Simon Pickl / Jonas Rumpf

Automatische Strukturanalyse von Sprachkarten – Ein neues statistisches Verfahren[1]

1. Einführung und Motivation

Im Jahr 1898 führte Carl Haag eine neue Methode ein, um Unterschiede zwischen Dialekten auf einer Karte darzustellen (HAAG 1898). Dabei gibt die Stärke einer Linie an, wie unterschiedlich die Dialekte an den Orten sind, zwischen denen sie verläuft. Dazu werden die Daten aus mehreren Sprachkarten, von denen jede die Realisierungen eines sprachlichen Merkmals im Raum darstellt (im Folgenden „Merkmalskarten“), zusammengefasst. Dies war der erste Schritt in Richtung einer quantitativen Dialektologie, die damit gegenüber anderen linguistischen Disziplinen wie Lexikographie, Phonetik oder historischer Linguistik nachzog (vgl. KÖHLER/ALTMANN/PIOTROWSKI 2005, BEST 2006). Die dialektometrischen Verfahrensweisen, die seitdem entwickelt wurden, haben eine Fülle an Methoden hervorgebracht, die ausnahmslos auf dem Messen von Dialektabständen (dem Grad der Unterschiedlichkeit zweier Ortsdialekte) basieren. Bis zu den 1970ern geschah dies, indem man die dialektalen Unterschiede zwischen allen Nachbarorten zählte. Diese Methode geht auf Haag zurück und wurde von Jean Séguy für den „Atlas linguistique et ethnographique de la Gascogne“ (SÉGUY 1965–1973), das erste größere dialektometrische Projekt, verwendet. In den 1970ern erweiterte Hans Goebl diese Methode, indem er die Unterschiede zwischen allen möglichen Ortspaaren – nicht nur Nachbarorten – zählte und ein breites Spektrum an differenzierten Darstellungsweisen und Analysemethoden in die Dialektometrie einführte, wie z.B. verschiedene Arten von Clusteranalysen oder Kohärenztests (siehe z.B. GOEBL 1994, 2001, 2006, 2007). Die darauf folgenden dialektometrischen Ansätze basieren alle im Wesentlichen auf Goebls Methode (KELLE 1986, HUMMEL 1993, SCHILTZ 1996 u.a.). Der nächste große Schritt wurde von John Nerbonne und Wilbert Heeringa vollzogen, die eine neue Form der Abstandsmessung in die Goebl’sche Dialektometrie einführten: Sie verwendeten eine angepasste Version des sogenannten Levenshtein-Abstandsmaßes, um phonetische Abstände zu berechnen, was es erstmals ermöglichte, graduelle Ähnlichkeiten zwischen Belegen zu berücksichtigen und nicht nur, wie bisher, Gleichheit oder Ungleichheit. Außerdem führten sie neue Analysemethoden wie multidimensionale Skalierung und Faktoranalyse ein (siehe z.B. NERBONNE/HEERINGA 1998, HEERINGA 2004, NERBONNE 2006).

Die klassische Dialektometrie befasst sich jedoch nach wie vor ausschließlich mit der globalen

1 Der vorliegende Beitrag ist in weiten Teilen eine gekürzte deutsche Fassung von RUMPF/PICKL/ELSPAß/KÖNIG/-SCHMIDT 2010a. Die vorgestellten Methoden stellen Ergebnisse des DFG-Forschungsprojekts „Neue Dialektomentrie mit Methoden der stochastischen Bildanalyse“ (http://www.philhist.uni-augsburg.de/de/lehrstuehle/germanistik/sprachwissenschaft/projekte/dialektometrie/) des Lehrstuhls für deutsche Sprachwissenschaft der Universität Augsburg und des Instituts für Stochastik der Universität Ulm dar.

Frage nach der dialektalen Einteilung des jeweiligen Untersuchungsgebiets oder mit den Beziehungen zwischen Dialekten oder Teilsystemen von Dialekten. Da hierbei immer die Daten aller Merkmalskarten des jeweiligen (Teil-)Korpus aggregiert werden, bleiben die strukturellen Charakteristiken der einzelnen Merkmalskarten unberücksichtigt. Beim Blättern durch die Karten eines Dialektatlas zeigt jedoch schon ein flüchtiger Blick, dass die Verteilungsmuster der einzelnen Varianten nicht nur leichte Schwankungen um die durch Aggregation erzielte Dialekteinteilung darstellen, sondern dass sie im Gegenteil individuelle, äußerst variable Muster ausbilden. Die Unterschiede zwischen einzelnen Karten sind mitunter so groß, dass sie nicht auf bloße zufällige Abweichungen zurückzuführen sein können.

Auf manchen Karten ergeben die Verteilungen der Varianten kleine, kompakte und kaum vermischte Gebiete, auf anderen hingegen größere, überlappende Gebiete, und es gibt auch Karten, auf denen man nicht viel mehr als Chaos ausmachen kann. Manche topologische Gegebenheiten wie Bergrücken oder Flüsse decken sich mit Grenzen zwischen sprachlichen Merkmalen, andere geographische Strukturen wiederum stimmen fast nie mit sprachlichen Grenzen[2] überein. Mehrere Dialektologen haben die verschiedenen Strukturen und Muster in den Verteilungen von Varianten untersucht und klassifiziert (z.B. WENZEL 1930, 107–110; FRINGS 1956; BACH 1969, 39–226; HILDEBRANDT 1983), doch bislang gibt es kein quantitatives Verfahren, das eine systematische Überprüfung der dort formulierten Hypothesen erlaubt.

Die strukturellen Eigenschaften einer sprachlichen Merkmalskarte sind das Ergebnis der Ausbreitung der dort verzeichneten Varianten im Raum. Verschiedene Faktoren, die dabei eine Rolle spielen, sind vorstellbar: Gebrauchshäufigkeit, semantische Gruppe oder geographische Bedingungen sind nur einige davon. In irgendeiner Weise muss das Aussehen einer Karte mit solchen Variablen zu tun haben. Kurz gesagt: Es muss Gründe dafür geben, dass die Karten so unterschiedlich ausfallen. Eine quantitative Analyse eines Korpus von Merkmalskarten sollte uns helfen, die Frage zu beantworten, welche Variablen eine Rolle beim Aufbau der Karten spielen und wie sie die Verteilungen der Varianten beeinflussen.

Um die Mechanismen zu erforschen, die die geographische Verbreitung der Varianten steuern, müssen wir zuerst einen Weg finden, um die strukturellen Charakteristiken der Karten zu beschreiben und zu quantifizieren. Konzepte wie „Komplexität“ oder „Homogenität“ können dabei hilfreich sein. Zunächst müssen sie definiert und Skalen festgelegt werden. Außerdem muss man einen Weg finden, sie möglichst objektiv und automatisch, d.h. reproduzierbar, zu messen. Wenn man eine große Anzahl Karten in Bezug auf diese Charakteristiken untersucht hat, kann man versuchen, die erhaltenen Werte mit linguistischen Eigenschaften der sprachlichen Merkmale zu

2 Wir vermeiden den Ausdruck „Isoglosse“, da er aus verschiedenen Gründen ungenau ist (vgl. HÄNDLER/WIEGAND 1982; SCHNEIDER 1988, 177–179), und verwenden stattdessen die weniger problematische Bezeichnung „Grenze“.

korrelieren, um über die Einflussfaktoren von Sprachvariation im Raum Aufschluss zu gewinnen.

2. Ansatz

Die in diesem Beitrag vorgestellte Methodik beruht auf einigen theoretischen Annahmen, die die Daten in Dialektatlanten betreffen. Da diese Atlanten auf Antworten beruhen, die ausgewählte Repräsentanten ausgewählter Orte auf ausgewählte Fragen gegeben haben, reflektieren sie nur einen Ausschnitt der sprachlichen Realität. Selbst wenn wir annehmen, dass die Belege mehr sind als einfache Stichproben, da eine Gewährsperson für mehr Personen als nur für sich selbst sprechen kann,[3] müssen wir akzeptieren, dass die Daten einem gewissen statistischen „Rauschen" unterliegen. Mit Hilfe statistischer Methoden kann jedoch eine Annäherung an die „tatsächliche" Situation erzielt werden, was sich als äußerst hilfreich für die Ermittlung der strukturellen Charakteristiken von Sprachkarten erweisen wird.

