Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede

John Nerbonne und Christine Siedle*
Alfa-informatica, Universität Groningen
Postbus 716, 9700 AS Groningen, Niederlande
{nerbonne,siedle}@let.rug.nl

28. Januar 2005

Zusammenfassung

Die Dialektgeographie betrachtet Sprachdaten normalerweise auf einem kategorialen (nominalen) Niveau, wobei die Einteilung in Dialekte auf der Grundlage einzelner sprachlicher Phänomene geschieht. Wenn aber die geographischen Verteilungen verschiedener sprachlicher Phänomene nun nicht übereinstimmen, dann erhält der Begriff 'Dialektgebiet' keine befriedigende analytische Basis—dies ist das Problem der nicht zusammenfallenden Isoglossen. Der vorliegende Beitrag schlägt eine Lösung dieses Problems vor durch die Verwendung eines Messinstruments für Ausspracheähnlichkeit, mit dem sich große Datenmengen systematisch verarbeiten lassen. Das Messinstrument ist numerisch (anstatt kategorial) und deswegen additiv, so dass man Unterschiede nicht nur auf dem Niveau des Wortes oder Phonems charakterisieren kann, sondern auch auf dem aggregierten Niveau der 'Varietät' – z.B. des Dialektes eines Dorfes. Auf dieser Basis lassen sich Dialektgebiete bestimmen, bei denen die Grenzen fließender sind als bei den Gebieten, die in den Dialekthandbüchern besprochen werden. Das Verfahren wird anhand von Daten 186 deutscher Varietäten illustriert.

1 Einleitung

Im Fokus des vorliegenden Berichts ist die Anwendung eines quantitativen Maßes f¨ur Aussprache¨ahnlichkeit auf eine relativ große Stichprobe phonetischer Transkriptionen von Dialekten aus dem Gebiet der heutigen Bundesrepublik. Die zentrale Frage unserer Arbeit lautet, ob zum Erkennen einer koh¨arenten Sprachgeographie tats¨achlich die subjektiv gelenkte Auswahl von Spracheigenschaften der traditionellen Dialektologie (siehe Bloomfield (1933, 340-341)) notwendig ist oder ob die von uns verwendete quantitative Methode eine echte Alternative sein kann. Mit Coseriu (1975, 50) suchen wir eine Alternative zum drohenden Atomismus in der Dialektforschung.

In den folgenden Abschnitten behandeln wir die anzuwendende Methode, die Datensammlung, die Resultate und schließlich die k¨unftigen M¨oglichkeiten einer auf Ausspracheabst¨anden basierenden Dialektologie.

*Dieser Bericht entstand in einem Projekt in Zusammenarbeit mit Prof.Dr. Hermann Niebaum. Peter Kleiweg hat die Programme implementiert, und Frits Steenhuizen und Elwin Koster haben die Konvertierung der geographischen Daten vorgenommen. Prof.Dr. Jan Goossens und Prof.Dr. J¨urgen-Erich Schmidt haben wertvolle Kommentare gegeben. Wir danken diesen Forschern und den Teilnehmern des 1. Kongresses der Internationalen Gesellschaft f¨ur Dialektologie des Deutschen in Marburg, 5.–8.3.2003 f¨ur ihre Verbesserungsvorschl¨age.

2 Methode

In diesem Abschnitt beschreiben wir den Hintergrund des vorliegenden Experiments, die Technik, die ihm zu Grunde liegt, und den experimentellen Entwurf mitsamt den zu erwartenden Resultaten.

2.1 Hintergrund

Der erste Bef¨urworter des Einsatzes quantitativer Techniken in der Dialektologie war Séguy (1971), wohingegen GOEBL (1982, 1984) viele neue Techniken untersucht und verfeinert hat. In de germanistischen Dialektologie waren Herrgen & Schmidt (1989) und Hummel (1993) fr¨uhe Anwendere dialektometrische Tehniken, die eine zunehmende Popularit¨at gerade in den letzten Jahren genie¨sen (siehe u.a. Lenz (2002), Lameli (2004), und M¨oller (2003)).

GOEBL betont, dass bez¨uglich bestimmter Fragen der Dialektgeographie, insbesondere der Definition von Dialektgebieten, das Problem der nicht-¨ubereinstimmenden Isoglossen nie eine befriedigende Antwort erhalten hat. Dies bedeutet wiederum, dass man Methoden entwickeln muss, um mit den durch Isoglossen indizierten widerspr¨uchlichen Tendenzen auszukommen. Sowohl SÉGUY als auch GOEBL behandeln Sprachdaten allerdings auf einem kategorialen Niveau, so dass Wahrnehmungen (Dialektdaten) letztendlich als gleich oder ungleich bewertet werden, nicht aber als *zu einem bestimmten Grad ähnlich bzw. unähnlich*. Nerbonne, Heeringa & Kleiweg (1999) zeigen, dass sich f¨ur das Messen von Ähnlichkeit bzw. Abstand von Dialekten anhand (transkribierter) Aussprachebeispiele besonders gut das *Levenshteinsche Sequenzabstandsmaß* eignet. Nerbonne & Heeringa (1998) wenden die Technik auf niederl¨andische Dialektatlasdaten an, und Bolognesi & Heeringa (2002) erweitern diese Technik durch den Gebrauch von statistisch zuverl¨assigeren Stichproben, die aus sardinischen Aussprachedaten gewonnen werden. Heeringa & Gooskens (2003) verarbeiten mit derselben Methode auch akustische Daten aus dem norwegischen Sprachraum; die Ergebnisse vergleichen sie anschließend mit perzeptionsbasierten Einsch¨atzungen der Dialektdistanzen durch die Dialektsprecher selber. F¨ur einen umfassenden Überblick ¨uber Technik und M¨oglichkeiten des Levenstheinschen Sequenzabstandsmaßes sei auf Heeringa (2004) verwiesen. Die vorliegende Arbeit ist die erste, die den Vergleich von deutschen Dialekten mittels dieser Technik zum Ziel hat.

