobile speakers, whose language behaviour allows linguistic forms to be attributed to certain places or regions. The distributions of these forms in the dimension of space do not normally constitute distinct dialect areas; more

Simon Pickl. 2016. Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects*, 75–98. Berlin: Language Science Press. DOI:10.17169/langsci.b81.84

often, they form a spatial continuum. Any efforts to divide space into dialect areas are therefore acts of deliberation, and they will inevitably lead to different results depending on who performs them and on the approach taken. In that sense, dialect areas are constructions.

Being ideational rather than factual entities, they have a very long tradition as conceptual realities. In the history of dialectology, the existence of dialect areas has been a permanent presupposition since its very beginnings. The cognitive organisation of dialectal variation in terms of areas or varieties seems to be virtually inevitable, or at any rate very compelling, when dealing with language in space. This can be illustrated with a passage from Chambers & Trudgill (1998), who state that they use such categories because they are handy, although they convey a strictly inaccurate picture of how language varieties are organised (in space or otherwise):

> We shall […] be using labels for linguistic varieties that may suggest that we regard them as discrete entities. It will be as well, nevertheless, to bear in mind that this will in most cases be simply an ad hoc device and that the use of labels such as 'language', 'dialect' and 'variety' does not imply that continua are not involved.
> (Chambers & Trudgill 1998: 12)

Also Peter Wiesinger (1983: 807) sees a general need or propensity to group similar ways of speaking together, which pertains to both linguists and non-linguists. Such groupings can be used, together with some salient linguistic features that are regarded as typical of them, to allocate speakers to a certain region. He stresses the practicality of regarding varieties as "discrete entities", as doing so makes it easier for speakers and for linguists alike to deal with the complexity of dialectal differences. Put more generally, dialect areas or varieties can be regarded as the expression of a mental requirement for categories, the result of a conscious or unconscious attempt to cognitively organise a large number of disparate but interrelated ways of speaking. Dialect areas are, like all kinds of categories, groupings of elements that are defined by certain traits.

There are different kinds of categories; one basic distinction often made is between crisp and fuzzy categories. Crisp categories are defined by certain traits or features that are either necessary or sufficient conditions. These categories make it very easy to decide whether a given element belongs to them or not: if the element has all the necessary or at least one of the sufficient features, then it is a member of the category. However, the definition of categories and their conditions might be regarded as arbitrary in the first place. Fuzzy categories, on

the other hand, are defined by a number of traits or features that serve as cues for that category, which are, however, neither necessary nor sufficient. If an element has many of the features associated with that category, then it is very likely that it belongs to it. Also the features themselves can have different degrees of importance for the category. Thus, no definitive answer is given as to whether an element belongs to a category or not; instead, membership is expressed as a gradual value or a probability.

Dialect areas, like other kinds of categories, can be modelled as crisp or fuzzy categories. The classical dialect area with sharp boundaries is a crisp category of local dialects; Girard & Larmouth (1993: 108–113), on the other hand, explicitly conceptualise dialect areas as "fuzzy sets", assigning local membership values between 0 and 1 to individual dialects. To obtain crisp dialect areas, specific defining features have to be selected. In this way, it is almost inevitable to pick those features that will reconstruct and thereby justify preconceived notions of areas. Therefore it is preferable not to preselect defining features, but to look at a large set of variables which may or may not be relevant. Dialect areas can then be delimited by looking for sets of bundling isoglosses. Depending on whether they coincide exactly or bundle together loosely, they delimit crisp or fuzzy dialect areas. This method is well-established in dialectology (cf. e.g. Hans Kurath 1972) and can be traced back to August Bielenstein (1892). Craig M. Carver (1987) used a similar approach for constructing dialect "layers" by combining features with similar geographic distribution. Also cluster analysis can be used to construct crisp or fuzzy dialect areas, depending on the method used (e.g. bootstrap clustering or noisy clustering; cf. Nerbonne et al. 2011: 83), without having to preselect defining features. Generally, it seems advisable to use tools that allow for fuzzy structures to emerge from the data and do not restrict the form of the outcomes to crisp structures. Later on in this paper, I will argue that factor analysis is a statistical tool particularly suited for identifying fuzzy dialect areas.

From a more theoretical perspective, viewing dialects against the background of fuzzy set theory (cf. Zadeh 1965) seems to provide a useful formalism for dealing with fuzzy dialect areas (cf. Girard & Larmouth 1993). Treating local dialects as elements that can have different degrees of belonging to an area implies that there are more and less TYPICAL examples of a dialect variety, which we could also call a DIALECT TYPE. This way of treating dialects has also been used for describing how dialects are organised cognitively by members of the speech community.

> Lectal categories, in short, constitute prototype categories. If lectal varieties constitute prototype categories, some realizations will be more 'typical' or

> 'central' or 'better examples' of a given variety than others. (Kristiansen 2008: 59)

This is one example of a number of attempts to apply prototype theory (Rosch 1973; Lakoff 1987) to dialect geography, all of which are, as far as I see, folk linguistic approaches (Christen 1998; 2010; Berthele 2006; Kristiansen 2008; Pustka 2009) in the sense that they deal with how speakers conceptualise language in space. Perceptual dialectology has produced many new insights in the cognitive perspective of language geography in the past decades, spearheaded by Dennis Preston (cf. e.g. Preston 1989; 1999; Anders, Hundt & Lasch 2010). It appears that prototype theory is a useful framework to describe how dialects are organised cognitively, and it is compatible with fuzzy set theory (see, however, Kretzschmar 2009: 218–250, who is critical of using prototype theory and favours schema theory). Prototype theory assumes that cognitive concepts are fundamentally fuzzy: examples for a concept can be more or less typical, depending on their traits or features (cf. also Labov 1973). Consequently, a specific way of speaking can be a more or less typical example of a dialect type. Folk linguistic dialect types are EMIC categories; they are cognitive concepts of the speakers whose speech is at the same time the object of linguistic investigation.

There is no reason why the fundamental linguistic concept of a dialect variety should differ from the folk linguistic concept. In other words, scholarly or ETIC ideas about geolinguistic entities can and should have the same principal structure as lay persons' implicit ideas about dialects in space while being based on transparent – and, as far as possible, objective – criteria that are not derived from the speakers' ideas, but from scientific reasoning. Only in this way, the question of to what extent emic and etic dialect types coincide and why (not) can be tackled: In what way do some folk linguistic ideas of space diverge from linguistic ones, and why? It appears worthwhile to use a linguistic (etic) notion of dialect types that is similar to the folk linguistic (emic) one, but based on intersubjective criteria. This way, it becomes possible to compare emic and etic dialect types directly and identify how and why they differ.

