Simon Pickl

2013. *Probabilistische Geolinguistik. Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben* [Probabilistic Geolinguistics. Analyses of Lexical Variation in the Region of Bayerisch-Schwaben (Bavaria)]. Stuttgart: Franz Steiner.

This book is Simon Pickl's slightly revised PhD thesis that he defended at the University of Augsburg (Germany) in 2012.

Pickl's general concern is to describe and explain patterns of diatopic lexical variation by applying innovative geostatistical methods for analysis. His work can thus be seen as part of a "statistical paradigm" that has grown tremendously (not only) in German dialectology during the past years (cf. Streck, 2012; Lameli, 2013; Szmrecsanyi, 2013; Schwarz, 2015). As Pickl points out, his approach is non-aggregate, thus aiming to preserve the individual variation of single variables (p. 16). Another important presupposition is the probabilistic (non-deterministic) nature of lexico-geographic variation that he aims to incorporate into a model. Pickl uses the lexical maps of the *Sprachatlas von Bayerisch-Schwaben* (Linguistic Atlas of Bayerisch-Schwaben) as his corpus. His area of investigation consists of "Bayerisch-Schwaben", which represents the westernmost part of the German state of Bavaria.

The book consists of seven chapters. Chapters 1 and 2 present relevant theoretical issues, as well as the main objectives of the study. Chapters 3 and 4 pave the way for the corpus study by introducing detailed information about the data and the methods used for analysis. Chapter 5 includes the analysis of data and Chapter 6 aims to construct a model for probabilistic lexico-geographic variation. The final 20 pages consist of color images representing results of the geostatistical analyses.

Chapter 1 ("Einleitung" 'Introduction') motivates Pickl's study against the background of desiderata that have been identified in previous socio-dialectological research. Of these, Pickl sees the following two as central (cf. pp. 14–16):

1. Especially in German dialectology there is still no theoretical framework of probabilistic theories and models for describing geolinguistic variation.
2. Quantitative studies in dialectology are mainly of an aggregate style. This implies the problem that variation within individual variables cannot be preserved, and thus leads to a collapse of variation ("Variationskollaps").

To illustrate the second point: if we have, for instance, different variables belonging to the semantic concept of "vegetable" (e.g., *onion, lettuce, cucumber*, etc.) and map the dialect words by which the speakers refer to a certain variable in different dialect areas, we will get a map showing the geographi-

cal distribution areas of the different words. If we then create additional maps for further variables, and put them all in layers on top of each other, we will probably not find the same geographical distribution for each of the incorporated variables (although they belong to the same semantic concept "vegetable"), but the lexical distributions will most likely differ from one another. These differences (which Pickl calls 'variation within variation' ["Variation in der Variation"], p. 16) are lost in an aggregate map as only one aggregate lexicogeographical distribution is visible, losing all information about the variation within the individual incorporated variables.

Pickl's study focuses on lexical distribution patterns, which, in contrast to patterns of phonological variation, still lack well-founded explanations. The author sees it as a challenge to have chosen lexical variation as the most chaotic part of geographical distribution patterns (phonological patterns of geolinguistic distribution are much more consistent). If his probabilistic model works out for explaining lexical variation, it should also be appropriate for the more regularly behaving parts of the language system (e. g., phonology, morphology).

Pickl's main research objectives are the following (cf. pp. 21–23):

1. Develop statistical methods to test hypotheses for the correlation of geolinguistic patterns with intra- and extralinguistic variables, and to discover new patterns.
2. Apply these methods to the data corpus of the *Sprachatlas von Bayerisch-Schwaben*. The approach is data-based on the one hand (top-down approach: are there semantic, grammatical, extralinguistic factors that have an influence on geolinguistic patterns?) and on the other hand data-driven (bottom-up approach: which data structures play a role in the constitution of the linguistic area investigated?).
3. Develop a geolinguistic-variationist model that is able to describe the dynamics of lexicogeographic structures against the background of existing theoretical models of language variation.

In Chapter 2 ("Theorie" 'Theory') some basic concepts and terms are clarified that are often used inconsistently in geolinguistic research. The author focuses on terms connected to language dynamics in space (e.g., "distribution" vs. "diffusion," "variation" vs. "change," etc.), concepts of space (e.g., Euclidian, social, and cognitive space), as well as on diatopic varieties (e. g. "lect" vs. "variety"). Pickl differentiates between intralinguistic and extralinguistic approaches to define a variety. His study is based on an intralinguistic approach of what constitutes a variety, and is modeled on the concept of Gaetano Berruto (2010).

