Vorstellung des Software-Pakets *GeoLing*

nebst kurzem Forschungsbericht zum Projekt
„Neue Dialektometrie mit Methoden der stochastischen Bildanalyse“

Wir stellen im Folgenden zentrale Verfahren und Ergebnisse des Projekts „Neue Dialektometrie mit Methoden der stochastischen Bildanalyse“ sowie das Software-Paket ***GeoLing*** vor, das aus dem Projekt erwachsen und über http://www.geoling.net frei verfügbar ist.

Das Projekt, das die DFG über 4 Jahre hin förderte**,** wurde getragen

- vom Lehrstuhl für deutsche Sprachwissenschaft an der Universität Augsburg (ab Herbst 2012 zusätzlich vom Fachbereich Germanistik der Universität Salzburg) von den Professoren Stephan Elspaß (Augsburg, dann Salzburg) und Werner König (Augsburg) sowie
- vom Institut für Stochastik der Universität Ulm von den Professoren Volker Schmidt und Evgeny Spodarev.

Basis für das Projekt waren die in digitaler Form vorliegenden Daten des *Sprachatlas von Bayerisch-Schwaben*. Grundlegend für alle Analysen war die Annahme, dass die vorliegenden Karten mit ihrem Variantenreichtum aus Stichproben bestehen, die trotz der Standardisierungsbemühungen bei der Aufnahme nicht in jedem Fall hätten so ausfallen müssen, wie sie ausgefallen sind.

Grundlegende Verfahren:

– *mit Blick auf Einzelvariablen*: Auf der Basis dieser Annahme wurden in einem ersten Arbeitsschritt aus den Punktsymbolkarten des Sprachatlasses Flächenkarten, d. h. farbige Polygonkarten erstellt. Sie stellen die Intensitäten, die Wahrscheinlichkeiten des Vorkommens der Varianten dar und werden mittels Intensitätsschätzung errechnet. Dieses Verfahren ermöglicht es, der Existenz mehrerer Varianten, die an einem Ort vorkommen, und ihrer geographischen Verteilung in der kartographischen Darstellung (in gradierten, d. h. abgestuften Flächenkarten) gerecht zu werden. Diese Karten bilden die Grundlage für viele weitere Analysen, die in dem Projekt vorgenommen wurden (s. u.).

– *mit Blick auf variablenübergreifende Strukturen*: Für die Auswertung globaler, latenter Strukturen und wiederkehrender Muster wurde die Faktorenanalyse eingesetzt, mit der die geographischen Verteilungen zahlreicher Varianten miteinander in Beziehung gesetzt werden und Zonen erhöhter Variantenkookkurrenzen extrahiert werden. Auf diese Weise entsteht eine neue, differenzierte Dialektgliederung, die nicht nur graduelle Zugehörigkeiten von Orten zu – unscharfen – Dialektgebieten zulässt, sondern auch solche Regionen ermittelt, die nur aufgrund weniger Varianten eine gewisse dialektale Eigenständigkeit aufweisen und bei üblichen Dialekteinteilungen durchs Raster fallen.

Wichtigste Ergebnisse:

– Es wurde eine neue differenzierte Dialektgliederung erarbeitet, die die Mundartregionen nicht mehr durch Grenzlinien an den Rändern zu bestimmen sucht, sondern durch die in ihnen vorhandenen Gemeinsamkeiten, die – ohne Vorgaben von außen – durch *data mining* gefunden wurden. Mit der entwickelten Methode lassen sich für jeden Ort Werte zur relativen Zugehörigkeit zu einer Dialektregion angeben.

– Es wurden Kennwerte erarbeitet, die die Struktur von Sprachkarten beschreiben. Diese quantifizierbaren Größen stellen Eigenschaften wie *Homogenität*, *Kompaktheit* und *Komplexität* einer Karte dar und dienen als Grundlagen für weiterführende statistische Analysen.