2.2 Technik

Grundlage f¨ur das Berechnen von (Aussprache-)Distanzen zwischen Dialekten ist das Berechnen der Distanzen zwischen W¨ortern, welches wiederum auf dem Vergleich der Phone basiert. Bei der einfachsten und bei hinreichender Datenmenge dennoch zuverl¨assigen Methode findet hier ein simpler Identit¨atsabgleich statt; d.h. sind die Phone identisch, betr¨agt der Abstand 0, sind sie nicht-identisch, ist er gleich 1. Um jedoch der Tatsache Rechnung zu tragen, dass z.B. ein [y] einem [i] ¨ahnlicher ist als ein [p], experimentieren wir auch mit einigen weitestgehend artikulatorisch motivierten Merkmalssystemen, wie z.B. VIEREGGE & CUCCHIARINI oder ALMEIDA & BRAUN, die beide entwickelt wurden, um die Qualit¨at von phonetischen Transkriptionen zu beurteilen (Vieregge, Rietveld & Jansen 1984, Almeida & Braun 1986, Heeringa & Braun 2003), und die beide zu ¨ahnlichen Ergebnissen kommen. Wir illustrieren das Verfahren zum Bestimmen des Abstandes zwischen Segmenten anhand des Systems f¨ur Vokale von VIERREGGE in Abbildung 1. Wir haben bereits Experimente auf der Grundlage einiger Merkmalssysteme durchgef¨uhrt, bei der Validierung unserer Ergebnisse sind wir jedoch zu dem Schluss gelangt, dass die Gewichtung der Merkmale f¨ur unsere Belange noch sorgf¨altig modifiziert werden muss, so dass die hier pr¨asentierten Ergebnisse auf dem einfacheren, f¨urs erste aber zuverl¨assigeren Identit¨atsabgleich der Phone basieren. Neben der Anpassung der Merkmalssysteme w¨are es sicherlich interessant, zum einen noch das Merkmalssystem von HERRGEN & SCHMIDT zu implementieren, das eigens dazu aufgestellt wurde,

	[i]	[e]	[u]	d([i],[e])	d([i],[u])
Lage	2 (vorn)	2 (vorn)	6 (hinten)	0	4
H¨ohe	4 (hoch)	3 (mittelhoch)	4 (hoch)	1	0
L¨ange	3 (kurz)	3 (kurz)	3 (kurz)	0	0
Rundung	0 (nicht-rund)	0 (nicht-rund)	1 (rund)	0	1
				1	5

Abbildung 1: Der Segmentabstand wird auf der Basis von Merkmalsbeschreibungen der Segmente errechnet. Die Distanz d zwischen zwei Lauten entspricht der Summe der Abstände zwischen einzelnen Merkmalsausprägungen (z.B. vorn – hinten). Innerhalb des Systems von VIEREGGE ist d([i],[e]) = 1 (d.h der Abstand zwischen [i] und [e] beträgt insgesamt eins) und d([i],[u]) = 5.

Dialektalit¨at zu messen (Herrgen & Schmidt 1989), und zum anderen direkt mit akustisch ermittelten Lautabst¨anden zu arbeiten, womit Heeringa (2004) f¨ur das Niederl¨andische noch bessere Ergebnisse erzielt hat als mit den artikulatorischen Merkmalssystemen. Nicht unbedingt geeignet sind phonologisch inspirierte Systeme, bei denen man sich vor einer Übergewichtung von Merkmalen h¨uten sollte, deren Motivation vor allem auf die Knappheit der Regeldarstellung zur¨uckgeht. So ist z.B. das Merkmal [± gespannt], das CHOMSKY und HALLE gebrauchen, um extrem vordere und extrem hintere Vokale von mittleren Vokalen zu unterscheiden, *prima facie* keine gute Basis f¨ur ein Abstandsmaß (Chomsky & Halle 1968).

Wie auch immer man nun die Abst¨ande zwischen den Lautsegmenten ermittelt hat, sie bilden die Basis f¨ur den Levenshteinschen Sequenzvergleich. Dieser definiert den Abstand zwischen zwei Sequenzen (hier: W¨ortern) als die Summe der *Kosten* aller Operationen, die n¨otig sind, um die eine in die andere Sequenz zu transformieren. Muss hierzu beispielsweise ein [b] durch ein [p] ersetzt werden, entsprechen die Kosten dieser Operation dem vorher definierten Segmentabstand zwischen den beiden Lauten. Ist nur diese eine Operation n¨otig, um die W¨orter zu transformieren, entspricht der Segmentabstand auch dem Sequenzabstand. Sind mehrere Operationen n¨otig, werden die jeweiligen Kosten der Operationen addiert. Um eine willk¨urliche Sequenz in andere zu transformieren, braucht man neben dem Ersetzen von Lauten auch die Operationen 'Tilgen' und 'Hinzuf¨ugen'. Die Kosten dieser Operationen leitet man ebenfalls aus dem Segmentabstand ab, wozu man eine Definition des Segments 'Stille' ben¨otigt. Bei der Verwendung von Merkmalssystemen wird hierf¨ur die etwas modifizierte Merkmalsdefinition eines artikulatorisch 'neutralen' Vokals bzw. Konsonanten ([ə] bzw. [ʔ]) verwendet. Die Kosten entsprechen dann dem Segmentabstand zwischen der so definierten 'Stille' und dem zu tilgenden bzw. hinzuzuf¨ugenden Segment. Das Beispiel in Abbildung 2 soll die Idee veranschaulichen. Hierin wird die Realisierung des Wortes *Durst* im Dialekt von Aachen [tʊəʃ] in die korrespondierende Realisierung des Dialekts von Vielbrunn [tɔʃt] ¨uberf¨uhrt (der Anschaulichkeit halber wird hier auf die Verwendung von Diakritika verzichtet). Die Kosten der Einzeloperationen, die hier auf dem Merkmalssystem von VIERREGGE basieren, werden zum

tʊəʃ	ersetze ʊ mit ɔ	1.5
tɔəʃ	tilge ə	1
tɔʃ	f¨uge t hinzu	5
tɔʃt		
		7.5

Abbildung 2: Levenshteinabstand entspricht den Kosten für die effizienteste Reihe von Operationen, die eine Sequenz in die andere transformiert.