Wenn man die aufgezeichneten sprachlichen Belege als Stichproben wertet, so kann man auf Grundlage der geographischen Verteilung einer Variante die Wahrscheinlichkeit schätzen, mit der diese an einem bestimmten Ort zu erwarten ist. Damit können wir einem beliebigen Ort Werte zuweisen, welche die geographische Verbreitung der Varianten in seiner Umgebung beschreiben, sogar wenn an dem Ort selbst keine Belege vorhanden sind. Ein Blick in die Nachbarschaft eines Ortes gibt uns dabei Aufschluss über die Aussagekraft eines einzelnen Beleges. Wenn beispielsweise ein einzelner „Ausreißer" in einem ansonsten gleichmäßigen Gebiet liegt, so ist es sehr wahrscheinlich, dass eine andere Gewährsperson am selben Ort (oder sogar dieselbe Person zu einer anderen Zeit oder in einer anderen Situation) die Variante der umliegenden Orte geliefert hätte. Gleichzeitig ist zu erwarten, dass einige wenige Leute der Nachbarorte auch mit der Variante des Ausreißers hätten aufwarten können. Dies ist ein sehr einfaches Beispiel, um unsere Grundannahme zu illustrieren: Je mehr Belege einer bestimmten Variante in der Umgebung eines Ortes auftreten, um so wahrscheinlicher ist es, dass diese Variante auch an dem Ort selbst vorkommt, auch wenn der tatsächlich befragte Informant eine andere angegeben hat. Dies bedeutet für die Schätzung der Auftretenswahrscheinlichkeit einer Variante, dass die Belegdaten der umliegenden Orte berücksichtigt werden; ihr Einfluss nimmt dabei mit wachsender Distanz ab. Dementsprechend können viele abweichende Varianten in der Umgebung eines Ortes dessen Belegsituation sogar „umkehren". Umgekehrt schwächt dies aber auch die umliegenden Varianten selbst, da der Ausreißer auch auf sie einen gewissen, wenn auch kleinen, Einfluss hat. Diese geographieabhängige Schätzung können wir vornehmen, da der Kontakt zwischen Sprechern nicht nur innerhalb von, sondern auch zwischen Ortschaften stattfindet. Je höher die geographische Distanz zwischen zwei Orten, um so niedriger ist der Grad an zu erwartendem Sprachkontakt.[4]

---

3 Vgl. „Der Informant als Experte" (SBS 1996–2009, Bd. 1, 20–21).

4 Das ist natürlich eine starke Vereinfachung, denn geographische Nähe steht nicht in unmittelbarem Zusammenhang

Die Ergebnisse dieser Methode können als ein geschätzter Prozentsatz für jede Variante an jedem Ort interpretiert werden, der angibt, wie wahrscheinlich es ist, dass eine Variante als Antwort einer Gewährsperson erscheinen würde, wenn man alle möglichen Gewährsperson befragte. Dies gilt natürlich nur für den Teil der Bevölkerung, der als Informantenpool bestimmt wurde. Die Berechnungen für die Schätzung basieren ausschließlich auf den tatsächlichen Belegen (die, von linguistischem Vorwissen gesteuert, in Varianten eingeteilt wurden) und den geographischen Koordinaten der Belegorte. Das genaue mathematische Vorgehen wird in Abschnitt 4 erläutert.

Wie sind die erhaltenen Werte zu interpretieren? Wenn der Prozentsatz für eine bestimmte Variante an einem Ort relativ klein ist, so können wir dennoch annehmen, dass sie zumindest im passiven Sprachgebrauch einer Handvoll möglicher Informanten vorkommt. Wenn keine der vorkommenden Varianten deutlich über die anderen dominiert, befinden sie sich in einem Zustand der Konkurrenz.[5] Somit dienen die Prozentsätze, die für jeden Ort und jede Variante ermittelt werden, einem doppelten Zweck: Erstens sind sie näher an der „tatsächlichen“ Situation, d.h. der Wahrscheinlichkeit, mit der eine bestimmte Variante an einem bestimmten Ort auftritt, und zweitens erlauben sie uns, die strukturellen Charakteristiken auf einer Karte lokal zu beschreiben. Ein Ort, an dem eine Variante deutlich dominiert, ist Teil eines eher homogenen Bereichs, während ein Ort, an dem mehrere Varianten ungefähr gleich stark vertreten sind, in einem mehr durchmischten, heterogenen Teil der Karte liegt. Diese Werte können dann zusammengenommen einen Eindruck von der Homogenität einer ganzen Karte vermitteln. Sie zeigen uns auch, welche Variante die vorherrschende in einem bestimmten Teilgebiet einer Karte ist, was uns erlaubt, die Karte in Dominanzareale von Varianten einzuteilen – und somit automatisch Flächenkarten aus den Daten zu erzeugen. Sie dienen als Zwischenschritt für weitergehende Analysen und bieten – sozusagen als Nebenprodukt – die Möglichkeit, reproduzierbar und konsistent Punktsymbolkarten in Flächenkarten zu zerlegen.[6]

---

mit Sprachkontakt. Es war ja gerade eines der wichtigsten Ziele der klassischen Dialektologie zu zeigen, was Sprachkontakt fördert und was ihn hemmt. Da diese Frage noch nicht abschließend beantwortet werden kann (zumindest nicht auf quantitativer Basis), haben wir keine solide Grundlage, um Sprachkontakt zu quantifizieren. Aus diesem Grund werden wir uns zunächst mit der geographischen Distanz als einem sehr groben, wenn auch nicht gänzlich ungeeigneten Indikator für Sprachkontakt begnügen müssen.

5 Ein solcher Wettbewerb zwischen zwei oder mehr Varianten in einer Sprachgemeinschaft wird üblicherweise als ein Indikator für gerade stattfindenden Sprachwandel gesehen: Eine Variante ist die ältere, die andere die jüngere. Früher oder später wird eine der beiden sich durchsetzen, entweder weil die jüngere die ältere verdrängt, oder weil die ältere sich behaupten kann (vgl. HAAS 1978, 34–80; KÖNIG 1982, 464). Einen stabilen Zustand gibt es nur dann, wenn eine Variante deutlich dominiert und keine ernstzunehmenden Konkurrenten hat. Ein Sprachwandelmodell, das die graduelle Verdrängung einer älteren Sprachform durch eine neue quantifiziert, ist als das Piotrowski-Gesetz bekannt (vgl. ALTMANN 1983; PIOTROWSKI/BEKTAEV/PIOTROWSKAJA 1985, 81–100; LEOPOLD 2005; BEST 2006, 106–123).

6 Ohne Frage sind die so erzeugten Flächenkarten vereinfachte Visualisierungen von geographischen Dialektdaten. Es gibt zweifellos detailliertere und besser geeignete Methoden, um sprachliche Daten auf einer Karte abzubilden (s. z.B. ELSPAß/KÖNIG (2008) und PRÖLL (i.V.)). Unsere Flächenkarten dienen jedoch in erster Linie der Überprüfung der Ergebnisse unseres Verfahrens und sind nicht für die klassische qualitative Interpretation von Dialektkarten gedacht, die sich oft mit Mikrovariation befasst. Sie sind aber reproduzierbar erzeugbar und bieten eine sehr

3. Daten

Alle Beispiele und Ergebnisse beruhen auf den Daten des Sprachatlas von Bayerisch-Schwaben (SBS), der in den Jahren 1984 bis 2005 unter der Leitung von Werner König an der Universität Augsburg erstellt wurde. Mit ca. 2.700 Karten in 14 Bänden ist er der umfangreichste Dialektatlas im deutschen Sprachraum. Das Untersuchungsgebiet erstreckt sich 90 km von West nach Ost und etwa 150 km von Nord nach Süd und umfasst 272 Erhebungsorte. Für jedes sprachliche Merkmal liegen im Normalfall für jeden Ort Antworten von einer Gewährsperson vor. Zusätzliche Informanten wurden jeweils nur dann hinzugezogen, wenn der erste Informant eine Frage nur unzureichend beantworten konnte. Da diese große Menge an Daten auch elektronisch vorliegt, ist der SBS ideal für die Erprobung neuer dialektometrischer Methoden.

Für jedes sprachliche Merkmal werden in der Datenbank jedem Ort so viele Einträge zugewiesen, wie verschiedene Antworten von den Gewährspersonen vorliegen. Ein Eintrag besteht aus einer phonetischen Transkription des Beleges und einem Code für das in der Kartierung verwendete Symbol. Dieser Code ist die Grundlage für die Identifikation der Varianten, denn er beinhaltet eine Vorklassifikation der Belege nach linguistischen Gesichtspunkten, was die Entscheidung, ob zwei Belege zur selben Variante gehören, erheblich erleichtert. Die Anzahl der Einträge entspricht dabei nicht der Zahl der Informanten, sondern der Anzahl der unterschiedlichen Varianten, die pro Ort vorliegen. Die Datenbank enthält keinerlei Informationen darüber, wieviele Gewährspersonen zu einer bestimmten Variante beigetragen haben, so dass sie in Ermangelung anderer Anhaltspunkte als gleichwertig behandelt werden müssen.

Für die vorliegende Studie haben wir uns auf Wortkarten beschränkt, da bei phonetischen und morphologischen Karten (syntaktische kommen im SBS nicht vor) zusätzliche Probleme bei der Einteilung der Varianten auftreten (z.B. Daten auf Ordinalniveau), die noch nicht gelöst sind. Das von uns betrachtete Korpus umfasst mit 823 Karten fast alle lexikalischen Karten des SBS.

4. Mathematische Methodik

4.1 Intensitätsschätzung und Erstellung von Flächenkarten

Unser erster Schritt zur Strukturanalyse von Dialektkarten besteht in der automatisierten Erstellung von Flächenkarten aus den ursprünglichen Daten. Da solche Daten meist – wie auch im SBS – in Form von Punktsymbolkarten vorliegen, liegt es nahe, sie als Gruppierungen von Punktmustern in der Ebene zu betrachten. Zur Untersuchung dieser Punktmuster verwenden wir Methoden der räumlichen Statistik (s. z.B. BADDELEY et al. 2006, DIGGLE 2003, ILLIAN et al. 2008, STOYAN/STOYAN 1994).

---

zugängliche Visualisierung von Mesovariation im SBS und Sprachatlanten ähnlicher Größe, indem sie den Grad an Variabilität in allen Teilen der Karte abbilden.