If dialect varieties – emic or etic ones – take the shape of fuzzy categories, then they have no clear-cut boundaries or distinct features; instead, their spatial distributions are fuzzy, and local dialects are typical of them to varying degrees. Also, linguistic features are not simply features of one variety or the other, but they have different degrees of relevance for them. Thus, individual dialects can be better or worse examples of a variety, i.e. they have different values of TYPICALITY for a dialect type. Typicality is a measure for the graded membership of dialects, thus providing the structure of a fuzzy set. If typicality or membership values of

a dialect type are projected into space, the result is a graded or fuzzy dialect area; being fuzzy, it can overlap with other dialect areas. Linguistic features, on the other hand, can be better or worse cues for a dialect type, i.e. they have varying degrees of cue validity or FEATURE VALIDITY. Consequently, the way in which dialect types are arranged in space takes the form of concretions with broad transition areas (cf. Figure 1). Note that there is no requirement for dialect types to have a core area in the sense that there must be locations that belong a hundred percent to them; instead, the core area of a dialect type can be defined as the area where it is DOMINANT, i.e. the area where it is the dialect type with the highest local typicalities. There may even be dialect types that are dominant nowhere. Because dialect types overlap in space, they appear layered; the individual layers consist of congruent distribution areas of co-occurring linguistic forms.

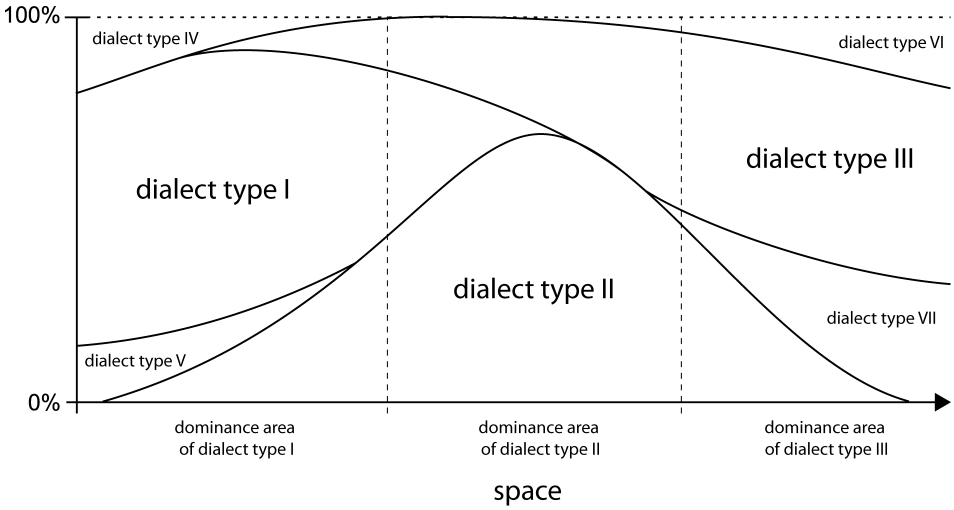


Figure 1: Layer model of dialect types in space (adapted from Pickl 2013a: 70).

This is in line with Gaetano Berruto's definition of varieties, which is based on the simple assumption that when a number of linguistic variants tend to occur together, then these variants constitute a variety:

> The tendential co-occurrence of variants gives rise to linguistic varieties. Therefore, a linguistic variety is conceivable as a set of co-occurring variants; it is identified simultaneously by both such a co-occurrence of variants, from the linguistic viewpoint, and the co-occurrence of these variants with extralinguistic, social features, from the external, societal viewpoint.
> (Berruto 2010: 229)

This notion captures several of the usual requirements for varieties: their relative internal homogeneity and their relative mutual dissimilarity, and also their association with language-external factors.[1] Varieties are thus condensations of co-occurring variants that can be pinned to a certain geographical expanse only to some extent – they are fuzzy and they overlap. Their arrangement in space is similar to the one depicted in Figure 1. According to Berruto, a variety as a condensation area is defined by certain co-occurring linguistic variants (its features). Depending on how many of these features are present in a given dialect, this dialect has a specific degree of membership between 0 (none of the features are present) and 1 (all of the features are present).

How are dialect types to be determined? Any method that is intended to identify linguistic varieties in the sense of Berruto's condensations will have to identify co-occurrences among linguistic variants. Craig M. Carver's (1987) approach did something similar in using lexical congruencies to establish 'layers' in American dialects,[2] "essentially speech areas characterised by sets of words with a similar geographic distribution" (Boberg 2005: 24). The resulting structure is quite similar to what is illustrated in Figure 1.

The remaining part of this paper is dedicated to demonstrating how the theoretical idea of fuzzy dialect types can be implemented methodologically and practically. It is to be shown that taking such an approach does not only reproduce well-established geolinguistic structures in a more nuanced way, but also that it yields new insights, e.g. regarding weaker, non-dominant structures.

2 A tool for identifying dialect types

There are various methods available for the identification of dialect areas (see Grieve 2014 for a more detailed comparison of popular statistical methods). Some of them, like fuzzy clustering, are suitable for identifying dialect types as fuzzy categories. However, I will argue that most of them are not suited for the identification of dialect types conceived of as layers of linguistic co-occurrence, either because of the structure of their outcomes or because of their internal working mechanisms, and that there are two options that are similarly well suited for this goal.

[1] It does not capture, however, their emic status, as required by Auer (1986: 99) and Lenz (2003: 389–390). As I treat emic and etic varieties separately, focussing on etic varieties, this is consistent and does not pose a problem.

[2] I would like to thank an anonymous referee for making me aware of this connection.

cluster analysis, the quantitative method that is to date the most popular tool to identify dialect areas (see e.g. Goebl 1983; Prokić & Nerbonne 2008; Prokić 2010: 17–29), analyses the aggregated similarities between local dialects to establish groups of local dialects that are relatively homogeneous internally and at the same time relatively distinct from each other. These groups or clusters are based on a measure of similarity between sites but not between distribution areas; it does not take into account the distribution patterns of individual variants and *their* mutual similarities, which would be a requirement for identifying condensations in Berruto's sense. cluster analysis does not identify types and their features, but clusters. For this reason, it is also impossible for a cluster analysis to come up with anything more subtle than global, exclusively dominant areas; subordinate, non-dominant areas that are determined by smaller numbers of features cannot be identified by cluster analysis. So, even though there are 'fuzzy' implementations of cluster analysis that yield overlapping clusters (e.g. bootstrap clustering or noisy clustering; cf. Nerbonne et al. 2011: 83), it is not a candidate for the operationalisation of dialect types. Bipartite spectral graph partitioning, which can also determine clusters of local dialects, simultaneously identifies the linguistic variants associated with these clusters (cf. Wieling & Nerbonne 2011) and is therefore in theory suitable for identifying areas together with their features. However, for our purpose this method has the disadvantage that it does not yield fuzzy areas but crisp clusters, at least as implemented by Wieling & Nerbonne (2011) or Wieling, Shackleton & Nerbonne (2013).