Sequenzabstand addiert.

Da die drei Arten von Operationen zur Transformation von Sequenzen beliebig oft und in beliebiger Reihenfolge angewendet werden k¨onnen, gibt es meist mehrere Transformationsreihen, die sich durchaus im 'Preis' unterscheiden k¨onnen. Als Levenshteinabstand werden jedoch nur die Gesamtkosten der 'kosteng¨unstigsten' Reihe von Operationen bezeichnet. Der so definierte Abstand entspricht allen mathematischen Definitionselementen des Abstands, d.h. Symmetrie, Nichtnegativit¨at, Dreiecksungleichheit sowie 0 genau dann wenn zwei Sequenzen gleich sind. Wichtig f¨ur unsere Zwecke ist, dass man die Resultate der Levenshteinmessung als echte Abst¨ande analysieren kann. Kruskal (1999) bietet eine zug¨angliche Einf¨uhrung in die Technik des Sequenzvergleichs.

Es sei noch kurz bemerkt, dass es inzwischen effiziente Algorithmen gibt, um den Levenshteinabstand sehr schnell zu berechnen (siehe u.a. `http://www.let.rug.nl/~kleiweg/lev` f¨ur ein public-domain-Paket f¨ur Levenshteinmessungen).

Obwohl ein ausf¨uhrlicher Vergleich zwischen dieser Methode und Alternativen den Rahmen dieses kurzen Beitrags sprengen w¨urde, k¨onnten einige Bemerkungen wohl n¨utzlich sein. Die meisten dialektometrischen Ans¨atze behandeln die Daten im Grunde genommen auf einem nominalen Niveau, d.h. als identisch mit anderen Daten oder nicht-identisch (SEGUY 1971; GOEBL 1982, 1984; MÖLLER 2003). Das innovative Element unserer Methode liegt darin, dass Ausspracheunterschiede als metrische Daten betrachtet werden, die erstens um einiges sensibeler sind, un die sich zweitens deswegen den numerischen Analysetechniken erschließen lassen. Denselben Vorteil d¨urften HERRGEN & SCHMIDT (1989) und andere, die sich ebenfalls von ihrer Methode bedienen, f¨ur sich behaupten (HUMMEL 1993; LENZ 2002; LAMELI 2004), allerdings mit der Einschr¨ankung, dass ihr Ansatz eines manuellen Eingriffs bedarf, um vergleichbare Laute zu betimmen, wogegen die hier beschriebene Methode vollautomatisch auf Aussprachelisten von vergleichbaren W¨ortern angewendet wird. Wir erwarten nicht, dass die Ergebnisse der zwei Methoden radikal anders sein werden. Allerdings m¨ochten wir betonen, dass unsere Methode auf Material aus Dialektatlanten angewendet werden kann, die aus anderen Gr¨unden entworfen wurden, ohne einen zus¨atzlichen Aufwand f¨ur die Selektion der Daten. Es scheint uns daher gerechtfertigt, von einer objektiven – oder, vorsichtiger ausgedr¨uckt, objektivierbaren – Methode zu sprechen.

2.3 Das aktuelle Experiment

Ziel des aktuellen Experiments ist es zu untersuchen, ob ein systematischer Vergleich vieler Wortrealisierungen verschiedener Dialekte im Aggregat zu einer zufriedenstellenden alternativen Basis f¨ur die Einteilung der deutschen Dialekte f¨uhren kann, die geeignet ist, die Probleme der subjektiv ausgew¨ahlten und letztendlich nie perfekt ¨ubereinstimmenden Isoglossen zu ¨uberwinden. Um eine Antwort liefern zu k¨onnen, werden wir geeignetes Material von unterschiedlichen Erhebungsorten mit dem oben beschriebenen Sequenzvergleichsalgorithmus analysieren. Unsere Hypothese ist, dass die Prozedur zeigen wird, dass Orte mit ¨ahnlicher Aussprache einander auch geographisch nah sind. Dies muss sich keineswegs zwangsl¨aufig ergeben. Wenn die Technik nicht ausreichend valide oder sensibel ist, oder wenn die Dialektgeographie des Deutschen letztlich ohne eine subjektiv gelenkte Auswahl von Unterscheidungskriterien nicht auskommen kann, werden sich die als aussprachenah ermittelten Erhebungsorte willk¨urlich ¨uber den Erhebungsraum verteilen. Wir werden also untersuchen, ob die Ergebnisse erstens geographische Koh¨arenz zeigen, und ob sie zweitens dem wissenschaftlichen Konsensus im Großen und Ganzen entsprechen.