Normalerweise dienen Flächenkarten dazu, dem Betrachter eine Aufteilung des Beobachtungsfensters in Flächen, die für einzelne Varianten stehen, zu bieten. Wir wollen sie darüber hinaus auch als Vorbereitung der Daten zur weiteren Untersuchung betrachten. In einem ersten Schritt teilen wir eine Punktsymbolkarte dazu in Vorkommenskarten für jede einzelne Variante auf. Auf jeder dieser Vorkommenskarten tritt dann nur eine einzige Variante auf, und zwar an den Orten, an denen sie belegt ist. An allen anderen Orten verzeichnet die entsprechende Vorkommenskarte keine Variante. Auf diese Weise erhalten wir Punktmuster von Auftretenspunkten für alle Varianten. Da es möglich ist, dass auf einer Punktsymbolkarte mehrere Symbole (also Varianten) an einem Ort verzeichnet sind, weisen wir jeder Variante x ein Gewicht $l_x(t)$ für jeden Befragungsort t zu: Ist Variante *x* die einzige, die an Ort *t* auftritt, so erhält sie das Gewicht $l_x(t) = 1$. Anderenfalls ist ihr Gewicht gegeben als ihr Anteil an den Gesamtbelegen am Ort, d.h. wenn *x* z.B. eine von drei am Ort *t* auftretenden Varianten ist, so ist $l_x(t) = \frac{1}{3}$. Entsprechend wird allen Orten, an denen *x* in den ursprünglichen Daten nicht auftritt, der Wert $l_x(t) = 0$ zugewiesen.

Nachdem in der beschriebenen Weise Vorkommenskarten für alle Varianten einer Karte erstellt wurden, wird nun für jede dieser Vorkommenskarten ein sogenanntes Intensitätsfeld geschätzt. Vereinfacht gesagt beschreibt das Intensitätsfeld einer Variante für jeden beliebigen Ort die Wahrscheinlichkeit, dass diese Variante dort auftritt. Die Intensität wird hier mittels eines sogenannten zweidimensionalen Kerndichteschätzers geschätzt. Detaillierte Darstellungen der Methoden der Kerndichteschätzung finden sich z.B. in SCOTT (1992) und SILVERMAN (1986). An dieser Stelle werden wir nur kurz auf die für unsere Untersuchungen relevanten Aspekte eingehen. Der Kerndichteschätzer $u_x(t_i)$ für die Intensität von Variante *x* an der Stelle t_i ist definiert als

(1)

Dabei stehen $t_1, \ldots t_n$ für die *n* Erhebungsorte und $d(t_i,t_j)$ für den geographischen Abstand der Orte t_i und t_j. Der Parameter $h > 0$ heißt Bandbreite des Kerndichteschätzers, und die monoton fallende Funktion *K* ist der sogenannte Kern. Das Prinzip der Kerndichteschätzung ist es, jedem Ort, an dem die Variante *x* aufgetreten ist, eine gewisse „Wahrscheinlichkeitsmasse“ zuzuweisen und diese Masse dann so zu verteilen, dass weiter entfernt gelegene Orte weniger Masse erhalten. Die Form der Kernfunktion bestimmt dabei, wie die Masse im Raum verteilt wird, während die Bandbreite *h* angibt, wie stark die Masse mit der Entfernung abnimmt. Der Schätzwert für die Intensität einer bestimmten Variante an einem bestimmten Ort ergibt sich nun, indem man an diesem Ort die von allen Erhebungsorten stammende Masse aufsummiert. In unserem Ansatz wird die Masse, die einem Ort t_i für Variante *x* zugewiesen wird, noch zusätzlich mit dem Auftretensgewicht $l_x(t_j)$ gewichtet.

So erzeugt das Auftreten von *x* gemeinsam mit einer weiteren Variante am Ort t_j nur die Hälfte der Masse am Ort t_i, die von einem alleinigen Auftreten von Variante *x* erzeugt würde.

Bei näherer Betrachtung von Formel (1) wird klar, dass sowohl für die Kernfunktion *K* als auch für die Bandbreite *h* eine geeignete Wahl getroffen werden muss. Einige Voruntersuchungen haben gezeigt, dass für unseren Zweck der sogenannte Normalverteilungskern am besten geeignet ist. Die Bandbreite bestimmt aus linguistischer Sicht, wie stark der Einfluss eines Befragungsortes auf seine Umgebung ist, d.h. wie stark der Einfluss ist, den eine an einem Ort vorkommende Variante auf dessen Umgebung ausübt. In Bezug auf den erwähnten Sprachkontakt beschreibt die Bandbreite, in welchem Ausmaß ein bestimmtes sprachliches Merkmal bei der Kommunikation zwischen verschiedenen Orten verwendet wird. Für unsere Untersuchungen haben wir das sogenannte Least-Squares-Cross-Validation-Verfahren zur Wahl einer „optimalen“ Bandbreite gewählt (s. z.B. SILVERMAN (1986, 87f.)). Damit könnte man für jede einzelne Variante eine individuelle Bandbreite erhalten. Da aber davon auszugehen ist, dass der Einfluss der Befragungsorte aufeinander nicht von der jeweiligen Variante, sondern von der jeweils betrachteten Variable abhängt, bestimmen wir stattdessen eine einzige globale Bandbreite für alle Varianten eines Merkmals.

In Abb. 1 ist die Punktsymbolkarte für ‘Kartoffelkraut’ aus dem SBS wiedergegeben. Abb. 2 zeigt den zugehörigen Intensitätsschätzer für die Variante *Kraut*; diese Variante wird durch ein gestreiftes gleichschenkliges Dreieck symbolisiert. Die Befragungsorte des SBS sind in Abb. 2 mit schwarzen Punkten markiert. Gebiete, die in dieser Abbildung mit dunkleren Blautönen markiert sind, weisen hohe geschätzte Intensitäten der Variante *Kraut* auf, hellere Werte stehen für geringere Intensitäten, und weiß erscheinende Bereiche haben Intensitätswerte nahe Null. Es ist zu beachten, dass die Intensität nur für die 272 Ortschaften des SBS geschätzt wird, nicht jedoch für Bereiche dazwischen. In der Darstellung gibt der Farbwert des einen Ort umgebenden Polygons den Intensitätswert des Ortes wieder. Die Polygone kommen durch die von allen Orten erzeugte Voronoi-Tessellation (auch bekannt als Thiessen-Polygone) zustande. Dieses Mosaik weist jeden Punkt im Beobachtungsfenster einfach demjenigen Befragungsort zu, der ihm am nächsten liegt. Angesichts der Tatsache, dass über den Raum zwischen den Befragungsorten keine weiteren Informationen vorliegen, erscheint dies die naheliegendste Vorgehensweise. Eine detaillierte Darstellung der Eigenschaften von Voronoi-Tesselationen findet sich z.B. in OKABE et al. (2000).

Bei einer Gegenüberstellung der Abb. 1 und 2 lässt sich gut erkennen, dass die höchsten geschätzten Intensitätswerte für die Variante *Kraut* in der Nordostecke des Untersuchungsgebiets auftreten, wo das entsprechende Symbol in Abb. 1 als einziges auftritt. In der Mitte oder im Nordwesten, wo *Kraut* mit anderen Varianten durchmischt ist, ist die Intensität niedriger, und im Süden, wo diese Variante nicht auftritt, ist die Intensität nahezu Null. Diese Beobachtung passen gut zu der oben

erwähnten Interpretation der Intensität einer Variante.

In einem nächsten Schritt werden Flächenkarten aus den geschätzten Intensitätsfeldern erstellt. An jedem Befragungsort im Untersuchungsgebiet liegen nun Intensitäten für alle Varianten vor. Wir weisen daher jeden Ort derjenigen Variante zu, die dort die höchste geschätzte Intensität hat. Dies ist naheliegend, da diese Variante ja diejenige ist, deren Auftreten am betrachteten Ort als am wahrscheinlichsten angesehen werden kann. Da somit jeder Befragungsort genau einer Variante zugewiesen ist, wird das Untersuchungsgebiet deterministisch in Flächen zerlegt, die für verschiedene Varianten stehen. Die Grenzen zwischen diesen Gebieten markieren dabei die Stellen, an denen eine Variante wahrscheinlicher wird als die andere. Im Folgenden werden wir mit $x(t)$ die einem Ort t zugewiesene Variante bezeichnen. Des Weiteren wird $T(x)$ für die Menge aller Orte stehen, die der Variante x zugewiesen wurden, und $|T(x)|$ für die Anzahl der Orte in $T(x)$. Die Grenzen ergeben sich dabei aus den Kanten zwischen den Voronoi-Zellen von Orten, die verschiedenen Varianten zugewiesen werden.

Abb. 3 zeigt die Flächenkarte, die sich aus den Daten für ‘Kartoffelkraut’ ergibt. Verschiedene Varianten sind durch verschiedene Farbtöne dargestellt; das Gebiet der Variante *Kraut* ist türkis markiert. Grenzen zwischen Gebieten sind orange gekennzeichnet. Obwohl in der ursprünglichen Punktsymbolkarte neun verschiedene Varianten auftreten, sind in Abb. 3 nur vier Gebiete zu sehen. Dies ist dadurch zu erklären, dass die restlichen fünf Varianten an keinem der Befragungsorte die höchste der geschätzten Intensitäten aufweisen. Mit anderen Worten: Diese Varianten sind nicht häufig genug oder nicht kompakt genug angeordnet, um ihre eigenen Gebiete auszubilden.