Multi-dimensional scaling (MDS) (see Wieling & Nerbonne 2015: 245 for an overview), "the de facto standard in dialectometry" according to an anonymous referee, arranges local dialects in a coordinate system of two or more dimensions, thus summarising the multiple similarities between local dialects. Again, the basis for the analysis are the linguistic similarities between sites, not the similarities between distribution areas. "MDS takes a site × site distance table as input and tries to assign the sites in the table to coordinates in a small-dimensional space, typically consisting of two or three dimensions" (Wieling & Nerbonne 2015: 245). Thus, it does not actually yield dialect areas but rather a dialect continuum without distinguishing condensations. Even if we took the axes as representing some sort of types, there would still be the problem, as with cluster analysis, that the results are based on global similarities between sites only, while similarities or differences between linguistic variants' spatial distributions are not taken into account. Therefore, the results of MDS cannot be interpreted as dialect types as discussed in the preceding section.

Two methods that are similar in the form of their results, but not in their internal functioning, are promising candidates for identifying fuzzy dialect areas as dialect types. Both Principal Component Analysis (PCA) and Factor Analysis (FA)[3] take feature × site matrices as data input and express recurring patterns in the data as principal components or FACTORS, usually producing a principal component/factor × site matrix as output. Additionally, a principal component/factor × linguistic feature matrix can be calculated. One of the earliest applications of PCA/FA in linguistics comes from Douglas Biber, who used it to analyse stylistic variation in written texts.

> In a factor analysis, a large number of original variables, in this case the frequencies of linguistic features, are reduced to a small set of derived variables, the 'factors'. Each factor represents some area in the original data that can be summarised or generalised. That is, each factor represents an area of high shared variance in the data, a grouping of linguistic features that co-occur with a high frequency. (Biber 1988: 79)

As a method for the reduction of high-dimensionality data, FA condenses the variation in a large data collection to a smaller number of underlying tendencies or factors. PCA does something very similar. By summarising large numbers of variants that have similar distributions, the variation in a data collection is condensed, providing a summary of predominant patterns in the data. Thus factors – "grouping[s] of linguistic features that co-occur with a high frequency" – or principal components are exactly what an operationalised method for identifying dialect types as condensations of co-occurring variants in the geographical dimension should output. The principal components or factors can be seen as condensations or layers because they are summaries of the distributions of co-occurring variants. Since co-occurrence is mathematically determined in terms of correlations, it is a technical requirement that the variant occurrences are given in the form of something like frequencies. Thus both PCA and FA meet the requirements of identifying linguistic layers as condensations of co-occurring variants and of yielding fuzzy areas as results. Hence, applying FA or PCA to geolinguistic data to find spatial patterns that qualify as varieties seems promising.

PCA and FA work quite similarly as far as their outcomes are concerned, but they function differently "under the hood". Both methods have been used several times before in dialectology.[4] For the present purpose, FA is favoured over PCA because FA is less susceptible to random variation and therefore "a more

[3] For a general introduction to both methods, see Tabachnick & Fidell (2012).

[4] See, among others, Shackleton (2005); Hyvönen, Leino & Salmenkivi (2007); Szmrecsanyi &

suitable method for identifying co-occurring linguistic features" (Leinonen 2010: 106). Leino & Hyvönen, comparing different component models including FA and PCA in an application to Finnish data, found that FA "gave solid and easily interpretable results" (2008: 186) and could be used as a default method.

The implementation of FA for dialectometric analyses presented in the following section was developed in the DFG-funded research project *New Dialectometry Using Methods of Stochastic Image Analysis*[5] (Department of German Linguistics, University of Augsburg, and Institute of Stochastics, Ulm University). It is included in *GeoLing – a software package for geolinguistic data*, which was developed in the project and is available as open source software (GPLv3) at www.geoling.net. The results reported in this article were obtained using this software.

3 Dialect types in Bavarian Swabia

In this section, the approach outlined in the previous sections will be exemplified with data from the *Sprachatlas von Bayerisch-Schwaben* (SBS, König 1996–2009), a dialect atlas that covers an area in the south of Germany. The area of investigation is delimited by the administrative region of Swabia in the south-west of Bavaria plus some adjoining stretches in the north and east, minus a part in the south that is already covered by the *Vorarlberger Sprachatlas* (VALTS). The data were collected under the direction of Werner König in the form of dialect interviews that were conducted at 272 record locations. The published version of the SBS contains approx. 2,700 maps covering lexical, morphological and phonetic variables in 14 volumes. Per location and map, up to three different variants are documented.

In previous research from the project that reported results from FA (Pickl 2013a,b; 2014; Pröll 2015), the individual subsets (lexicon, morphology, phonetics) were analysed either separately or all combined. In this article, the morphological and phonetic subsets will be analysed together, excluding the lexical subset to provide an additional angle. The rationale behind this is that morphology and

Wolk (2011); Wieling, Shackleton & Nerbonne (2013) for PCA and e.g. Clopper & Paolillo (2006); Nerbonne (2006); Grieve, Speelman & Geeraerts (2011) for FA. Grieve (2009) and Leinonen (2010) use both, while Leino & Hyvönen (2008) compare PCA and FA with other component models. The approach and data used in this paper are based on previous research by Pickl (2013a,b; 2014); Pröll, Pickl & Spettl (2015) and Pröll (2015).

[5] *Neue Dialektometrie mit Methoden der stochastischen Bildanalyse* (http://www.philhist.uni-augsburg.de/de/lehrstuehle/germanistik/sprachwissenschaft/projekte/dialektometrie/)

phonetics are usually seen to be more systematically organised and thus more relevant for geolinguistic abstractions (cf. Francis 1983: 20; Labov, Ash & Boberg 2006: 41, 119).[6] For a more detailed comparison for FA based on different subsets, see Pröll (2015: 84–132); generally, he finds that morphological and phonetic variation can be slightly better summarised (61% and 64% explained variance, respectively) than lexical variation (57% explained variance).