Abend	acht	Affe	alle	als	alte	alter
andern	anf¨angt	Apfelb¨aume	Äpfelchen	Äpfeln	auch	auf
Augenblick	austrinken	bald	bauen	Bauern	bei	beißen
Berge	besser	bestellt	Bl¨atter	bleib	b¨ose	b¨osen
braune	Brot	Bruder	das	deiner	dich	die
Dienstag	Donnerstag	Dorf	drei	dreschen	du	durchgelaufen
d¨urft	Durst	Ei	Eier	eingebrochen	eingeschlafen	eins
Eis	Eise	elf	erz¨ahlt	f¨angt	Feld	Felde
fest	Feuer	Fleisch	fliegen	Frau	Freitag	f¨unf
f¨ur	G¨anse	Garten	geblieben	gebracht	gebrannt	gefahren
gefallen	gefunden	gekannt	gekommen	genug	geschlafen	geschmolzen
gestern	gestohlen	gestorben	glaube	gleich	groß	gr¨oßer
gut	hat	Haus	Hause	H¨auser	heiß	heiße
Herz	Herzen	hoch	h¨oher	h¨ort	ich	isst
ist	ja	kalte	kein	Kind	Kochl¨offel	Kohlen
K¨uhe	lauter	Leute	Leuten	liebes	liegen	Luft
machen	Mann	mein	meinem	Milch	mit	mitgehen
Mittwoch	Montag	m¨ude	muss	m¨usst	Mutter	n¨ahen
neue	neun	oben	Ochsen	Ofen	ohne	Ohren
Pfeffer	Pfund	recht	rot	roten	Salz	Samstag
schlechte	schlechten	Schnee	schneien	sch¨one	schwarz	Schwester
sechs	Seife	sein	selbst	sich	sie	sieben
so	sollen	Sonntag	St¨uckchen	Tisch	Tochter	tot
treiben	trinken	tu	tut	¨uber	um	und
uns	unsere	unserem	unten	verkaufen	versteht	viel
vier	von	vor	wachsen	w¨are	was	Wasser
weh	Wein	weiße	wem	werden	Wetter	wieviel
will	Winter	wo	Wochen	wollte	Wort	Wurst
zehn	Zeiten	zum	zwei	zw¨olf		

Tabelle 1: Die 201 vorstehenden Wörter aus den Wenkersätzen wurden für den Kleinen Deutscher Lautatlas – Phonetik (siehe Text) erhoben und bilden die Grundlage für die Messungen in dieser Arbeit.

3 Daten

Um den Ausspracheabstand innerhalb der deutschen Dialekte zu bestimmen, wurden die Wortabst¨ande zwischen phonetischen Transkriptionen von 201 W¨ortern aus 186 Erhebungsorten auf dem Gebiet der heutigen Bundesrepublik gemessen. Die Summe dieser 201 Wortabst¨ande[1] gilt als das Maß des Dialektabstands zwischen den jeweiligen Erhebungsorten.

Die Daten wurden im Rahmen des am Forschungsinstitut f¨ur deutsche Sprache "Deutscher Sprachatlas" in Marburg initiierten Projekts "Kleiner Deutscher Lautatlas – Phonetik" (auch bekannt als "Phonetischer Atlas der Bundesrepublik Deutschland") erhoben (G¨oschel 1992). Weil diese qualitativ hochwertigen Daten bis jetzt allerdings relativ unbekannt geblieben sind, ist es sinnvoll, sie hier n¨aher zu beschreiben.

Die Transkriptionen wurden auf der Basis von Aufnahmen gemacht, die unter der Leitung von JOACHIM GÖSCHEL zuerst in den 1960er und 1970er Jahren in der alten BRD erhoben worden waren (G¨oschel 1992, S. 64-70). Nach der Wende konnte man vergleichbare Erhebungen auch auf dem Gebiet der ehemaligen DDR durchf¨uhren, so dass das Netz von Erhebungsorten innerhalb der heutigen Staatsgrenzen relativ vollst¨andig ist. Die elizitierten W¨orter (siehe Tabelle 1) stammen aus den f¨ur die Aufnahmen verwendeten 'Wenker-S¨atzen', d.h. den S¨atzen, die GEORG WENKER 1879–87 durch die Lehrer von ca. 40.000 Schulorten des damaligen Deutschen Reichs in den jeweiligen Ortsdialekt ¨ubertragen ließ. Der R¨uckgriff auf die Wenker-S¨atze war durch den Wunsch motiviert, die historische Entwicklung

[1] Tats¨achlich sind es jeweils etwas weniger, da i.d.R. nicht f¨ur jeden Dialekt alle Wortrealisierungen vorliegen. Zur Berechnung des Abstandes zwischen zwei Dialekten werden ausschließlich die W¨orter verwendet, f¨ur die in beiden Dialekten Realisierungen vorhanden sind.

¨uber ein Jahrhundert hinweg n¨aher zu studieren. Die Erhebungsorte sind in Abbildung 3 eingetragen; die Motivation f¨ur ihre Auswahl scheint jedoch undokumentiert geblieben zu sein.

Die Daten wurden von einer Gruppe in Marburg in IPA transkribiert (IPA 1949), wobei alle Transkriptionen dem Konsens der Mitarbeiter entsprechen mussten. Die Transkribenten waren ANTONIO ALMEIDA, ANGELIKA BRAUN, RAPHAELA LAUF und KLAUS MONTERMANN. Im Jahr 2002 haben ROGIER NIEUWEBOER, SAAKJE VAN DELLEN und BERTHIEN MARKVOORT im Rahmen einer Kooperation zwischen den Universit¨aten Groningen und Marburg[2] die Transkriptionen digitalisiert, wobei eine etwas modifizierte Version von X-SAMPA als maschinenlesbare Kodierung verwendet wurde (f¨ur Informationen ¨uber X-SAMPA, siehe `http://coral.lili.uni-bielefeld.de/LangDoc/EGA/Formats/Sampa/sampa.html`).

Das Material wird vom Deutschen Sprachatlas in Marburg (`http://www.uni-marburg.de/dsa/`) verwaltet und ist bislang nicht ¨offentlich zug¨anglich.

4 Resultate

Das unmittelbare Resultat der Abstandsbestimmung ist eine Abstandstabelle, in der die ermittelten Ausspracheabst¨ande in den Zellen stehen (anstatt von z.B. Kilometerangaben in Tabellen f¨ur St¨adteentfernungen). Weil diese Tabelle aber viel zu groß ($\binom{186}{2} = 17.205$ Abst¨ande) ist, um hier abgedruckt zu werden, bedienen wir uns einer Visualisierung, um die Resultate global zu verstehen. Abbildung 4 zeigt im Prinzip *alle* phonetischen Abst¨ande dadurch, dass kleine phonetische Abst¨ande zwischen Orten dunklen und große Abst¨ande hellen Linien entsprechen. Hierbei fallen vier gr¨oßere Gebiete ins Auge. Dies w¨aren zum einen ein jeweils relativ homogenes s¨udliches, ¨ostliches und n¨ordliches Dialektgebiet und zum anderen ein offensichtlich extrem heterogenes westliches Gebiet.