Würde man einen Ort und seine Umgebung jeweils einfach einer Variante zuweisen, würde man jegliche Information über die Höhe der jeweils geschätzten Intensitäten der anderen Varianten verlieren. Daher variieren wir die Helligkeit der Farbe an einem Ort gemäß der Intensität der jeweils dominanten Variante: Je höher die Dominanz, desto dunkler die Farbe. Genauer gesagt wählen wir die Helligkeit an einem Ort t proportional zu

$$b(t) = \frac{\max_x u_x(t)}{\sum_x u_x(t)} \,. \tag{2}$$

Diese Größe $b(t)$ kann beschrieben werden als derjenige Anteil der gesamten am Ort t geschätzten Intensitäten, der von $x(t)$ stammt. Bei einer parallelen Betrachtung der Abb. 1 und 3 kann man erkennnen, dass die dunkelsten Farben in den Bereichen auftreten, wo eine Variante nahezu frei von jeder Durchmischung mit anderen Varianten auftritt, so z.B. im Südosten und Nordosten des Untersuchungsgebietes. Dort, wo mehrere Varianten auftreten, sind die Farben deutlich blasser. Wie gut zu sehen ist, ist dies auch der Fall entlang der Grenzen. Dies war natürlich zu erwarten, denn solche Grenzen „[do] not mark a sharp switch from one word to the other, but the center of a

transitional area where one comes to be somewhat favored over the other“ (vgl. FRANCIS 1983, 5).

4.2 Charakteristiken von Flächenkarten

Flächenkarten sind ein etabliertes Werkzeug für dialektologische Untersuchungen. Flächenkarten, die wie in Abschnitt 4.1 beschrieben erzeugt werden, können darüber hinaus als Vorbereitung der Daten der Punktsymbolkarten für eine weitergehende statistische Untersuchung dienen. Charakteristiken wie $b(t)$, die bei der Erstellung der Flächenkarten berechnet werden, können dabei helfen, die strukturellen Eigenschaften dieser Karten zu beschreiben. Freilich ist eine Betrachtung dieser Kennzahlen nur dann sinnvoll, wenn eine sinnvolle linguistische Interpretation für sie gefunden werden kann.

Als erste Kennzahl zur Charakterisierung einer Flächenkarte berechnen wir die Gesamtgrenzlänge zwischen den Flächen verschiedener Varianten. Es erscheint einleuchtend, diese Größe als Indikator für die Komplexität einer Karte zu betrachten, denn eine größere Gesamtgrenzlänge bedeutet häufigere Wechsel zwischen dominanten Varianten; ein Umstand, der die Karte im Ganzen komplexer erscheinen lässt. Nicht berücksichtigt werden dabei Intensitätsschwankungen innerhalb der entstandenen Flächen, deren Untersuchung mithilfe einer anderen Kenngröße weiter unten erläutert wird. Es ist wichtig zu beachten, dass selbst Karten mit einer identischen Anzahl von Flächen deutlich unterschiedliche Gesamtgrenzlängen aufweisen können: Bei glatteren und zusammenhängenden Gebieten wird eine deutlich geringere Gesamtgrenzlänge entstehen als bei Gebieten, bei denen dies nicht der Fall ist. Die Grenzen, die auf der Flächenkarte in Abb. 3 orange markiert sind, haben eine Gesamtlänge von 240,8 km. Theoretisch könnten die Grenzen auf den Karten des SBS bis zu ca. 8.644 km lang werden, was aber in der Praxis nie auftritt, da nie zwei Orte mit derselben dominanten Variante benachbart sein dürften. In Abschnitt 5 werden wir Ergebnisse diskutieren, die eine bessere Einordnung dieser Zahlen erlauben.

Eine zweite Strukturkenngröße für Flächenkarten ist das mittlere Auftretensgewicht $\bar{l}_x$ einer Variante x in dem ihr zugeordneten Gebiet $T(x)$, also

$$\bar{l}_x = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} l_x(t_j) \,. \tag{3}$$

Diese Kennzahl kann beschrieben werden als der Bruchteil der insgesamt möglichen Gewichte in $T(x)$ (vgl. Abschnitt 4.1), der tatsächlich von x stammt. Der Extremfall $\bar{l}_x = 1$ tritt nur dann auf, wenn an allen Befragungsorten in $T(x)$ nur eine einzige Variante, nämlich x, auftritt. Andernfalls wird ein Teil der in $T(x)$ auftretenden Varianten durch die Zuordnung zu x nicht wiedergegeben. Treten beispielsweise an einem Ort t_i die Varianten x_1 und x_2 mit $\bar{l}_{x_1} = \frac{1}{2} = \bar{l}_{x_2}$ auf, und ist $x(t_i) = x_1$,

so wird das Auftretensgewicht $\bar{l}_{x_2} = \frac{1}{2}$ durch die Zuordnung von t_i zu $T(x_1)$ nicht berücksichtigt. Grob gesagt ist der Wert von $\bar{l}_x$ umso größer, je weniger von x verschiedene Varianten in $T(x)$ auftreten. Wir bezeichnen $\bar{l}_x$ daher als die „Gebietskompaktheit der Fläche der Variante x". Entsprechend definieren wir die „Gesamtkompaktheit einer Karte" über das gewichtete Mittel

$$\bar{L} = \sum_x \frac{|T(x)|}{n} \cdot \bar{l}_x = \frac{1}{n} \sum_x \sum_{t_j \in T(x)} l_x(t_j) \,. \tag{4}$$

Die Gewichte werden dabei als die relative Zahl von Orten in den Gebieten der einzelnen Varianten gewählt. Diese Wahl ergibt sich aus der Überlegung, dass diese Zahl, multipliziert mit n, das insgesamt mögliche Auftretensgewicht im jeweiligen Gebiet ergibt, denn die Gesamtgewichte der Varianten an einem gegebenen Ort sind immer 1. Die Gesamtkompaktheit einer Karte könnte man auch als die „Wiedergabetreue der Flächenkarte zu den Originaldaten" bezeichnen, da sie angibt, in welchem Umfang die ursprünglichen Varianten durch die Aufteilung in Flächen wiedergegeben werden. Die in Abb. 3 dargestellte Karte hat $\bar{l}_x$-Werte von 0,6 für das grüne Gebiet im Westen des Untersuchungsgebiets, bis hin zu 1,0 für das rote im Osten. Die Gesamtkompaktheit der Karte ist $\bar{L} = 0{,}72$, d.h. 72 % der Belege auf der Karte werden durch die Zuordnung der Orte zu den einzelnen Gebieten wiedergegeben. Auch für dieses Charakteristikum werden wir in Abschnitt 5 diverse Beispiele anführen, welche die Zahlen besser interpretierbar machen.

Schließlich sind mit

$$\bar{b}_x = \frac{1}{|T(x)|} \sum_{t_j \in T(x)} b(t_j) \tag{5}$$

und

$$\bar{B} = \sum_x \frac{|T(x)|}{n} \cdot \bar{b}_x \tag{6}$$

Kennzahlen für die „Homogenität" von Gebieten bzw. einer gesamten Flächenkarte verfügbar. Begründen lässt sich dies damit, dass hohe Werte von $b(t)$ anzeigen, dass die geschätzte Intensität von $x(t)$ deutlich höher ist als die anderer Varianten (vgl. Formel (2)). Mit anderen Worten: Große Werte für $\bar{b}_x$ bedeuten, dass im Gebiet der Variante x keine große Durchmischung mit anderen Varianten auftritt, dieses Gebiet also homogen ist. Im Unterschied zu $\bar{l}_x$ wird bei der Berechnung von $\bar{b}_x$ die geschätzte Auftretenswahrscheinlichkeit der Variante x benutzt; es geht hier also nicht um die tatsächlichen Belege. Das grüne Gebiet in Abb. 3 hat eine Homogenität von 0,44, während die Homogenität des violetten Gebietes 0,75 beträgt, und die der gesamten Flächenkarte 0,70. Wie für die beiden vorgenannten Charakteristiken werden auch hier die Beispiele aus Abschnitt 5 einen Kontext schaffen, der diese Zahlen interpretierbar macht.

Natürlich sind zahlreiche weitere Kennzahlen für strukturelle Charakteristiken von Flächenkarten

denkbar; einige davon werden wir in Abschnitt 6 erwähnen. Nicht nur aus Gründen der Übersichtlichkeit beschränken wir uns hier aber auf die drei genannten Kennzahlen, sondern auch, weil wir glauben, dass diese in ihrer Einfachheit und leichten Interpretierbarkeit anderen, komplizierteren Kennzahlen überlegen sind.

5. Einige Ergebnisse

Wir haben die in Abschnitt 4 vorgestellten Methoden zur Erzeugung und Charakterisierung von Flächenkarten auf ein großes Teilkorpus von Karten angewandt. Dieses Teilkorpus enthält mit 823 Karten einen Großteil der Wortkarten des SBS. Abb. 4 zeigt die Histogramme der mithilfe unserer Analysemethoden ermittelten Charakteristiken C (entspricht der Gesamtgrenzlänge), $\bar{L}$ und $\bar{B}$ dieser Karten. In Tab. 1 sind die Werte dieser Charakteristiken für die Beispielkarten in Abb. 3 und 5 bis 8 abgebildet, die weiter unten in diesem Abschnitt besprochen werden.