– Unter Anwendung von Bilderkennungsverfahren können erstmalig gezielt große Kartenkorpora nach bestimmten Strukturen (etwa Kreise oder Ellipsen) einzelner Varianten durchsucht werden.
– Erstmals können geographische Linien wie Flüsse oder politische Grenzen statistisch auf ihre sprachgeographische Relevanz als Barrieren (oder auch Verstärker) hin statistisch getestet werden.

Ein großer Teil der erarbeiteten Programme wurde als geostatistisches **Software-Paket *GeoLing*** ins Netz gestellt und ist damit für die Forschung *frei und kostenlos benutzbar*, siehe die Seite http://www.geoling.net. Dieses Paket ist so gehalten, dass es auch für LinguistInnen ohne Programmiererfahrung benutzbar ist. Der zur Verfügung gestellte Quellcode ermöglicht es Anwendern mit entsprechendem Hintergrund, die Programme für ihre Zwecke zu adaptieren. Die gewählte Programmiersprache *Java* macht sie auf einer Vielzahl von Plattformen lauffähig. Zu diesem Programmpaket, bei dem man die *Arbeitssprachen Deutsch und Englisch* wählen kann, werden auch *Anleitungen in beiden Sprachen* zur Verfügung gestellt.

Nähere Beschreibungen der genannten und weiterer zentraler Methoden, die im Projekt entwickelt und angewandt wurden, sowie der damit erzielten Ergebnisse finden sich in den Dissertationen von Simon PICKL (2013) und Simon PRÖLL (2015) sowie in verschiedenen Aufsätzen:

- Man kann den geographischen Übergang von zwei oder mehr Varianten quantitativ beschreiben, auch visualisieren mit Farbintensitäten und dadurch die Wahrscheinlichkeiten des Auftretens einer bestimmten Variante kartographisch darstellen (RUMPF et al. 2009; PICKL / RUMPF 2011; PICKL 2013a; PICKL et al. 2014; PRÖLL 2015).
- Es wurden Kennwerte entwickelt, die die Struktur von Sprachkarten beschreiben. Diese quantifizierbaren Größen stellen Eigenschaften wie *Homogenität*, *Kompaktheit* und *Komplexität* einer Karte dar und dienen als Grundlage für weiterführende statistische Analysen (RUMPF et al. 2009, 2010; PICKL / RUMPF 2011; PICKL 2013a, 2013b).
- Die statistische Auswertung der oben genannten Kennwerte und zusätzlicher Eigenschaften erlaubt es, den Einfluss sprachlicher Eigenschaften auf die räumliche Gliederung der entsprechenden Karten zu untersuchen (vgl. PICKL 2013a, 2013b; PRÖLL 2015).
- Mittels Clustern können Gruppierungen aus denjenigen Karten gebildet werden, die sich in ihrem Raumbild entsprechen (vgl. RUMPF et al. 2010; MESCHENMOSER / PRÖLL 2012a; PRÖLL 2015) – dabei sind durch die Methode des „Fuzzy Clustering“ auch graduelle Aussagen möglich.
- Unter Anwendung von Bilderkennungsverfahren konnten erstmalig gezielt große Kartenkorpora nach bestimmten Strukturen einzelner Varianten durchsucht werden (z. B. Kreise/Ellipsen, vgl. MESCHENMOSER / PRÖLL 2012b).
- Man kann die Signifikanz von Isoglossenbündeln prüfen, indem festgestellt wird, inwieweit diese höhere Werte besitzen, als sie im Durchschnitt zufällig auftreten. So lässt sich ein beliebiger Grenzverlauf (Fluss, alte Territorialgrenze o. ä.) darauf hin untersuchen, ob die dort auftretenden Isoglossen mehr als zufällig sind oder nicht (PICKL 2013a).