The data for this study consist of 831 phonetic and 541 morphological maps (1,372 in total) containing data from 272 locations. There are a total of 14,825 linguistic variants in the data,[7] i.e. each of the maps, representing an individual linguistic variable, contains on average 10.8 variants. In order to be workable for FA, these data have to be pre-processed. This is done by converting their occurrences at each location into 'weights' ranging between 0 and 1; the weight is the fraction of times a variant has been recorded at a location in relation to all records of variants at that location. Thus 0 means that a variant is not recorded at a location, 1 means that it is the only variant recorded there, 0.5 means that the variant has been recorded there together with 1 other variant, and so on, so that the values of all variants at a location add up to 1 for each variable. This seems to be the easiest and most straightforward way to deal with the non-frequency data while at the same time providing something that can be used by FA and interpreted as relative frequencies (even though as record frequencies and not necessarily as usage frequencies).

The local variant weights are filled into a location × variant matrix, which forms the basis for the analysis. Usually FA in dialectology is performed as an R-type FA, which means that spatial correlations among linguistic variants are identified. In order to perform an R-type FA, the number of cases (= sites) has to be larger than the number of items (= variants), which is clearly not the case with our data. The alternative, Q-type FA, looks for correlations among cases across items, identifying linguistic patterning of sites. The difference between Q-type FA and R-type FA is that the matrix is transposed prior to analysis, and that consequently the results are agglomerations of cases, not of items. While this is conceptually different, the outcome is very similar. "The choice of R or its transpose […] is […] not a matter of end goal but of convenience and of the ease of

[6] For a complementary analysis, looking at the lexicon alone, see Pickl (2013a,b); for an integrated analysis, looking at all linguistic subsets together, see Pröll (2015).

[7] The exact number of variants depends, of course, on the granularity of the classification of the records. For the SBS data, three levels of granularity have been defined, each being more general than the one before, thus aggregating more records together (cf. Pickl 2013a: 75–78; Pröll 2015: 47–48). For the analysis, I use Level 1 with the finest granularity, which means most of the differences between records are rendered as different variants.

meeting statistical requirements" (Cattell 1978: 326). As FA requires the number of cases to be larger than the number of items, Q-type FA has to be applied for the data used in this study to identify types of local dialects. In consequence, the FACTOR LOADINGS matrix contains the values specifying the relations between factors and locations. Varimax rotation is applied to optimise the results. Additionally, a FACTOR SCORES matrix is calculated using Bartlett's method to specify the relations between factors and variants. Both factor loadings and factor scores can take on positive and negative values.

A further parameter to be specified is the number of factors to be extracted. This choice is much less crucial as the number of clusters for cluster analysis, since from a certain number onwards, the preceding larger factors change only very little when more factors are added. This is because each additional factor explains less variance than the ones preceding it. A popular guideline is the Kaiser criterion, which admits only factors with eigenvalues greater than or equal to 1. In this case, this means that it explains the equivalent of the variance of one location.

In the present application, the Kaiser criterion leads to a total of 16 factors. These factors account for 62.21% of the variance in the data, i.e. 62.21% of the data can be explained with recurring patterns, which is in line with Pröll's (2015) findings regarding phonetics and morphology separately. The remaining 37.79% cannot be summarised by the FA. While the number of factors may seem surprisingly high, it should be borne in mind that the number of items is also very high (14,825 variants). Even the smallest factors, with well below 1% of explained variance, still contain the same amount of variation as about a hundred variants. The amount of variance accounted for has to be seen in relation to the absolute numbers; even one of the dominant factors (Factor 11) has an explained variance of less than 1%, which illustrates that even factors this small can be indispensable for getting a complete picture.

Figure 2 shows the first factor, i.e. the factor with the highest explained variance. Each location is coloured depending on its factor loading (the darker the colour, the higher the loading). The total variance explained by this factor is 15.68%. The maximal local explained variance (which is the square of the local factor loading) is 62.41% at location 163 (Olgishofen). This factor's expanse coincides roughly with an area that is traditionally identified as the Middle East Swabian dialect area (cf. e.g. Nübling 1988: 118). As a number of variants with high scores for Factor 1 are associated with this area, Middle East Swabian can be seen as a dialect type constituted by these variants. Therefore, a variant's relevance as feature, its feature validity, is specified by its factor score from FA.

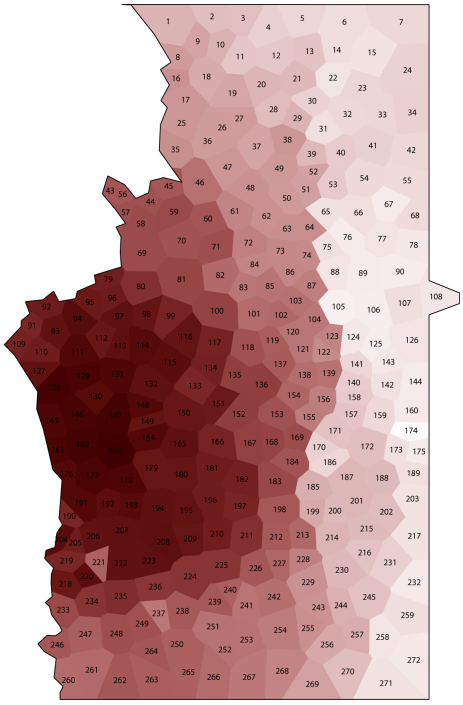


Figure 2: Factor 1 (15.68%).

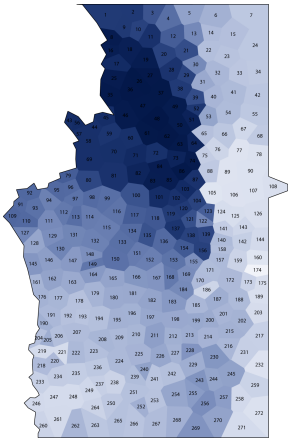
Table 1 shows the 20 variants with the highest factor score for factor 1, or the top 20 of the features of Middle East Swabian. Even though these are only the top 20 out of 2,557 variants with positive factor scores (most of them have scores close to zero), some linguistic phenomena can be ascribed to this factor: the loss of *h*, *ch* and *g* in certain positions (5, 8, 9, 14, 17, 19), the realisation of MHG *ou* as *ao* (3, 4, 7, 10, 11), and the preservation of vowel length (8, 12, 14). A deeper look at the variants with high factor scores can lead to additional insights in the linguistic make-up of this dialect type and in the alignment between variants and their distributions, but is skipped here for reasons of brevity.