Wie Abb. 4(a) zeigt, beträgt die durchschnittliche Gesamtgrenzlänge zwischen den Gebieten auf einer Karte ca. 389 km, während es auch einige Karten mit weniger als 100 km oder auch überhaupt keinen Grenzlinien gibt. Diese erste Säule des Diagramms umfasst also auch diejenigen Karten, auf denen eine einzelne Variante so häufig vorkommt, dass nur ein großes Gebiet und somit keinerlei Grenzen entstehen. Die größte Gesamtgrenzlänge auf allen Karten ist noch unter 1.100 km, und Werte über 700 km sind schon relativ selten. Vor diesem Hintergrund ist der Wert C = 240,8 km der Karte in Abb. 3 recht niedrig. Ein weiteres extremes Beispiel liegt in Abb. 5 vor, wo die Varianten für ‘Rosenkranz’ verzeichnet sind. Hier ist C = 929,7 km. Die folgenden beiden Beispiele (Abb. 6 und 7) haben beide C-Werte, die über dem Durchschnitt liegen, aber nicht extrem sind. Abb. 8, die Karte für ‘dürres Reisig’, hat Grenzen mit einer Gesamtlänge von 391,4 km, was in etwa dem Durchschnitt entspricht. Diese Zahlen helfen uns also, die Komplexität einer Karte zu beurteilen. Ein Vergleich der Abb. 3 und 5 zeigt, dass diesbezüglich mitunter sehr große Unterschiede zwischen einzelnen Karten auftreten, ein Umstand, der mit Methoden der klassischen Dialektometrie nicht erfasst werden kann. Eine Erklärung für diese Unterschiede steht jedoch nach wie vor aus. Als nächsten Schritt könnte man nun beispielsweise Karten mit ähnlichen Werten gruppieren (z.B. mithilfe von Clusteranalyse), um herauszufinden, welche Bedingungen (wie z.B. Gebrauchsfrequenz oder Alter des Begriffs) vergleichbare Karten gemeinsam haben, um so ihre Ähnlichkeit zu erklären.[7]

Der Mittelwert von $\bar{L}$ über alle Karten ist ca. 0,62, was bedeutet, dass im Schnitt 62 Prozent der Belege in einem Gebiet zu der Variante gehören, der das Gebiet zugewiesen wurde. Das entsprechende Histogramm ist in Abb. 4(b) abgebildet. Die Beispielkarte in Abb. 3 hat einen Wert

7 Erste Ergebnisse von mithilfe von Clusteranalyse ermittelten Gruppierungen erscheinen in RUMPF/PICKL/ELSPAß/-KÖNIG/SCHMIDT (2010b).

von $\bar{L}$ = 0,72 und somit einen Wert, der etwas über dem Durchschnitt liegt. Die Karte in Abb. 6 hat den sehr kleinen $\bar{L}$-Wert 0,31, d.h. die Gebiete stimmen im Schnitt nur mit 31 Prozent der Belege überein, die an den zu ihnen gehörenden Orten auftreten. Die Karte in Abb. 8 weist wiederum einen Wert auf, der relativ nahe am Durchschnitt liegt. Der $\bar{L}$-Wert ist eine Maßzahl, die angibt, wie geeignet eine Karte ist, in Flächen eingeteilt zu werden, indem sie die Abweichungen von den vereinfachenden Flächen quantifiziert. Sie liefert uns somit Informationen über die strukturellen Eigenschaften einer Karte. Karten, die sehr kompakt sind, erlauben eine hohe Wiedergabetreue in den abgeleiteten Flächenkarten, was bedeutet, dass hierfür nur wenig von den Originaldaten abstrahiert werden muss. Daher dient die Kompaktheit $\bar{L}$ in erster Linie der Bewertung der Eignung einer Flächenkarte zur Visualisierung der zugrundeliegenden Punktsymbolkarte.

Ein mit $\bar{L}$ in engem Zusammenhang stehender, wenngleich aussagekräftigerer Wert ist $\bar{B}$, die mittlere Dominanz der jeweils durchsetzungsstärksten Variante. $\bar{B}$ ist ein Indikator für die sogenannte Homogenität einer Karte. Mehr Durchmischung der Varianten im kleinen Maßstab führt zu einem kleineren $\bar{B}$-Wert. Vor dem Hintergrund der Ausführungen in Abschnitt 2 kann man auch schließen, dass bei einem kleinen $\bar{B}$ der lokale Wettbewerb zwischen den Varianten eher groß ist, was auf eine weniger stabile linguistische Situation hindeuten würde. Für $\bar{B}$ wurde ein Mittelwert von 0,58 ermittelt. Wie das Histogramm aller $\bar{B}$-Werte (Abb. 4(c)) zeigt, treten keine Werte unter 0,2 auf. Dies ist auch zu erwarten, denn in Anbetracht von Formeln (2), (5) und (6) müsste, damit $\bar{B}$ gegen 0 ginge, die Anzahl der Varianten auf einer Karte gegen unendlich gehen. Der $\bar{B}$-Wert 0,71 für die Karte in Abb. 3 liegt signifikant über dem Durchschnitt, was sich in mehrheitlich dunklen Farbtönen auf der Flächenkarte mit nur wenigen helleren Stellen niederschlägt. Dies spricht für eine relativ stabile Verteilung der Varianten. Abb. 7 ist insgesamt noch dunkler als Abb. 3, mit nur sehr wenigen helleren Stellen innerhalb der Gebiete, was sich durch ein höheres $\bar{B}$ ausdrückt. Dass dies trotz einer größeren Gesamtgrenzlänge der Fall ist, zeigt, dass der Einfluss der Grenzen auf $\bar{B}$ nicht allzu groß ist. Abb. 6 hingegen ist insgesamt sehr hell, $\bar{B}$ dementsprechend nur 0,24. Wiederum liegt die Karte in Abb. 8 nahe am Durchschnitt.

Um die Aussagekraft dieser Werte einschätzen zu können, ist es natürlich wichtig zu wissen, inwieweit sie miteinander in Beziehung stehen. Das gebräuchlichste Hilfsmittel zur Untersuchung dieser Frage ist der Pearson'sche Korrelationskoeffizient ρ, dessen Werte zwischen 1 und –1 liegen. Extreme Werte von ρ in Bezug auf ein Paar von Charakteristiken deuten darauf hin, dass die eine der beiden betrachteten Kenngrößen zu einem großen Anteil von der jeweils anderen bestimmt werden. Umgekehrt bedeuten Werte nahe bei 0, dass ein solcher Zusammenhang nicht besteht. Mehr Details zum Korrelationskoeffizienten findet man z.B. in RODGERS/NICEWANDER (1988). Die Werte von ρ für die von uns betrachteten Charakteristiken sind in Tab. 2 gegeben. Die recht hohe positive Korrelation zwischen $\bar{L}$ und $\bar{B}$ zeigt, dass Karten, auf denen hohe Werte von $\bar{L}$ auftreten,

tendenziell auch hohe Werte für $\bar{B}$ haben – und umgekehrt. Mit anderen Worten: Homogenität und Kompaktheit tendieren dazu, Werte in ähnlichen Größenordnungen zu haben, was man beispielsweise auch an den Werten in Tab. 1 leicht erkennen kann. Diese Erkenntnis ist nicht überraschend, wenn man sich die Definitionen von $\bar{L}$ und $\bar{B}$ noch einmal vor Augen führt: Sind die Varianten auf einer Punktsymbolkarte recht homogen verteilt (was sich in einem großen $\bar{B}$ niederschlägt), so lässt sich diese leicht in eine Flächenkarte transformieren, was durch klar abgegrenzte Gebiete und große $\bar{L}$ zum Ausdruck kommt.

Die Gesamtgrenzlänge C ist sowohl mit $\bar{L}$ also auch mit $\bar{B}$ negativ korreliert, wenngleich betragsmäßig nicht so deutlich wie $\bar{L}$ mit $\bar{B}$. Trotzdem bedeutet dies, dass Karten mit höherer Komplexität dazu tendieren, eine geringe Homogenität und eine geringe Kompaktheit zu haben. Dieses Ergebnis erscheint sehr plausibel, wenn man bedenkt, dass eine größere Grenzlänge bedeutet, dass es mehr Orte gibt, die nahe an Grenzen liegen und an denen die Zuordnung zu einer Variante daher nicht sehr deutlich und somit $b(t)$ eher klein ist.

6. Schluss und Ausblick

In diesem Beitrag haben wir ein neues Verfahren zur computergestützten Analyse von geographischen Sprachdaten vorgestellt. Der erste Schritt bestand in der Erzeugung von Flächenkarten aus den Rohdaten. Dazu wurden mittels räumlich-statistischer Methoden für alle Varianten des betreffenden Merkmals im gesamten Untersuchungsgebiet Intensitäten, d.h. die erwarteten Auftretenshäufigkeiten, geschätzt. Die durch die Kombination der erhaltenen Intensitätsfelder entstehenden Flächenkarten sollen als Grundlage für die weitere Analyse dienen, u.a. indem Werte für jede Karte ermittelt werden, die ihre Struktureigenschaften beschreiben. Die drei vorgestellten Kennwerte, C, $\bar{L}$ und $\bar{B}$, ermöglichen es, den Aufbau einer Karte quantitativ zu erfassen und dabei die intuitiv verständlichen Konzepte *Komplexität*, *Kompaktheit* und *Homogenität* zu bedienen. Anhand verschiedener Beispiele haben wir gezeigt, dass sowohl die automatisierte Gebietseinteilung als auch die entsprechenden C-, $\bar{L}$- und $\bar{B}$-Werte die visuellen Eindrücke gut wiedergeben, die ein Dialektologe mehr oder weniger intuitiv – und damit subjektiv – beschreiben müsste. Diese neue Methodik ist der erste Schritt in Richtung eines Softwaresystems, das auf verschiedenste Arten von geographischen Sprachdaten anwendbar sein soll, und dem Dialektologen auch dann eine problemorientierte Analyse von Sprachkarten ermöglichen soll, wenn er die verwendeten Algorithmen nicht im Detail kennt.