Die Anwendung der *Faktorenanalyse* auf räumlich aufgelöste Daten (und die Kartierung ihrer Ergebnisse) bietet einen neuen, differenzierteren Blick in die Zusammenhänge der einzelnen Varianten (in der Fläche betrachtet) und die Zusammensetzung lokaler Variation (am Ort betrachtet):

- In der Fläche konstituieren sich Dialektgebiete in der Faktorenanalyse direkt aus statistischen Zusammenhängen (Kookkurrenzen) in den Daten (vgl. PICKL 2013a, 2013c; PRÖLL 2015). Die Mundartregionen („Faktoren“) definieren sich nicht mehr wie bisher üblich durch Grenzlinien an den Rändern, sondern durch die in ihnen vorhandenen Gemeinsamkeiten, die ohne Vorgaben von außen durch *data mining* gefunden wurden. Dabei erfahren wir auch quantitativ, wie groß die „Stärke“ des Faktors, das heißt, wie hoch die Anzahl der sprachlichen Kookkurrenzen im Vergleich zu den anderen Gebieten, die ermittelt wurden, ist. Die Gebiete sind dadurch unscharf umrissen und können überlappen. So können wir die vorhandenen Sachverhalte sehr viel besser beschreiben als andere Modelle der Dialektgliederung.
- Am Ort erhalten wir Vergleichswerte, die darstellen, in welchem Maße die Faktoren den vorhandenen Ort beschreiben können, d. h. wie groß der Anteil der jeweiligen Faktoren in einem Ortsdialekt ist. (Ottobeuren hat z. B. die folgenden Werte: Mittelostschwäbisch 34%, Allgäuisch 19%, Memminger Raum 4%, Nordostschwäbisch 2 %, sonstige 2%, unbestimmt 39%.) Mit diesen Werten lässt sich sehr schön und besser als bisher möglich das Kontinuum der Varietäten im geographischen Raum darstellen, z. B. in Form von Querschnittsgraphiken (PICKL 2013a, 2013c; PRÖLL 2015; PICKL / PRÖLL i. E.).
- Schließlich erlaubt die Faktorenanalyse die Rückbeziehung der Gebiete auf die damit assoziierten Varianten, was nicht nur Rückschlüsse auf die Relevanz einzelner Varianten für die globale Struktur zulässt, sondern auch Einblicke in die Beziehungen zwischen sprachlichen Varianten in sprachgeographischer Hinsicht, d. h. welche Varianten sich im Raum ähnlich verhalten, welche nicht, was wiederum Aufschluss über sprachgeographisch relevante linguistische Relationen gibt (PICKL 2013a, c; PRÖLL 2015; PICKL / PRÖLL i. E.).

Die hier dargestellten Verfahren und Methoden sind in ihrer Mehrzahl in unser Programmsystem *GeoLing* eingegangen und stehen über dieses zugänglich interessierten WissenschaftlerInnen zur Verfügung. Sie sind bisher vor allem auf Daten des SBS (wie hier beschrieben), aber auch des SDS (s. http://www.dialektkarten.ch/dominance.de.html) erfolgreich angewandt worden und werden zurzeit z. B. an Daten des Projekts *Deutsch Heute* und dem *AdA* getestet.

Publikationen aus dem Projekt

Dissertationen

PICKL, Simon (2013a): *Probabilistische Geolinguistik. Geostatistische Analysen lexikalischer Distribution in Bayerisch-Schwaben.* Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte, 154).

PRÖLL, Simon (2015): *Raumvariation zwischen Muster und Zufall. Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben.* Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte, 160).

RUMPF, Jonas (2010): *Statistical models for geographically referenced data. Applications in tropical cyclone modelling and dialectology*. Dissertation, Universität Ulm.

Aufsätze

MESCHENMOSER, Daniel / PRÖLL, Simon (2012a): Using fuzzy clustering to reveal recurring spatial patterns in corpora of dialect maps. In: *International Journal of Corpus Linguistics* 17/2, 176–197.

MESCHENMOSER, Daniel / PRÖLL, Simon (2012b): Automatic detection of radial structures in dialect maps: determining diffusion centers. In: *Dialectologia et Geolinguistica* 20, 71–83.

PICKL, Simon (2013b): Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry. In: *Literary and Linguistic Computing* 28/1, 63–81.