According to this, a variety is shaped by a set of variables that each, again, consist of a set of variants that can cooccur. For instance, of a set of words A, B, C, D, E [variants], the words A and B cooccur if especially these two are frequently used within the same area in order to refer to a certain concept [variable]. Through their cooccurrence they contribute to the formation of a variety in this area. Through the use of this bottom-up approach, which creates so-called 'dialect types' ("Dialekttypen") with a high amount of cooccurrences, Pickl claims to open up a new perspective on the structure of language in space. He criticizes top-down approaches (cf., e.g., old-school dialectology) that often fit linguistic data to predetermined and pre-expected geographic structures.

Chapter 3 ("Daten" 'Data') contains information about the data used for the study. Pickl carries out a secondary analysis of existing data from the *Sprachatlas of Bayerisch-Schwaben* and presents information about the data structure and about cartographic and lexical classification methods applied by the editors of the atlas.

Chapter 4 ("Methodik" 'Method') contains methodological considerations of how lexical variation in space can best be presented. This chapter introduces and discusses the fundamental measures chosen for mapping lexical variation in space. Pickl here again criticizes the aggregative approach of quantitative geolinguistics, and argues for leaving the identity of individual linguistic variants untouched as far as possible by using non-aggregating methods. He then points out how lexical data can be mapped in a way that is in line with his approach. First, he splits up all occurring variants and exemplifies this for seven variants of the variable *Kartoffelkraut* ('haulm'). A major consideration is how to estimate the spatial density of attested variants. This is done by first interpolating the probability of occurrence for each location, even for locations without data. In a second step, "dominant areas" are mapped out, yielding a cumulative map for all variants of a variable. This indicates where a certain variant is most frequent without deleting information about the other included variants.

A very important part of Pickl's methodology is the establishing of three so-called 'distribution values' ("Distributionskennwerte"). These values are numerical operationalizations of the following geolinguistic features represented in maps:

1. Complexity: this refers to the level of dissection into subareas, measured in the overall length of all isoglosses.
2. Compactness: weight/density of variants within their assigned area. If a map has a high level of compactness, the variants that constitute the various

subareas are very frequent within their respective subareas. Hence, these mapped subareas reflect the diatopic reality to a very high extent.

3. Homogeneity: frequency of one variant among other variants within its related area (dominant area).

These three distribution values are crucial in the course of Pickl's statistical analyses on the basis of the *Sprachatlas von Bayerisch-Schwaben* in Chapter 5 ("Analyse" 'Analysis'). The chapter is divided into two sections, concentrating on a data-based (5.1) and a data-driven (5.2) analysis respectively.

In Section 5.1 Pickl tests various hypotheses about the correlation between features of geolinguistic structures (complexity, compactness, homogeneity) and various endogenous (frequency, word class, semantic topic) and exogenous factors (rivers, political borders).

Frequency: In dialectological literature the size of the distribution area of a variant has often been viewed as a decisive factor in determining the lexical complexity of an area (Bach, 1950): very common concepts are referred to by only a few words (each with high frequency) that are used in a broad geographical area, whereas the more special and emotional concepts are referred to by many different words (with low frequency), each used only in a small area. Pickl tests the hypothesis whether the size of a certain word's area correlates with the respective word's frequency. However, the results show no significant correlation between absolute variant frequency and any of the three geolinguistic features (complexity, compactness, homogeneity).

Word class: Regarding the influence of the word class on spatial lexical distribution, Pickl only reports significant results for interjections. This word class implies low complexity on the one hand and high compactness/homogeneity on the other hand (the interjections discussed here are oral salutations and acclamations). In his interpretation, Pickl argues that salutations are very frequent in conversation, and can spread out through language contact very quickly. Once established, they remain very stable. The argumentation here seems quite contradictory: if such salutations and acclamations spread very easily, how can they remain stable then? As to acclamations, Pickl explains their low complexity as due to the occupational group of wagoners, who were a mobile part of society and distributed this kind of expressions very rapidly throughout the country.