Dabei sind die hier vorgestellten Verfahren nur ein erster Schritt. Damit sie zu brauchbaren Ergebnissen führen, müssen die ermittelten Werte in sinnvoller Weise ausgewertet werden. So sollen die Karten etwa rechnerisch klassifiziert werden, was beispielsweise durch Clusteranalyse

erfolgen kann.[8] Die entstehenden Gruppierungen können dann bei der Suche nach Faktoren der geographischen Variation hilfreich sein, indem systematisch untersucht wird, was strukturell ähnliche Karten auch linguistisch gemeinsam haben. Über die Betrachtung dieser die ganzen Karten betreffenden Struktureigenschaften hinaus kann auch die Suche nach geographisch definierten Strukturen interessant sein. Hier sei nur auf die in der dialektologischen Literatur viel diskutierten Trichter-, Keil- oder Kreisformationen verwiesen. Bei der Untersuchung solcher geometrischen Muster werden spezielle Methoden der räumlichen Statistik eine wichtige Rolle spielen. Das Ziel all dessen ist es, empirisch fundierte Aussagen über die den geographischen Strukturen zugrundeliegende Sprachvariation und den stattfindenden Sprachwandel treffen zu können.

Die erhaltenen Messwerte können in Verbindung mit der Einteilung der Karten in Gebiete auch die Grundlage für eine Forschung sein, die über die Untersuchung einzelner Merkmalskarten hinausgeht. So können sie etwa dazu dienen, Hypothesen über sprachliche Grenzen zu testen, indem überprüft wird, ob eine bestimmte vorgegebene Linie (z.B. Flüsse oder politische Grenzen) signifikant oft mit einer sprachlichen Grenze übereinstimmt. Für linguistisch definierte Merkmalsgruppen, wie z.B. Romanismen oder Standardismen, können kumulative Vorkommenskarten erzeugt werden, um die Suche nach horizontalen oder vertikalen Einflusssphären zu erleichtern.

Literatur

ALTMANN, GABRIEL (1983): Das Piotrovski-Gesetz und seine Verallgemeinerungen. In: BEST, KARL-HEINZ / JÖRG KOHLHASE (Hg.): Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte. Göttingen: Edition Herodot, 59–102.

BACH, ADOLF (1969): Deutsche Mundartforschung. Ihre Wege, Ergebnisse und Aufgaben. 3. Auflage. Heidelberg: Winter.

BADDELEY, ADRIAN / PABLO GREGORI / JORGE MATEU / RADU STOICA / DIETRICH STOYAN (Hg.) (2006): Case Studies in Spatial Point Process Modeling. New York: Springer (Lecture Notes in Statistics. 185).

BEST, KARL-HEINZ (2006): Quantitative Linguistik. Eine Annäherung. 3. Auflage. Göttingen: Peust & Gutschmidt.

DIGGLE, PETER J. (2003): Statistical Analysis of Spatial Point Patterns. 2. Auflage. London: Arnold.

ELSPAß, STEPHAN / WERNER KÖNIG (Hg.) (2008): Sprachgeographie digital. Die neue Generation der Sprachatlanten. Hildesheim/Zürich/New York: Olms (Germanistische Linguistik. 190–191).

8 Vgl. RUMPF/PICKL/ELSPAß/KÖNIG/SCHMIDT 2010b.

FRANCIS, W. NELSON (1983): Dialectology. An Introduction. London: Longman.

FRINGS, THEODOR (1956): Sprache und Geschichte II. Halle (Saale): Niemeyer (Mitteldeutsche Studien. 17).

GOEBL, HANS (1994): Dialektometrie und Dialektgeographie. Ergebnisse und Desiderata. In: MATTHEIER, KLAUS / PETER WIESINGER (Hg.): Dialektologie des Deutschen. Forschungsstand und Entwicklungstendenzen. Tübingen: Niemeyer, 171–191.

GOEBL, HANS (2001): Arealtypologie und Dialektologie. In: HASPELMATH, MARTIN / EKKEHARD KÖNIG / WULF OESTERREICHER / WOLFGANG RAIBLE (Hg.): Sprachtypologie und sprachliche Universalien. Ein internationales Handbuch. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 20.2), 1471–1491.

GOEBL, HANS (2006): Recent Advances in Salzburg Dialectometry. In: Literary & Linguistic Computing 21, 411–435.

GOEBL, HANS (2007): Kurzvorstellung der Korrelativen Dialektometrie. In: GRZYBEK, PETER (Hg.): Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday. Berlin: Mouton de Gruyter, 165–180.

HAAG, CARL (1898): Die Mundarten des oberen Neckar- und Donaulandes. Reutlingen: Hutzler.

HAAS, WALTER (1978): Sprachwandel und Sprachgeographie. Untersuchungen zur Struktur der Dialektverschiedenheit am Beispiele der schweizerdeutschen Vokalsysteme. Stuttgart: Steiner (Zeitschrift für Dialektologie und Linguistik. Beihefte. 30).

HÄNDLER, HARALD / HERBERT ERNST WIEGAND (1982): Das Konzept der Isoglosse: methodische und terminologische Probleme. In: BESCH, WERNER/ ULRICH KNOOP / WOLFGANG PUTSCHKE / HERBERT ERNST WIEGAND (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 1.1), 501–527.

HEERINGA, WILBERT (2004): Measuring dialect pronunciation differences using Levenshtein distance. Rijksuniversiteit Groningen (Groningen Dissertations in Linguistics. 46).

HILDEBRANDT, REINER (1983): Typologie der arealen lexikalischen Gliederung deutscher Dialekte aufgrund des Deutschen Wortatlasses. In: BESCH, WERNER / ULRICH KNOOP / WOLFGANG PUTSCHKE / HERBERT ERNST WIEGAND (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 1.2), 1331–1367.

HUMMEL, LUTZ (1993): Dialektometrische Analysen zum Kleinen Deutschen Sprachatlas (KDSA).

Experimentelle Untersuchungen zu taxometrischen Ordnungsstrukturen als dialektaler Gliederung des deutschen Sprachraums. Tübingen: Niemeyer (Studien zum Kleinen Deutschen Sprachatlas. 4).

ILLIAN, JANINE / ANTTI PENTTINEN / HELGA STOYAN / DIETRICH STOYAN (2008): Statistical Analysis and Modelling of Spatial Point Patterns. Chichester: Wiley.

KELLE, BERNHARD (1986): Die typologische Raumgliederung von Mundarten. Eine quantitative Analyse ausgewählter Daten des Südwestdeutschen Sprachatlasses. Marburg: Elwert (Studien zur Dialektologie in Südwestdeutschland. 2).

KÖHLER, REINHARD / GABRIEL ALTMANN / RAJMUND G. PIOTROWSKI (Hg.) (2005): Quantitative Linguistik. Ein internationales Handbuch. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 27).

KÖNIG, WERNER (1982): Probleme der Repräsentativität in der Dialektologie. In: BESCH, WERNER / ULRICH KNOOP / WOLFGANG PUTSCHKE / HERBERT ERNST WIEGAND (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 1.1), 463–485.

LEOPOLD, EDDA (2005): Das Piotrowski-Gesetz. In: KÖHLER, REINHARD / GABRIEL ALTMANN / RAJMUND G. PIOTROWSKI (Hg.): Quantitative Linguistik. Ein internationales Handbuch. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft. 27), 627–633.

NERBONNE, JOHN / WILBERT HEERINGA (1998): Computationele vergelijking en classificatie van dialecten. In: Taal en Tongval, Tijdschrift voor Dialectologie 20, 164–193.

NERBONNE, JOHN (2006): Identifying Linguistic Structure in Aggregate Comparison. In: Literary & Linguistic Computing 21, 463–475.

OKABE, ATSUYUKI / BARRY BOOTS / KOKICHI SUGIHARA / SUNG NOK CHIU (2000): Spatial tessellations: concepts and applications of Voronoi diagrams. 2. Auflage. Chichester: Wiley.

PIOTROWSKI, RAJMUND G. / KALDYBAY B. BEKTAEV / ANNA A. PIOTROWSKAJA (1985): Mathematische Linguistik. Bochum: Brockmeyer (Quantitative Linguistics. 27).

PRÖLL, SIMON (i.V.): Wahrnehmungspsychologische Aspekte der Sprachkartographie.

RODGERS, JOSEPH LEE / W. ALAN NICEWANDER (1988): Thirteen ways to look at the correlation coefficient. In: The American Statistician 42 (1), 59–66.