Die 'Linienkarte' liefert also offenbar eine Antwort auf die Frage nach der geographischen Koh¨arenz. Wenn die geographisch entfernteren Orte nach dem Levenshteinschen Maß nicht auch phonetisch un¨ahnlicher w¨aren, w¨urden wir u.a. viele dunklere Linien zwischen relativ weit auseinanderliegenden Orten sehen. Wie wir aber Abbildung 4 entnehmen k¨onnen, kommt signifikantere phonetische Ähnlichkeit praktisch nur f¨ur Paare von Orten vor, die auch geographische Nachbarn sind.

Der zweite Aspekt der Untersuchungsfrage gilt dem Vergleich der aktuellen Ergebnisse mit der traditionellen Dialekteinteilung durch die heutige Dialektwissenschaft. Weil diese vor allem mit Begriffen wie 'Dialektgebiet' arbeitet, ist es sinnvoll, die Gebiete, die implizit in der Tabelle der phonetischen Abst¨ande vorhanden sind und die sich bereits in der Linienkarte andeuten, klarer zum Vorschein zu bringen. Eine hierf¨ur intuitiv einleuchtende Methode ist das sogenannte *Clustering* (genauer: hierarchisches agglomeratives Clustering) (Jain & Dubes 1988). Beim Clustering identifiziert man innerhalb einer Abstandstabelle stets die beiden Elemente mit dem geringsten Abstand, die man anschließend zusammenf¨ugt und damit eine kleinere Tabelle erstellt. Wir illustrieren das Verfahren in Tabelle 2.

Dem Clustering wird zwar vorgeworfen, dass es mathematisch nicht 'stabil' sei, d.h., dass kleine Unterschiede in den Ausgangsdaten mitunter zu großen Unterschieden in den Ergebnissen f¨uhren k¨onnen. Daf¨ur ist diese Methode anschaulich und f¨uhrt oft zu sehr respektablen Ergebnissen. Es gibt, in Abh¨angigkeit von der Bestimmung der restlichen Abst¨ande zum neu zusammengef¨ugten Element (Aldenderfer & Blashfield 1984, Heeringa 2004), sehr viele Varianten des Clusterings, die ebenfalls zu deutlich unterschiedlichen Ergebnissen kommen k¨onnen. Wir haben aus diesem Grund mehrere Cluster-Verfahren getestet und sind dabei zu dem Schluss gelangt, dass mit dem Clustering durch gewogenes arithmetisches Mittel (*Weighted arithmetic average clustering*) die am einfachsten zu interpretierenden

[2] Die Kooperation wurde von HERMANN NIEBAUM initiiert; ANGELIKA BRAUN, JÜRGEN-ERICH SCHMIDT, JOHN NERBONNE und WILBERT HEERINGA sind bislang die weiteren Teilnehmer.