Figures 3–10 show the geographic distributions of all the other dominant factors (Factors 2–7, 10–11), i.e. of all the factors that are strongest at one location at least. The explained variances are given in brackets. Divergent colours in the individual maps represent negative loadings.

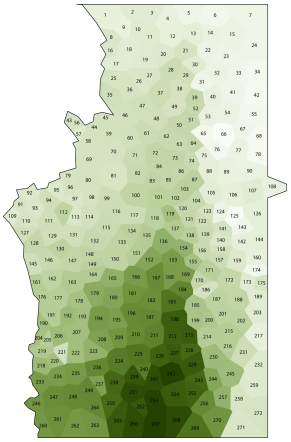
Most of these factors can be associated with traditional dialect areas: Factor 2 with North East Swabian, Factor 3 with East Algovian, Factor 4 with Central Bavarian, Factor 5 with Lechrainian, Factor 6 with Northern Bavarian, Factor 7

Table 1: Top 20 features of Factor 1. MHG: Middle High German. OHG: Old High German.

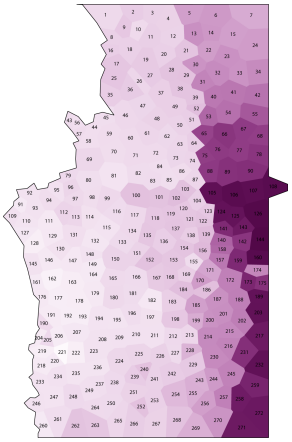
	Variable (map)	Variant	Score
1	*man* (before stress) (9.275)	*mɑ/mə*	6.226463
2	MHG *z* (germ. **t*) in *heraußen* (7.143)	fricative, lenis	6.112138
3	MHG *ou* in *auch* (5.118)	*ao*	6.045917
4	MHG *ou* in *(ein)kaufen* (5.124)	*ao*	5.823385
5	OHG Strong verbs, Class V (*siehst*, 2nd sg.) (6.58)	*sīš* (*h* not realised)	5.800744
6	MHG *ë* in *Besen* (4.57)	*ēə*	5.787916
7	MHG *ou* in *laufen* (5.125)	*ao*	5.748020
8	MHG *h* in *siehst* (7.201)	*h* not realised (long vowel)	5.721860
9	MHG *h* (germ. **h*) in *hoh-* (7.193)	*h* not realised	5.715720
10	MHG *ou* in *Auge(n)* (5.121)	*ao*	5.677732
11	MHG *ou* in *glauben/Glaube(n)* (5.119)	*ao*	5.645496
12	MHG *b* in *geglaubt* (7.14)	*b* (long vowel)	5.631118
13	*(voll)er* (*deine Hose ist … Dreck*) (9.310)	*ə*	5.576183
14	MHG *ch* in *Furche* (7.190)	*ch* not realised (long vowel)	5.575329
15	OHG Strong verbs, class Ib (*geschneit*, participle) (6.36)	*gšnīə̯*	5.570381
16	OHG Strong verbs, class VI (*trägst*, 2nd sg.) (6.75)	*drâeš*	5.562552
17	MHG *h* in *(ich) sehe*, (*er) sieht* (7.197)	*h* not realised	5.534073
18	MHG *û* (*iu*) before *l* in *Säulen* (5.K3)	*ə̯i* (first element closed)	5.521782
19	*sagst/sagt* (6.K23)	*sē(š)(d)/sâe(š)(d)* (*g* not realised)	5.508044
20	MHG *k* in *Onkel* (7.93c)	unaspirated, fortis	5.423593



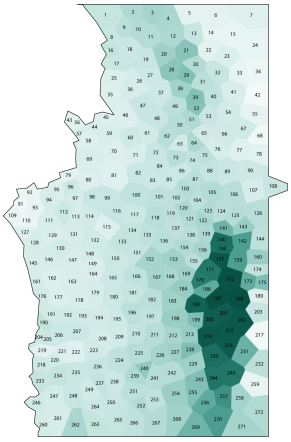
(a) Figure 3a: Factor 2 (14.04%).



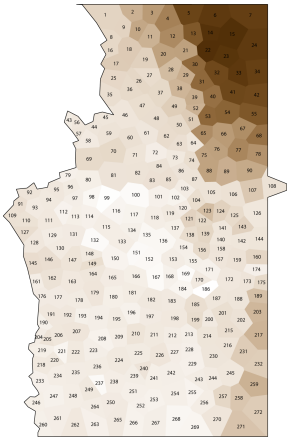
(b) Figure 3b: Factor 3 (8.98%).



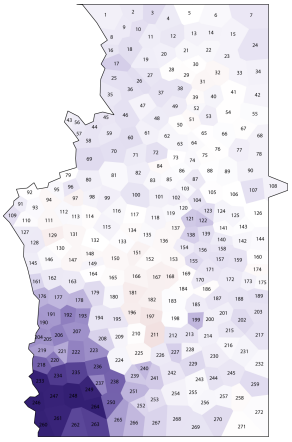
(c) Figure 3c: Factor 4 (5.74%).



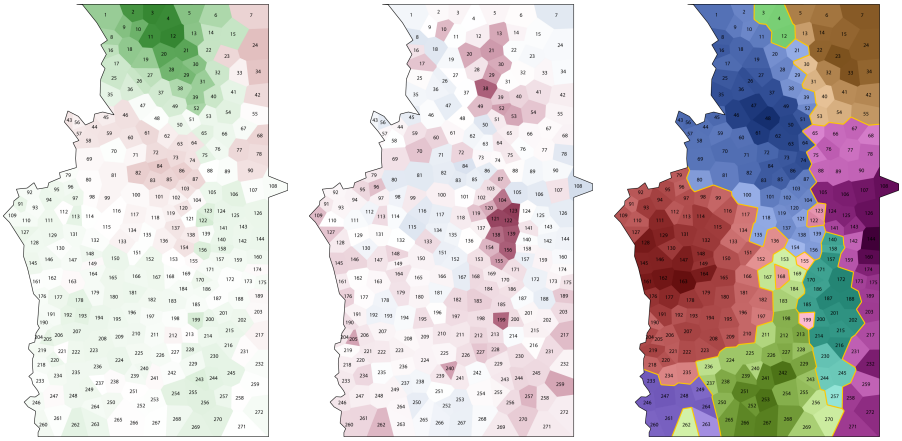
(d) Figure 3d: Factor 5 (4.68%).



(e) Figure 3e: Factor 6 (3.88%).