RUMPF, JONAS / SIMON PICKL / STEPHAN ELSPAß / WERNER KÖNIG / VOLKER SCHMIDT (2010a): Structural Analysis of Dialect Maps Using Methods from Spatial Statistics. In: Zeitschrift für Dialektologie und Linguistik 76 (3), 280–308.

RUMPF, JONAS / SIMON PICKL / STEPHAN ELSPAß / WERNER KÖNIG / VOLKER SCHMIDT (2010b): Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-class Maps. In: Dialectologia et Geolinguistica 18 (im Druck).

SBS = WERNER KÖNIG (Hg.) (1996–2009): Sprachatlas von Bayerisch-Schwaben. 14 Bände. Heidelberg: Winter (Bayerischer Sprachatlas. Regionalteil 1).

SCHILTZ, GUILLAUME (1996): Der dialektometrische Atlas von Südwest-Baden (DASB). Konzepte eines dialektometrischen Informationssystems. Marburg: Elwert (Studien zur Dialektologie in Südwestdeutschland. 5).

SCHNEIDER, EDGAR W. (1988): Qualitative vs. Quantitative Methods of Area Delimitation in Dialectology: A Comparison Based on Lexical Data from Georgia and Alabama. In: Journal of English Linguistics 21 (2), 175–212.

SCOTT, DAVID W. (1992): Multivariate Density Estimation: Theory, Practice, and Visualization. New York: Wiley.

SEGUY, JEAN (1965–1973): Atlas linguistique et ethnographique de la Gascogne. Paris: Centre National de la Recherche Scientifique.

SILVERMAN, BERNARD W. (1986): Density Estimation for Statistics and Data Analysis. New York: Chapman & Hall.

STOYAN, DIETRICH / HELGA STOYAN (1994): Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics. Chichester: J. Wiley & Sons.

WENZEL, WALTER (1930): Wortatlas des Kreises Wetzlar und der umliegenden Gebiete. Marburg: Elwert (Deutsche Dialektgeographie. 28).

Abbildungen

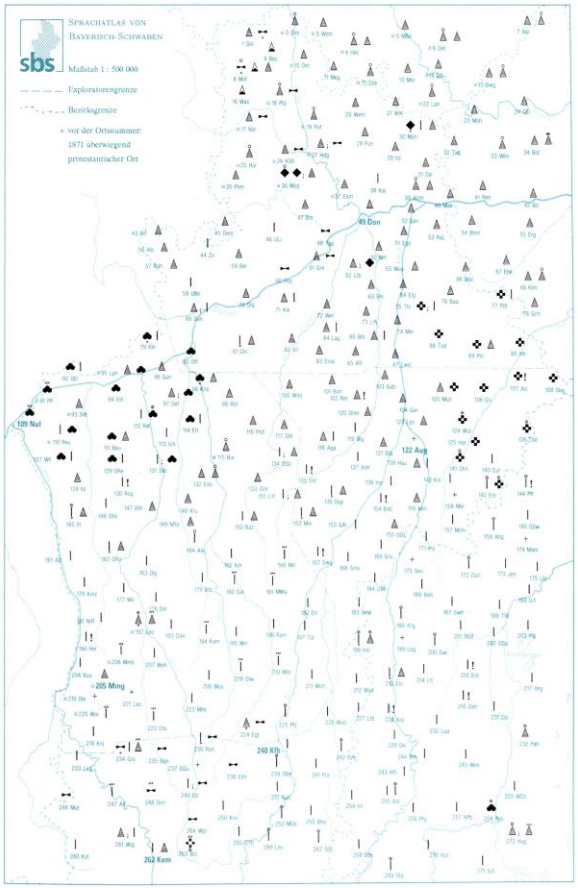


Abb. 1: Punktsymbolkarte 80 „Kartoffelkraut“ aus dem SBS (Band 8, 295).

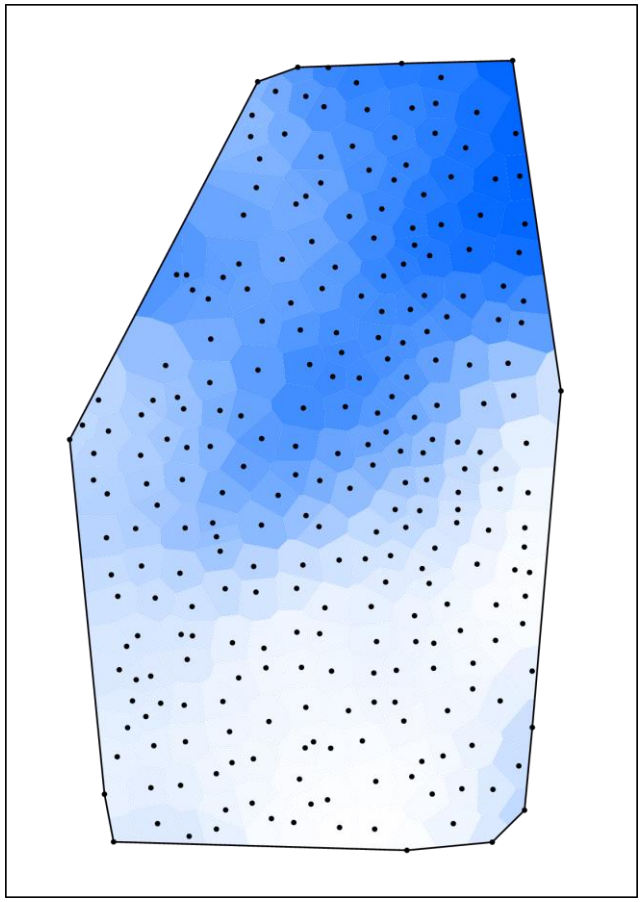


Abb. 2: Beispiel: geschätzte Intensitätskarte der Variante *Kraut* aus Karte 80 „Kartoffelkraut“ (SBS, Band 8, 294f.). In Abb. 1 wird diese Variante durch ein Dreieck markiert. Das zugehörige Gebiet in Abb. 3 ist türkis.

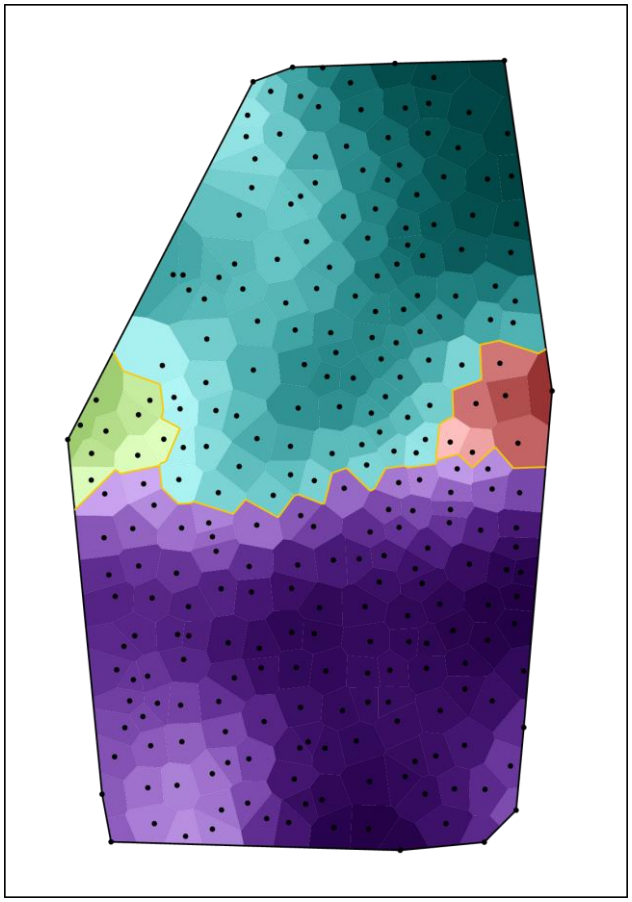
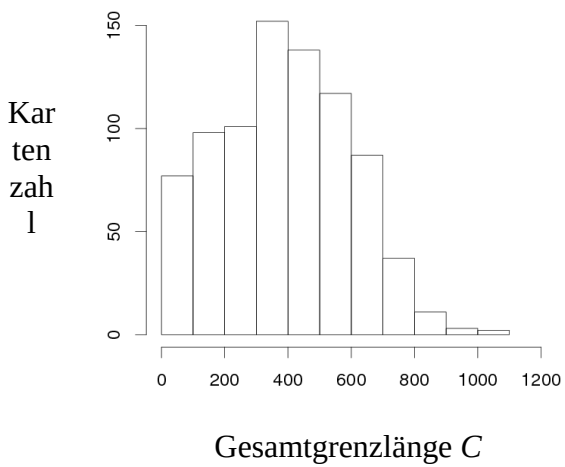
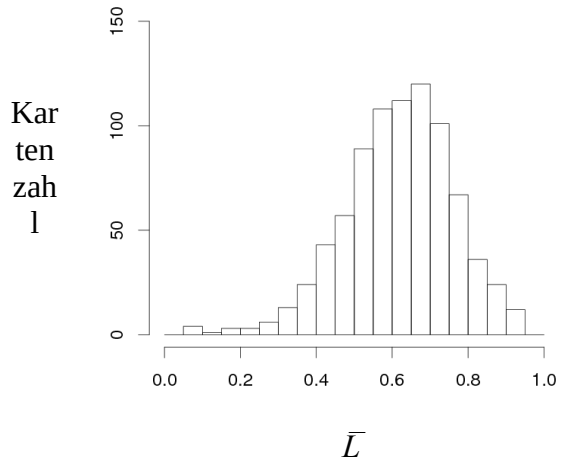


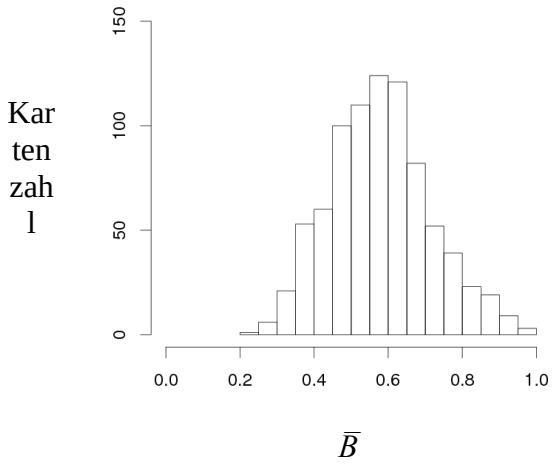
Abb. 3: Beispiel: Flächenkarte der Daten zu Karte 80 „Kartoffelkraut“ (SBS, Band 8, 294f.).