Semantic topic: The volumes of the *Atlas of Bayerisch-Schwaben* present their material by semantic topics. Thus, Pickl can check whether different semantic concepts are aligned with particular geolinguistic structures. Such concepts are, e.g., "the human body," "weather phenomena," or "animals." In eleven cases Pickl finds significant relations of these semantic topics with one

or more of the three geolinguistic features he considers (complexity, compactness, and homogeneity). The interpretations of the results are in most cases plausible but not always convincing, and sometimes they are mere speculation. For example, the expressions for "to rain heavily"/"to rain lightly" reach high complexity values, while the expressions for different wind directions do not. This result is interpreted as based on the higher affective/emotional potential of rain expressions. Personally, I cannot see why expressions for different kinds of rain should be more affective/emotional than those for wind. It has to be stated, though, that Pickl is well aware of the problem of such ad-hoc interpretations. In an attempt to validate his assumptions, he seeks to establish some basic tendencies that link social, phatic or emotional concepts to the complexity of geolinguistic structures. One example is whether a semantic concept shows a tendency towards formal innovation ("Innovationsaffinität" 'inclination to innovate'). It turns out that innovations are often linked to emotional concepts and features that show a low level of geolinguistic compactness and homogeneity on the one hand, and a high level of complexity on the other. Another example is the tendency of a variant to diffuse in geographic space ("Diffusionsaffinität" 'inclination to diffuse'). A high tendency to diffuse can be found for semantic concepts with a high degree of social relevance and for concepts related to the mobility of the speaker. According to Pickl's analyses, this results in low geolinguistic complexity. He summarizes that his geolinguistic features (complexity, compactness, and homogeneity) can best be explained by basic mechanisms like innovation and diffusion.

Exogenous factors: In Subsection 5.1.3, Pickl focuses on rivers and political borders as exogenous factors affecting linguistic divergence and convergence. His aim is to establish a statistical method that predicts whether linguistic isogloss bundles just run at random throughout the country, or whether they have a significant tendency to be situated along nonlinguistic borders. The first exogenous factor tested is a set of eleven rivers running through the area of investigation of the Linguistic Atlas of Bayerisch-Schwaben. The results are significant only for three rivers. The river *Lech* reaches a very high significance, followed by the rivers *Danube* and *Zusam*. Thus, if the number of significant results is judged against the total number of analyzed rivers, it appears that rivers generally tend not to be a very strong predictor of linguistic divergence. The second exogenous factor Pickl analyzes is political borders. His results show that such borders can only become statistically relevant if they are old, or if they have geographically more or less identical predecessors that reach far into the past. For political borders with an age up to about 200 years, no significant values appear, while borders with an age about 400–500 years become significant predictors of linguistic divergence.

Section 5.2 is dedicated to a data-driven approach to geolinguistic patterns. As a methodological tool, Pickl uses factor analysis, a data-mining method that allows identifying similar structures within a dataset without having to define these similar structures previously (as factors). This bottom-up approach follows Berruto (2010), who defines varieties as a set of cooccurring linguistic features. The geolinguistic outcomes of factor analysis can then be compared with corresponding and known diatopic patterns. Pickl describes the adjustment of a number of values before carrying out the factor analysis; for example, he sets the number of factors to 20, a number that explains 59% of overall variation. These factors sum up those parts of a variant's geographical variation that have a similar distribution (Pickl's dataset consists of a matrix of 12,341 variants distributed over 272 locations).

In Subsection 5.2.6, Pickl discusses the geolinguistic results for each of the 20 factors and the corresponding lexical variants that constitute these factors. Most factors produce prominent geolinguistic areas that are clearly coherent. They correspond, to a certain extent, to dialect areas that have been described earlier by traditional dialectology. Besides the possibility of illustrating coherent areas that are constituted by single factors, Pickl also presents profiles of individual locations (p. 265). These profiles are constituted by a "sandwich" of factors at this very location, i.e., the profiles show each factor and to what extent it explains variation at the respective location. Each of the 20 factors has a different degree of explanatory power as to the variance at a given location (for all factors together, it sums up to about 60%, while 40% of variance remains idiosyncratic). This representation makes it easy to find out which factors play a role in the varietal setup of a concrete location (and to what degree they do this).

In Chapter 6 ("Modell" 'Model') Pickl creates a model that aims to connect a theory of probabilistic geolinguistics to geostatistical methods of empirical data analysis. In a first step, Pickl examines the variation of variants in space and tries to set up a statistical model that allows forecasting instances of language change on the basis of the pattern of variation among variants. In Pickl's view, three major processes are responsible for such changes of variants:

1. Innovation: This can be described as random appearance of new variants on the basis of a given variation in space. Based on a stochastic model introduced by Altmann (1985), Pickl develops an extended version that allows him to calculate the development of new variants in time. A methodological problem of Pickl's calculations, however, is the missing diachronic data. As the data of the *Sprachatlas von Bayerisch-Schwaben* is synchronic, we

do not have direct information about the extent of innovation in a certain area. Pickl uses the aforementioned geolinguistic feature "homogeneity" as an indicator for probability (affinity) of innovation.