(f) Figure 3f: Factor 7 (1.91%).



(a) Figure 4a: Factor 10 (1.02%). (b) Figure 4b: Factor 11 (0.85%). (c) Figure 4c: Combined factor map.

with West Algovian, Factor 10 with East Franconian. Factor 11 appears to capture variants that are characteristic for towns and cities: The agglomeration around location 122 is the metropolitan area of Augsburg, and most of the other locations with high loadings are larger towns: Landsberg am Lech (199), Memmingen (205), Kaufbeuren (240), Weilheim (Oberbayern) (259), Neu-Ulm (109), Günzburg (96), Dillingen (70), Kaisheim (38), Rain am Lech (53), Donauwörth (49), Nördlingen (17), Oettingen (10), Monheim (30) with its boroghs Itzing (29) and Weilheim (21), Wemding (20), Möhren (13) (borough of Treuchtlingen). The correlation between Factor 11's loadings and the populations[8] of all 272 locations is 0.45, which corresponds to an explained variance of R^2 = 20%; the logarithmic relation is somewhat stronger (R^2 = 28%). Factor 11, therefore, can be interpreted as a geographically discontinuous urban variety; it captures variants that are used predominantly and typically in larger towns and cities. Table 2 lists the top 20 features for this type. The preservation of vowel shortness (2, 3, 4, 5, 6, 8, 12, 15, 16, 17, 18, 20) seems to be especially characteristic of this factor. It does not come as a surprise that almost all of the variants are identical with the respective standard variants. The lenition of plosives and fricatives (1, 10, 19, 20) seems to be an exception (except for 10, where lenition occurs also in the standard), which would qualify it as a unique feature of regional urbanity that is distinct from the standard.

[8] Figures for 1971 are taken from: Bayerisches Statistisches Landesamt (1972).

Table 2: Top 20 features of Factor 11.

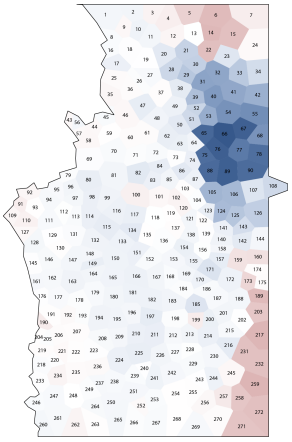
	Variable (map)	Variant	Score
1	MHG *pf* after *m*, word-final (*Dampf*, *Strumpf*) (7.215)	lenis affricate	8.066857
2	*i* before *ch* (*Stich(e)*) (sg./pl.) (3.2)	short vowel	7.848610
3	Vowel quantity in *Stich(e)* (sg./pl.) (9.30)	short vowel in singular and plural	7.548978
4	*o* before fortis fricative (*Frosch*) (3.8)	short vowel	7.436807
5	Vowel quantity in *Darm* (3.66)	short vowel	7.253476
6	MHG *i*/*u* in *Zinken* (4.47)	short open *i* (monophthong)	7.102994
7	MHG *u*/*o* in *donnern* (4.49)	neutral/closed *o*	7.009078
8	Vowel quantity in *First* (3.42)	short vowel	6.999477
9	Gender of *Teller* (9.165)	masculine	6.919323
10	MHG *t* after nasal, word-final (*tausend*) (7.K68c)	lenis plosive	6.914453
11	*im (Bett)* (9.373)	*i̯m*	6.901412
12	*a*/*o* before *ch* (*Bach*/*Dach*/*Loch*) (3.1)	short vowel	6.759506
13	MHG *â* in *Salat* (5.55)	neutral *ā*	6.635629
14	*-ig* in *König* (9.26)	*-ig*	6.581771
15	MHG *o* before *pf* in *Kopf* (4.100a)	short closed *o* (monophthong)	6.542050
16	Vowel quantity in *Stall* (3.26)	short vowel	6.491926
17	MHG *o* before *pf* in *Zopf* (4.100c)	short closed *o* (monophthong)	6.490776
18	unorganic *r* in *waten* and *Schatten* (7.254)	no *r*, short vowel	6.323896
19	MHG *t* in *Feiertag* (7.74)	lenis plosive (*r* realised)	6.303051
20	MHG *pf* in *Kopf* (7.216)	lenis affricate, short vowel	6.246637

In Figure 11, all dominant factors are combined into one map, with each location assigned to the locally dominant factor. Consequently, only information about the locally dominant factors is depicted, which means that only the surface of the dialectal landscape is visible. The resulting division into areas is in principle comparable to classifications obtained using cluster analysis or similar methods. A distinction of the present map lies in the colour shades of the individual locations, which represent the different degrees of dialect area membership. Another benefit of these results is that they retain variation 'below' the threshold of dominance, which is not visible in Figure 11 but latently present. This variation belongs firstly to the locally non-dominant parts of the globally dominant factors: each factor has loadings other than zero outside of its dominance area, but these proportions are hidden. However, they can be viewed by regarding one factor at a time (Figures 2–10).

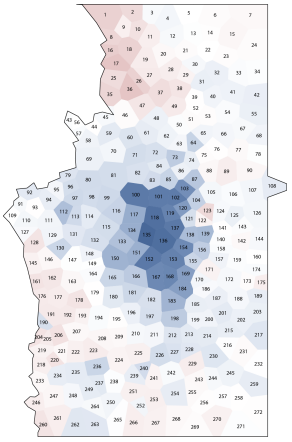
There are also factors that are dominant nowhere in the area under investigation. They do not show up in Figure 11 at all, but again, they are latently present and can be viewed individually (Figures 12–18). Summarising small fractions of the data, they contribute to a more complete picture of overall variation and the dialectal landscape, even though they represent non-dominant dialect types, dialect areas without a core area. Many of the factors can be interpreted in a meaningful way. Several of the factors shown in Figures 12–18 seem to be related to (former) market towns: their central areas (and in some cases also their counter-centres with negative values, in red) coincide with the respective market towns' catchment areas (as documented in Volume 1 of the SBS). For Factor 12, the blue centre correlates with the catchment area of (Neu-)Ulm (109), the red centre with the catchment area of Mindelheim (195); for Factor 13, the blue centre correlates with the catchment area of Lauingen (without number); for Factor 14, the blue centre correlates with the catchment area of Nördlingen (17), the red centre with that of Wertingen (72); for Factor 15, the blue centre correlates with the catchment area of Jettingen (without number), the red centre with the catchment area of Memmingen (205), for Factor 16, the blue centre correlates with the catchment areas of Schongau and Weilheim (Oberbayern) (259), the red centre with the catchment area of Mering (158).[9] These effects are relatively weak – the factors have between 0.51% and 0.85% of explained variance, which is, however, still the equivalent of 76 to 126 variants and their distributions, and they are clearly associated with their respective counterparts. Hence it is justified to speak of non-dominant dialect types that are constituted by features character-

[9] For similar findings for lexically-based factors and a more in-depth discussion, see Pickl (2013a: 170–196); Pickl (2014).