PICKL, Simon (2013c): Verdichtungen im sprachgeographischen Kontinuum. In: *Zeitschrift für Dialektologie und Linguistik* 80/1, 1–35.

PICKL, Simon (2014). Dialekträume 'unter der Oberfläche'. Nicht-dominante wortgeographische Strukturen in Bayerisch-Schwaben. In: BÜHLER, Rudolf / BÜRKLE, Rebekka / LEONHARDT, Nina Kim (Hrsg.): *Sprachkultur – Regionalkultur. Neue Felder kulturwissenschaftlicher Dialektforschung.* Tübingen: TVV (Studien & Materialien des Ludwig-Uhland-Instituts der Universität Tübingen, 49), 198–217.

PICKL, Simon / RUMPF, Jonas (2011): Automatische Strukturanalyse von Sprachkarten. Ein neues statistisches Verfahren. In: GLASER, Elvira / SCHMIDT, Jürgen Erich / FREY, Natascha (Hrsg.): *Dynamik des Dialekts – Wandel und Variation. Akten des 3. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD)*. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte, 144), 267–285.

PICKL, Simon / RUMPF, Jonas (2012): Dialectometric Concepts of Space: Towards a Variant-Based Dialectometry. In: HANSEN, Sandra / SCHWARZ, Christian / STOECKLE, Philipp / STRECK, Tobias (Hrsg.): *Dialectological and folk dialectological concepts of space.* Berlin, Boston: de Gruyter (linguae & litterae, 17), 199–214.

PICKL, Simon / PRÖLL, Simon (i. E.). Die Dialekte Bayerisch-Schwabens als Spiegel historischer Kommunikationsräume. In: DAUSER, Regina / FASSL, Peter / SCHILLING, Lothar (Hrsg.): *Wissenszirkulation auf dem Land vor der Industrialisierung.* Augsburg: Wißner (Documenta Augustana).

PICKL, Simon / SPETTL, Aaron / PRÖLL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2014): Linguistic distances in dialectometric intensity estimation. In: *Journal of Linguistic Geography* 2/1, 25–40.

PRÖLL, Simon (2013): Detecting structures in linguistic maps – Fuzzy clustering for pattern recognition in geostatistical dialectometry. In: *Literary and Linguistic Computing* 28/1, 108–118.

PRÖLL, Simon (2014): Stochastisch gestützte Methoden der Dialektdifferenzierung. In: HUCK, Dominique (Hrsg.): *Dialekte im Kontakt. Beiträge zur 17. Arbeitstagung für alemannische Dialektologie in Straßburg vom 26.–28.10.2011*. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte, 155), 233–246.

PRÖLL, Simon / PICKL, Simon / SPETTL, Aaron (2015): Latente Strukturen in geolinguistischen Korpora. In: ELMENTALER, Michael / HUNDT, Markus / SCHMIDT, Jürgen Erich (Hrsg.): *Deutsche Dialekte. Konzepte, Probleme, Handlungsfelder. Akten des 4. Kongresses der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD).* Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik, Beihefte, 158), 247–258.

PRÖLL, Simon / PICKL, Simon / SPETTL, Aaron / SCHMIDT, Volker / SPODAREV, Evgeny / ELSPAß, Stephan / KÖNIG, Werner (i. E.): Neue Dialektometrie mit Methoden der stochastischen Bildanalyse. In: KEHREIN, Roland / LAMELI, Alfred / RABANUS, Stefan (Hrsg.): *Regionale Variation des Deutschen – Projekte und Perspektiven*. Berlin, New York: de Gruyter.

RUMPF, Jonas / PICKL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2009): Structural analysis of dialect maps using methods from spatial statistics. In: *Zeitschrift für Dialektologie und Linguistik* 76/3, 280–308.

RUMPF, Jonas / PICKL, Simon / ELSPAß, Stephan / KÖNIG, Werner / SCHMIDT, Volker (2010): Quantification and statistical analysis of structural similarities in dialectological area-class maps. In: *Dialectologia et Geolinguistica* 18, 73–100.