2. Entrenchment: This term refers to the development of actual usage of a certain variant once it came into play by an innovation process. Some variants might disappear very quickly, while others get established within a given (homogeneous) society. The stochastic formulation of this measurement is based on Altmann et al. (1983), but it allows more than just two competing variants in the calculation.

3. Diffusion: This concept brings in the concept of space. Here, size (number of speakers) and spatial constellation (distance, geographical distribution of speakers) of locations are taken into account. The following parameters are considered in detail: number and geographical distribution of speakers, the probability of interaction between speakers from different locations, as well as the distance between locations. Pickl incorporates these parameters into a stochastic model that calculates the likelihood of diffusion of variants as a function of geographic space.

Having compiled these three processes into formulas, Pickl runs a hypothetical simulation with randomly set parameters and without real data on the basis of 20 locations and 62 chronological iterations. The outcome is presented in a colored schema on pages 268 and 269, and then the effects and visual clarity of the three randomly set parameters (innovation, entrenchment, diffusion) are discussed. Pickl states that this simulation is just a test run of the model. Whether this model is valid can only be confirmed by feeding it with real data and comparing its outcomes to geographically and diachronically distributed empirical data. Until this comparison has been carried out, Pickl's model (which he admits himself) has to remain hypothetical for now.

In sum, Pickl's work represents an innovative (geo)statistical approach aiming at a deeper understanding of why lexical data is distributed in space the way it is and how varieties form, seen from a bottom-up perspective, and how this distribution of varieties and variants might develop in time. The major achievements of this book are threefold: 1) It presents an elaborate stochastic method of preparing linguistic data in space, 2) it carries out concrete statistical analyses that reveal insights into the formation of geolinguistic structures from a data-based and data-driven perspective, and 3) it contributes a stochastic approach that provides a deductive "forecast" tool for the geolinguistic development of variants and varieties. However, it has to be stated that this "forecast" tool is quite hypothetical and mechanistic. In my view it seems applicable only to "idealized" linguistic data as we find it among most dialect atlases in the

German-speaking area.[1] Exposed to more spontaneous and socially complex data, the model with its set of incorporated processes (innovation, entrenchment, diffusion) could probably not adequately display (geo)linguistic developments. Pickl is well aware of this problem (see footnote 138, p. 217), and as the aim of his work is focused on the object of traditional dialects and not on the whole spectrum of linguistic variation, one cannot blame him for having ignored this methodological issue.

Pickl's work contains a lot of statistics, mathematical derivations and formulas. This does not make the methodology easily accessible, especially not for those linguists who have not worked with statistical methods. On the other hand, some parts of Pickl's book can be recommended as reading in undergraduate and graduate courses. Especially Chapter 2 offers a very thorough, yet short and easy-to-read overview of different approaches and terms concerning the field of language and space.

Altogether this book is a remarkable study of lexico-diatopic variation. Hopefully, the developed stochastic methods will soon be applied to other linguistic levels of diatopic variation, e.g., phonology, morphology or syntax.

Christian Schwarz
University of Münster
ch.schwarz@uni-muenster.de

References

Altmann, Gabriel. 1985. Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft* 4/2: 139–155.

Altmann, Gabriel, Haro von Buttlar, Walther Rott, and Udo Strauss. 1983. A law of change in language. In Barron Brainerd (ed.), *Historical Linguistics*, 104–115. Bochum: Brockmeyer.

Bach, Adolf. 1950. *Deutsche Mundartforschung. Ihre Wege, Ergebnisse und Aufgaben.* Heidelberg: Winter.

Berruto, Gaetano. 2010. Identifying dimensions of linguistic variation in a language space. In Peter Auer and Jürgen E. Schmidt (eds.), *Language and Space: An International Handbook of Linguistic Variation*, Vol. 1, 226–241. Berlin: de Gruyter.

---

1 Here "idealized" means that the data of these atlases is knowledge-based (not spontaneous) and derives from carefully selected NORMs and NORFs (non-mobile older rural males/females), representing a high level of sociolinguistic homogeneity.

Lameli, Alfred. 2013. *Strukturen im Sprachraum. Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Berlin: de Gruyter.

Schwarz, Christian. 2015. *Phonologischer Dialektwandel in den alemannischen Basisdialekten Südwestdeutschlands im 20. Jahrhundert. Eine empirische Untersuchung zum Vokalismus*. Stuttgart: Steiner.

Streck, Tobias. 2012. *Phonologischer Wandel im Konsonantismus der alemannischen Dialekte Baden-Württembergs. Sprachatlasvergleich, Spontansprache und dialektometrische Studien*. Stuttgart: Steiner.

Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-based Dialectometry*. Cambridge: Cambridge University Press.