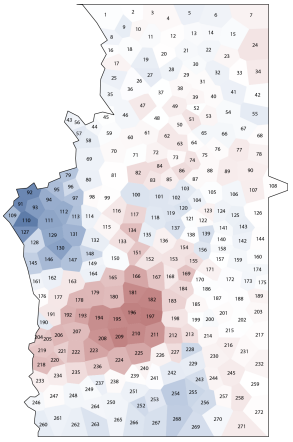
istic of these towns' surrounding areas. With these findings, a level of detail and depth is reached that goes beyond what has been attainable with previous methods of dialect classification.



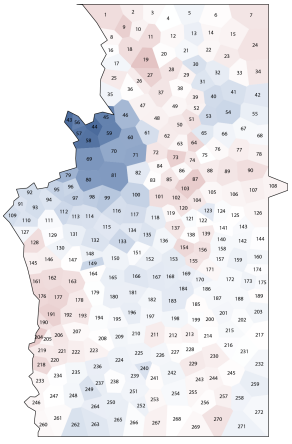
(a) Figure 5a: Factor 8 (1.11%).



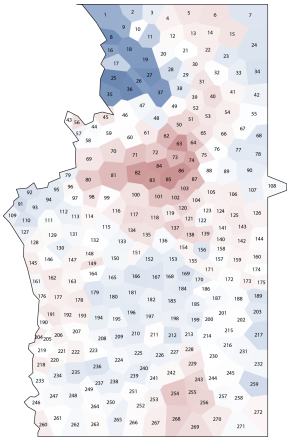
(b) Figure 5b: Factor 9 (1.03%).



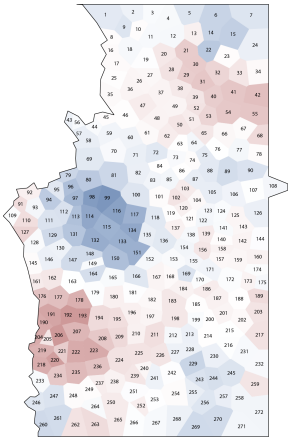
(c) Figure 5c: Factor 12 (0.85%).



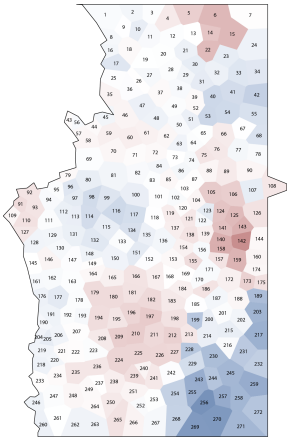
(d) Figure 5d: Factor 13 (0.71%).



(e) Figure 5e: Factor 14 (0.62%).



(f) Figure 5f: Factor 15 (0.60%).



(a) Figure 6a: Factor 16 (0.51%).

4 Conclusion

It has been demonstrated how prototype theory can be used not only to describe emic, folk ideas of dialects, but also to establish a scholarly, etic notion of dialect areas. Since the two are conceptually similar, they can be compared in a straightforward way in the future to gain insights in the relative importance of individual variants and their evaluation and assessment, e.g. based on their salience.

In this paper, it was argued that emic and etic dialect areas alike are best viewed as fuzzy dialect types, which can be described in terms of prototype theory. Dialect types have an unsharp spatial expanse, individual locations exhibiting differential membership values, and are characterised by linguistic features that have individual degrees of relevance for a type.

Following this approach, dialect areas or types are constituted by sets of co-occurring features. It was argued that factor analysis, which has been used before in dialectology, is a suitable method for the identification of such sets and thus of dialect types. Its expedience was demonstrated using data from the *Sprachatlas von Bayerisch-Schwaben* (SBS), yielding 16 factors representing dialect types. Nine of them are locally dominant within the area of investigation and lead to a classification into fuzzy dialect areas, with broad spans of overlap. Seven are non-dominant, because everywhere other factors are stronger; they, too, represent meaningful patterns and can be interpreted, as was illustrated using cities' and towns' catchment areas.

Acknowledgments

This work has been supported by the *Deutsche Forschungsgemeinschaft* (DFG), which funded the joint research project *Neue Dialektometrie mit Methoden der stochastischen Bildanalyse* of the Department of German Linguistics (University of Augsburg) and the Institute of Stochastics (Ulm University). My cordial thanks go to all of my colleagues who have contributed to this work, as well as to the editors and referees of this volume.

References

Anders, Christina Ada, Markus Hundt & Alexander Lasch (eds.). 2010. *Perceptual dialectology. Neue Wege der Dialektologie*. Berlin; New York: De Gruyter.

Auer, Peter. 1986. Konversationelle Standard-Dialekt-Kontinua (Code-Shifting). *Deutsche Sprache* 14. 97–124.

Berruto, Gaetano. 2010. Identifying dimensions of linguistic variation in a language space. In Peter Auer & Jürgen Erich Schmidt (eds.), *Language and space an international handbook of linguistic variation. Vol. 1*, vol. 1, 226–241. Berlin; New York: De Gruyter Mouton.

Berthele, Raphael. 2006. Wie sieht das Berndeutsche so ungefähr aus? über den Nutzen von Visualisierungen für die kognitive Laienlinguistik. In Hubert Klausmann (ed.), *Raumstrukturen im Alemannischen: Beiträge der 15. Arbeitstagung zur alemannischen Dialektologie*, 163–175. Graz: Neugebauer.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bielenstein, August Johann Gottfried. 1892. *Die Grenzen des lettischen Volksstammes und der lettischen Sprache: In der Gegenwart und im 13. Jahrhundert : Ein Beitrag zur ethnologischen Geographie und Geschichte Russlands*. St. Petersburg: Eggers & Co.

Boberg, Charles. 2005. The North American regional vocabulary survey: New variables and methods in the study of North American English. *American Speech* 80(1). 22–60.

Cattell, Raymond Bernard. 1978. *The scientific use of factor analysis in behavioral and life sciences*. New York, N.Y., [etc.]: Plenum Press.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge, New York: Cambridge University Press.