Abbildung 3: Verteilung der 186 Aufnahmeorte

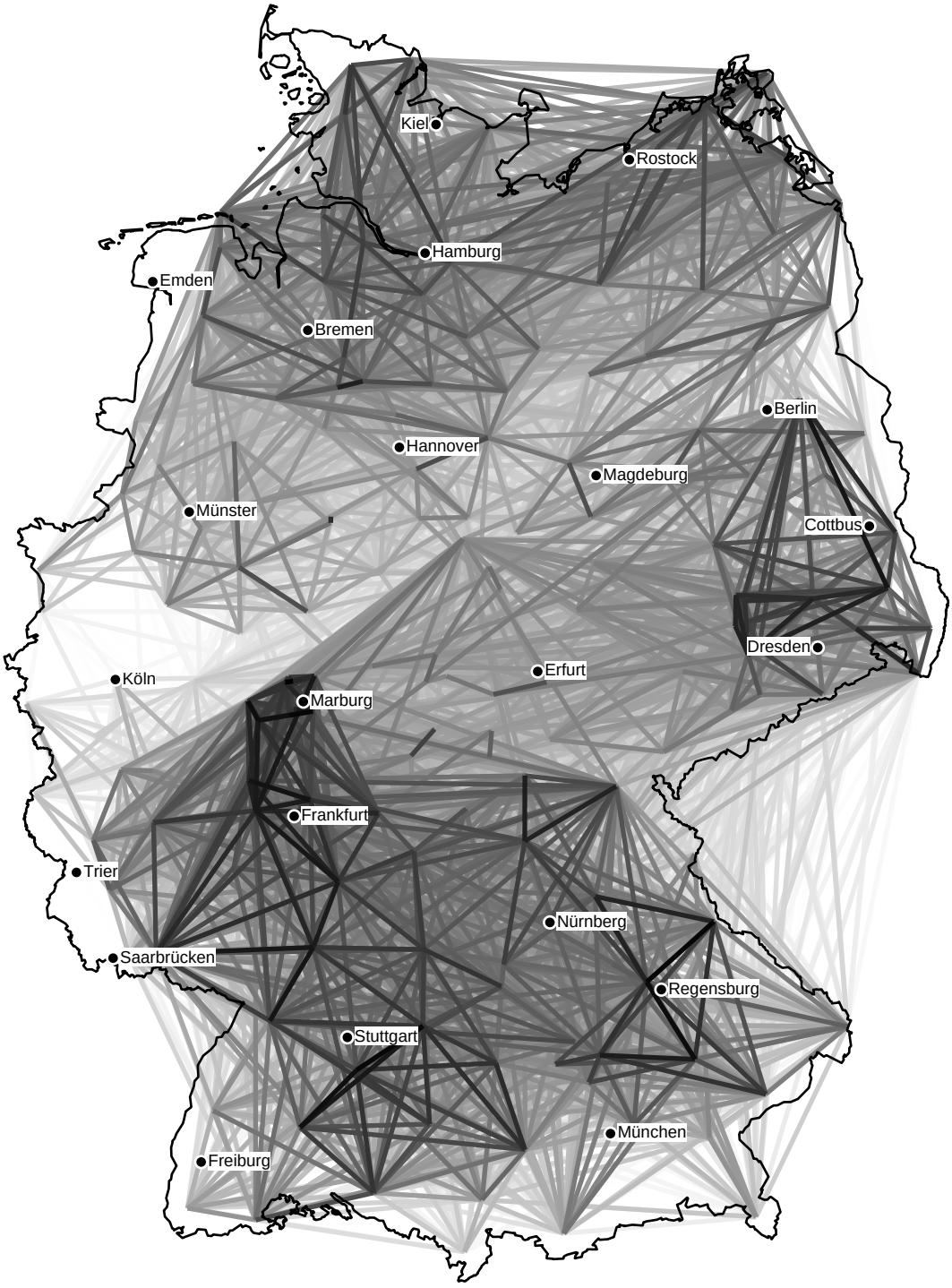

Abbildung 4: Ausspracheabstände zwischen *allen* Orten, wobei gilt: Je ähnlicher die beiden Dialekte, desto dunkler ihre Verbindungslinie. Zur Orientierung sind einige größere Städte eingezeichnet, von denen jedoch (mit Ausnahme von Köln und Nürnberg) keine Aufnahmen existieren.

	Jever	Oberau	K¨oln	Aachen	Hagen
Jever	0	73	64	67	79
Oberau		0	81	74	68
K¨oln			0	43	91
Aachen				0	86
Hagen					0

	Jever	Oberau	K¨oln-Aachen	Hagen
Jever	0	73	65.5	79
Oberau		0	77.5	68
K¨oln-Aachen			0	88.5
Hagen				0

Tabelle 2: Clustering identifiziert Gruppen in einer Abstandstabelle durch wiederholtes Zusammenfügen der jeweils nächsten Elemente. In der ersten Abstandstabelle weisen Köln und Aachen die geringsten Ausspracheunterschiede auf, weswegen sie in der unteren Tabelle zusammengefügt werden. Die Abstände der anderen Dialekte zu dem neuen Element Köln-Aachen entsprechen in diesem Beispiel einfach dem Durchschnitt der Abstände zu den Komponenten Köln und Aachen.

Resultate zu erzielen sind; diese Methode wird daher im Folgenden verwendet. Beim Clustering durch gewogenes arithmetisches Mittel wird die Anzahl der Elemente in den Clustern nicht ber¨ucksichtigt, weswegen die einzelnen Elemente in einem Cluster mit vielen Elementen weniger stark gewichtet werden als in einem Cluster mit wenigen Elementen. Dies mag auf den ersten Blick als wenig sinnvoll erscheinen, ist f¨ur unsere Zwecke jedoch von Bedeutung, da die Verteilung der Aufnahmeorte nicht v¨ollig gleichm¨aßig ist, so dass sich deren Dichte in den sp¨ateren Clustern erheblich unterscheiden kann. W¨urde die Anzahl der Elemente in die Cluster-Analyse einfließen, w¨urden Gebiete mit einer hohen Aufnahmeortdichte (z.B. der Raum Marburg) anders verarbeitet als Gebiete mit einer niedrigen Aufnahmeortdichte (z.B. der Raum Freiburg oder insbesondere das Gebiet n¨ordlich von K¨oln), was zu einer unrealistischen Klassifikation f¨uhren k¨onnte.

Wenn wir die Resultate des Clusterings auf einer Karte abbilden (Abb. 5), so erhalten wir n¨ahere Informationen zu unserer Untersuchungsfrage, ob eine systematische Messung von aggregierten Ausspracheunterschieden im Stande ist, die Dialektunterschiede im Deutschen zu erfassen. Wie sich bereits in der Linienkarte andeutete, finden wir eine geographisch koh¨arente Dreiteilung, die beim Clustering deutlich zu Tage tritt, sowie ein Gebiet im Westen Deutschlands, das offenbar durch eine große Heterogenit¨at gekennzeichnet ist. Auf die Gefahr hin, uns zu wiederholen, sei hier bemerkt, dass das Clustering in dieser Anwendung keineswegs geographisch inspiriert ist. Wenn die Levenshteinsche Methode nicht geeignet w¨are oder der Dialektsystematik des Deutschen keine geographische Struktur zu Grunde l¨age, w¨aren die Schattierungen in Abbildung 5 willk¨urlich im Raum verteilt.

Ein zweiter Aspekt der Untersuchungsfrage war, das Verh¨altnis zwischen den Ergebnissen der Levenshteinschen Methode und der Sicht der aktuellen Dialektwissenschaft zu vergleichen. Abbildung 5 zeigt eine Verteilung der großen Dialektgebiete, die nur wenig von der der Standardwerke abweicht (vgl. z.B. K¨onig (1994) oder Niebaum & Macha (1999)). Auch die ’Benrather Linie’ ist anhand typischer Merkmale (z.B. der ’Berliner Trichter’ oder die Harzer Dialektenklave) eindeutig zu erkennen, und auch die große Heterogenit¨at der Dialekte im Westen Deutschlands findet sich in der traditionellen Dialektologie als ’Rheinischer F¨acher’ wieder. Die Hierarchie der Dialektgebiete, die im Dendrogramm (Abb. 6)

Abbildung 5: Ergebnis des Clusterings als Karte dargestellt. Erkennbar sind drei Hauptgebiete, die sich im Wesentlichen mit den Verteilungen des Nieder-, Ostmittel- und Oberdeutschen (Cluster 1, 4 und 5) nach traditioneller Einteilung decken, sowie ein heterogenes Gebiet im Westen, das in etwa Ripuarisch (Cluster 3) und Niederrheinisch-Westmünsterländisch (Cluster 2) entspricht.