(a) Gesamte Grenzlänge C (*Komplexität*) zwischen den Gebieten auf einer Karte.



(b) Gesamte *Gebietskompaktheit* $\overline{L}$ von Karten.



(c) Gesamte *Homogenität* $\overline{B}$ von Karten.

Abb. 4: Histogramme verschiedener Kartencharakteristiken, berechnet aus 823 Wortgeographiekarten aus dem SBS.

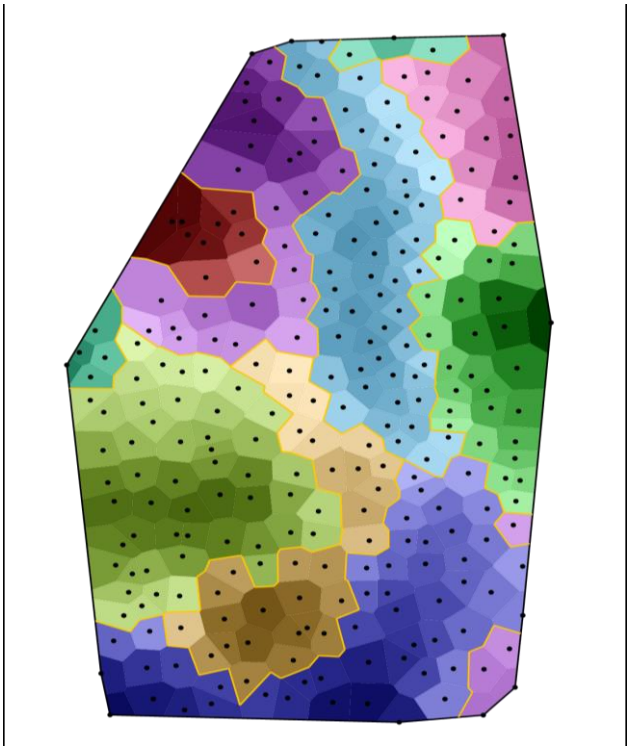


Abb. 5: Beispiel: Flächenkarte der Daten zu Karte 126 „Rosenkranz“ (SBS, Band 2, 532f.).

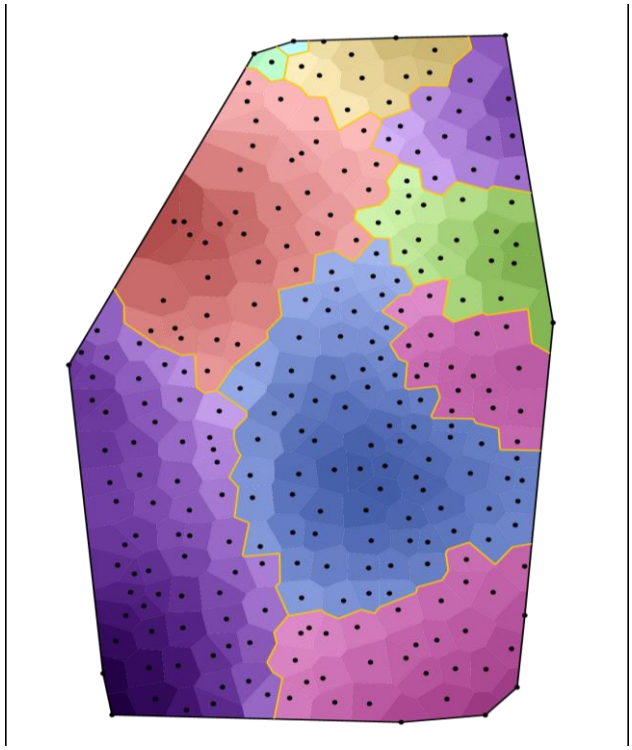


Abb. 6: Beispiel: Flächenkarte der Daten zu Karte 15 „die kleinen Hinterklauen der Kuh“ (SBS, Band 11, 52f.).

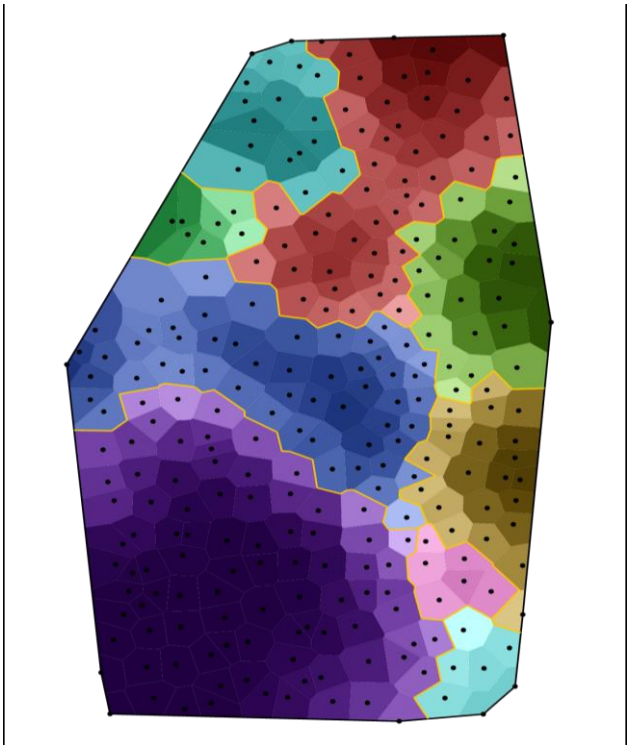


Abb. 7: Beispiel: Flächenkarte der Daten zu Karte 71 „Heuhaufen bei drohendem Regen“ (SBS, Band 12, 220f.).

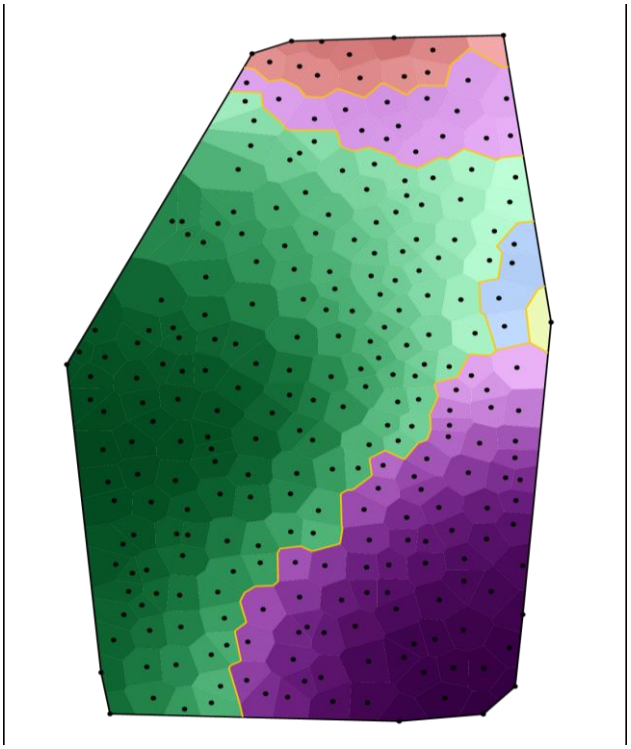


Abb. 8: Beispiel: Flächenkarte der Daten zu Karte 13 „dürres Reisig“ (SBS, Band 13, 532f.).

Tabellen

Tab. 1: Werte verschiedener Charakteristiken für die Beispielkarten.

Abbildung	**Titel**	***C***	$\bar{L}$	$\bar{B}$

Abb. 3	‘Kartoffelkraut’	240,8 km	0,72	0,71
Abb. 5	‘Rosenkranz’	929,7 km	0,53	0,40
Abb. 6	‘die kleinen Hinterklauen der Kuh’	637,2 km	0,31	0,24
Abb. 7	‘Heuhaufen bei drohendem Regen’	653,3 km	0,80	0,67
Abb. 8	‘dürres Reisig’	391,4 km	0,65	0,60
Durchschnittswerte		388,6 km	0,62	0,58

Tab. 2: Empirische Korrelationskoeffizienten zwischen verschiedenen Kartencharakteristiken, berechnet aus 823 Wortgeographiekarten aus dem SBS.

	C	$\bar{L}$	$\bar{B}$
C	1	-0,26	-0,58
$\bar{L}$		1	0,75
$\bar{B}$			1