Christen, Helen. 1998. *Dialekt im Alltag: Eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten*. Tübingen: Niemeyer.

Christen, Helen. 2010. Was Dialektbezeichnungen und Dialektattribuierungen über alltagsweltliche Konzeptualisierungen sprachlicher Heterogenität verraten. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), *Perceptual dialectology: Neue Wege der Dialektologie*, 269–290. Berlin; New York: De Gruyter.

Clopper, Cynthia & John C. Paolillo. 2006. North American English vowels: A factoranalytic perspective. *Literary and Linguistic Computing* 21(4). 445–462.

Girard, Dennis & Donald Larmouth. 1993. Some applications of mathematical and statistical models in dialect geography. In Dennis Preston (ed.), *American dialect research celebrating the 100th anniversary of the American Dialect Society, 1889-1989*, 107–131. Amsterdam/Philadelphia: John Benjamins Pub. Co.

Goebl, Hans. 1983. Stammbaum und Welle. *Zeitschrift für Sprachwissenschaft* 2. 3–44.

Grieve, Jack. 2009. *A corpus-based regional dialect survey of grammatical variation in written Standard American English*. Flagstaff: Northern Arizona University PhD dissertation.

Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (Lingua & Litterae 28), 53–88. Berlin, New York: Walter de Gruyter.

Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2). 193–221.

Hyvönen, Saara, Antti Leino & Marko Salmenkivi. 2007. Multivariate analysis of Finnish dialect data — An overview of lexical variation. *Literary and Linguistic Computing* 22(3). 271–290.

Kretzschmar, William A. 2009. *The linguistics of speech*. New York: Cambridge University Press.

Kristiansen, Gitte. 2008. Style-shifting and shifting styles: A socio-cognitive approach to lectal variation. In René Dirven & Gitte Kristiansen (eds.), *Cognitive sociolinguistics language variation, cultural models, social systems*, 45–88. Berlin; New York: Mouton de Gruyter.

Kurath, Hans. 1972. *Studies in area linguistics*. Bloomington/London: Indiana University Press.

König, Werner (ed.). 1996–2009. *Sprachatlas von Bayerisch-Schwaben*. Heidelberg: Winter.

Labov, William. 1973. The boundaries of words and their meanings. In Charles James Nice Bailey & Roger W Shuy (eds.), *New ways of analyzing variation in English*, 340–373. Washington: Georgetown University Press.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin, New York: Mouton de Gruyter.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, Ill., [etc.]: The University of Chicago Press.

Leino, Antti & Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialectal features. *International Journal of Humanities and Arts Computing* 2(1–2). 173–187.

Leinonen, Therese Nanette. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. Groningen: Rijksuniversiteit Groningen Ph.D. dissertation.

Lenz, Alexandra N. 2003. *Struktur und Dynamik des Substandards: Eine Studie zum Westmitteldeutschen (Wittlich/Eifel)*. Stuttgart, Wiesbaden: F. Steiner.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4). 463–475.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia* Special Issue II. 65–89.

Nübling, Eduard. 1988. *Studien und Berichte zur Geschichts-, Mundart- und Namenforschung Bayerisch-Schwabens: Festgabe zum 80. Geburtstag des Verfassers*. Augsburg; Weissenhorn: Schwäbische Forschungsgemeinschaft ; In Kommission bei A.H. Konrad.

Pickl, Simon. 2013a. *Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*. Stuttgart: Steiner.

Pickl, Simon. 2013b. Verdichtungen im sprachgeografischen Kontinuum. *Zeitschrift fur Dialektologie und Linguistik* 80(1). 1–35.

Pickl, Simon. 2014. Dialekträume 'unter der Oberfläche'. Nicht-dominante wortgeographische Strukturen in Bayerisch-Schwaben. In Rudolf Bühler, Rebekka Bürkle & Nina Kim Leonhardt (eds.), *Sprachkultur, Regionalkultur: Neue Felder kulturwissenschaftlicher Dialektforschung*, 198–217. Tübingen: Tübinger Vereinigung für Volkskunde e. V.

Preston, Dennis. 1989. *Perceptual dialectology nonlinguists' views of areal linguistics*. Dordrecht: Foris Publications.

Preston, Dennis. 1999. *Handbook of perceptual dialectology*. Vol. 1. Amsterdam; Philadelphia: J. Benjamins.

Prokić, Jelena. 2010. *Families and resemblances*. Groningen: Rijksuniversiteit Groningen Ph.D. dissertation.

Prokić, Jelena & John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 2. 153–171.

Pröll, Simon. 2015. *Raumvariation zwischen Muster und Zufall. Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben*. Stuttgart: Steiner.

Pröll, Simon, Simon Pickl & Aaron Spettl. 2015. Latente Strukturen in geolinguistischen Korpora. In Michael Elmentaler, Markus Hundt & Jürgen Erich Schmidt (eds.), *Deutsche Dialekte - Konzepte, Probleme, Handlungsfelder: Akten des 4. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD)*, 247–258. Stuttgart: Steiner.

Pustka, Elissa. 2009. A prototype-theoretic model of Southern French. In Kate Beeching, Nigel R. Armstrong & Françoise Gadet (eds.), *Sociolinguistic variation in contemporary French*, 77–94. Amsterdam; Philadelphia: John Benjamins Pub. Co.

Rosch, Eleanor. 1973. Natural categories. *Cognitive Psychology* 4. 328–350.

Shackleton, Robert. 2005. English-American speech relationships. *Journal of English Linguistics* 33(2). 99–160.

Szmrecsanyi, Benedikt & Christoph Wolk. 2011. Holistic corpus-based dialectology. *Revista Brasileira de Linguística Aplicada* 11(2). 561–592.

Tabachnick, Barbara G & Linda S. Fidell. 2012. *Using multivariate statistics*. Boston, Munich: Pearson.

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

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1. 243–264.

Wieling, Martijn, Robert Shackleton & John Nerbonne. 2013. Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialects and phonetic features. *Literary and Linguistic Computing* 28(1). 31–41.

Wiesinger, Peter. 1983. Die Einteilung der deutschen Dialekte. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert Ernst Wiegand (eds.), *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Zweiter Halbband*, 807–900. Berlin/New York: de Gruyter.

Zadeh, Lotfi A. 1965. Fuzzy Sets. *Information and Control* 8. 338–353.

Bayerisches Statistisches Landesamt (ed.). 1972. *Einwohnerzahlen am 31. Dezember 1971. Jährliches Ergänzungsheft zum Amtlichen Gemeindeverzeichnis für Bayern*.