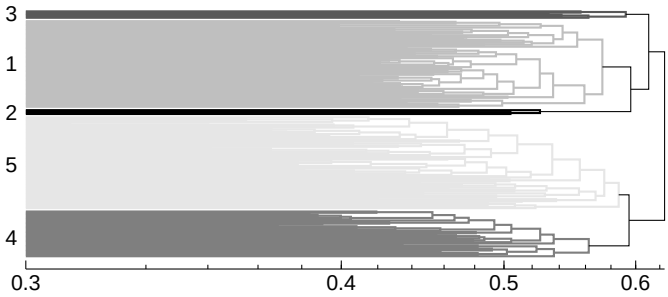

Abbildung 6: Resultat des Clusterings als Dendrogramm dargestellt. Die Grautöne und Clusternummern stimmen mit denen in Abbildung 5 überein. Die Grenze zwischen den nieder- und hochdeutschen Dialekten ist dominant. Das ripuarische Cluster wird mit dem hier verwendeten Clustering durch gewichtetes arithmetisches Mittel dem Niederdeutschen zugeordnet (Diskussion hierzu im Text).

dargestellt ist, deckt sich ebenfalls mit der herk¨ommlichen Klassifikation. Die ostmittel- und oberdeutschen Dialekte (Cluster 4 und 5) werden auch durch das Clustering zu 'Hochdeutsch' zusammengefasst; die Grenze zu den niederdeutschen Dialekten (Cluster 1-3) ist ausgepr¨agt. Der wesentliche Unterschied zur traditionellen Einteilung besteht darin, dass auf der Basis unserer bisherigen Resultate ein koh¨arentes westmitteldeutsches Dialektgebiet nicht auszumachen ist. Einerseits werden die traditionell als 'Westmitteldeutsch' klassifizierten Dialekte wie Hessisch, Mosel- und Rheinfr¨ankisch zwar dem Hochdeutschen zugeordnet, sie werden jedoch nicht, oder zumindest nicht im gleichen Ausmaß vom Oberdeutschen abgegrenzt wie die ostmitteldeutschen Dialekte. Andererseits wird Ripuarisch (Cluster 3) als Niederdeutsch klassifiziert (s. Dendrogramm), wobei allerdings anzumerken ist, dass dieser Zusammenschluss auf einer relativ hohen Ebene in der Hierarchie stattfindet, was darauf hindeutet, dass dieses Gebiet eine Sonderstellung zwischen den hoch- und niederdeutschen Dialekten einnimmt. Verwendet man ein anderes geeignetes Cluster-Verfahren, wie z.B. Clustering durch ungewichtetes arithmetisches Mittel, wird das ripuarische Cluster dem Hochdeutschen zugeordnet. Es ist jedoch nicht unwahrscheinlich, dass die Verwendung einer anderen Methode der Lautabstandsbestimmung, die die Konsonanten st¨arker gewichtet als dies der einfache Phon-Identit¨atsabgleich tut, was sich nat¨urlich auf die Dialektabst¨ande auswirkt, bei einem Gebiet mit einer so kritischen Stellung in der Hierarchie durchaus zu einer anderen Klassifikation f¨uhren kann. Generell sollte die Verwendung einer anderen Methode zur Bestimmung der Lautabst¨ande allerdings eher zu quantitativen als qualitativen Unterschieden f¨uhren; d.h. es kann zu kleinen Ver¨anderungen im Grenzverlauf kommen, was in einigen Gebieten auch noch notwendig erscheint. Zu den Grenzen allgemein sei noch angemerkt, dass ihr Verlauf nat¨urlich auch durch die Anzahl an Messpunkten mitbestimmt wird. Um die Punkte herum werden Polygone definiert, also kleine Areale, f¨ur die derselbe Dialekt angenommen wird, der in dem jeweiligen Referenzort gesprochen wird. Bei abnehmender Messpunkt-Dichte kommt es demnach zu einer zunehmend ungenauen Abbildung der realen Verteilung der Dialekte. Unser Messpunkt-Raster ist relativ grob, d.h. Feinheiten k¨onnen im Allgemeinen nicht wiedergegeben werden. F¨ur unsere Zwecke entscheidend ist jedoch, dass es wie demonstriert fein genug ist, um die wesentlichen Charakteristika der Verteilung wiederzugegeben.

Die klaren Grenzen in den Cluster-Karten dienen vor allem der Verdeutlichung, entsprechen jedoch nicht unbedingt der Realit¨at. Wie bereits erw¨ahnt, verlaufen die Grenzen zwischen den Dialekten tats¨achlich oft deutlich fließender, so dass man insgesamt eher von einem Dialektkontinuum sprechen sollte. Mit Hilfe des Levenshteinschen Sequenzabstandsmaßes kann man nun auch dieses Kontinuum darstellen. Wir bedienen uns hierf¨ur der sogenannten *multidimensionalen Skalierung* (Tourgerson 1952, Heeringa 2004). Zur Illustration dieser Methode stelle man sich z.B. einen Quader vor, in dem jede Ecke durch ihren Abstand zu allen anderen Ecken definiert ist. St¨unden einem nun nur die Abst¨ande zur Verf¨ugung, k¨onnte man den Quader allein aus diesen Angaben rekonstruieren. Der Unterschied bei der Anwendung auf Dialekte liegt haupts¨achlich darin, dass diese nicht von vorneherein dreidimensional sind. Nichtsdestotrotz werden sie bei einer multidimensionalen Skalierung so in drei Dimensionen[3] angeordnet, dass ihre jeweiligen Abst¨ande zu allen anderen Dialekten m¨oglichst genau wiedergegeben werden. Es findet also eine, nat¨urlich nicht verlustfreie, Datenkompression statt. Ordnet man nun jeder Dimension eine Farbe zu (hier rot, gr¨un und blau), erh¨alt jeder Dialekt auf Grund seiner Lage in den drei Dimensionen eine f¨ur ihn typische Farbe.Ähnliche Dialekte, die im dreidimensionalen Raum nah beieinander liegen, erhalten demnach auch ¨ahnliche Farben. Diese lassen sich wiederum in einer (zweidimensionalen) Karte darstellen. Abbildung 4 zeigt solch eine Regenbogenkarte als Ergebnis einer multidimensionalen Skalierung. Nun liegt es in der Natur der Methode, dass Grenzen kaum auszumachen sind, dennoch lassen sich die wichtigsten Unterscheidungen (Oberdeutsch gr¨un, Ostmitteldeutsch

[3]Die Beschr¨ankung auf drei Dimensionen ist nicht theoretisch begr¨undet. M¨oglich w¨aren auch zwei oder mehr Dimensionen, wobei gegen zwei Dimensionen jedoch die schlechteren Ergebnisse und gegen mehr Dimensionen die Probleme bei Darstell- und Verst¨andlichkeit sprechen.

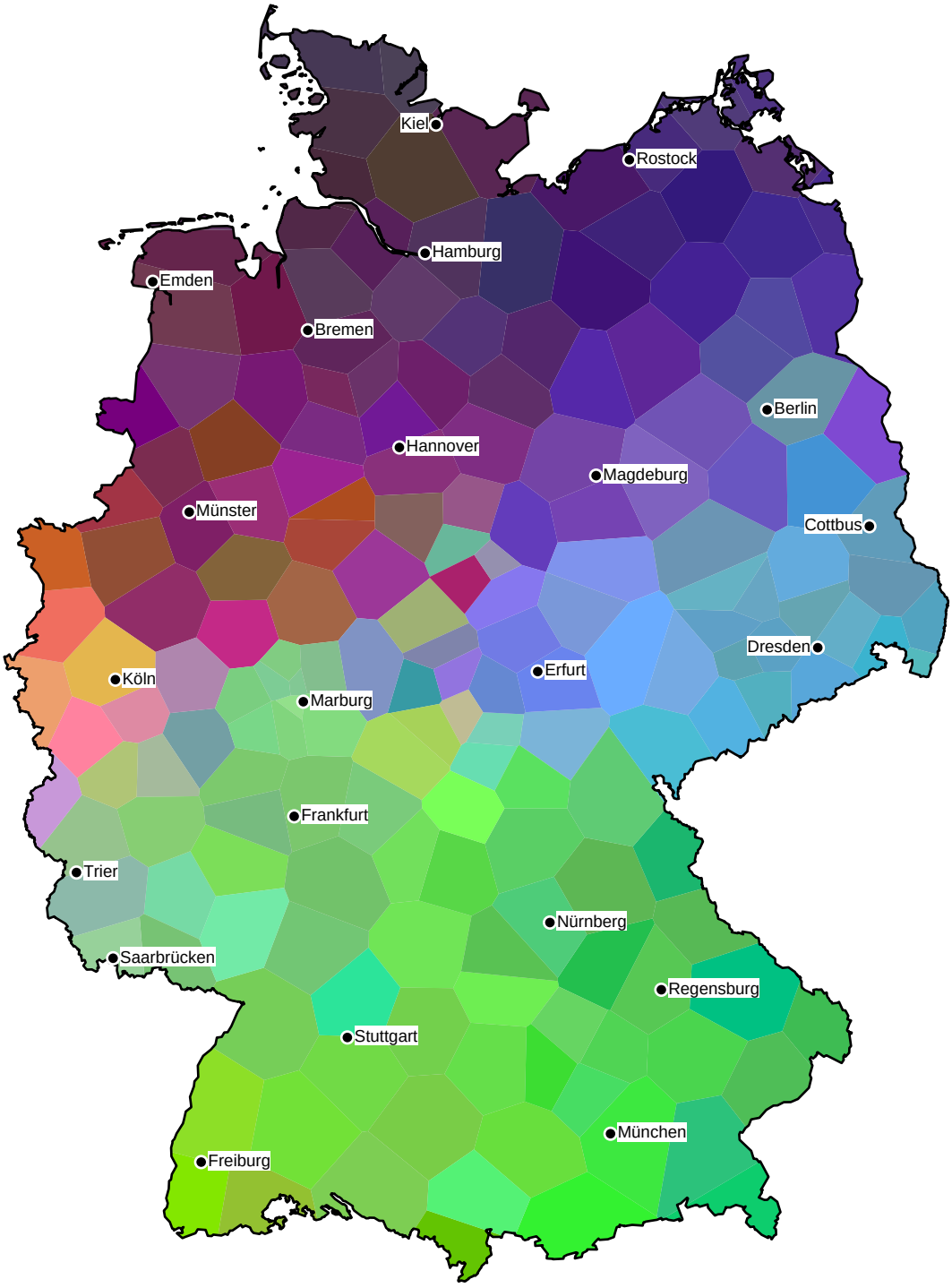

Abbildung 7: Visualisierung des Dialektkontinuums mittels multidimensionaler Skalierung. Die Farbbezeichnungen sind u.U. diskussionsbedürftig, um jedoch Missverständnissen im Text vorzubeugen, legen wir uns an dieser Stelle auf folgende Termini fest: Oberdeutsch entspricht grün, Ostmitteldeutsch hellblau, Ostniederdeutsch dunkel-lila und Westniederdeutsch weitestgehend rötlich-violett, wobei auch Westniederdeutsch in Richtung des rheinischen Fächers zunehmend heterogener wird, für den wir die Bezeichnung '(rötlich) bunt' verwenden.

+

hellblau, Niederdeutsch dunkel-lila bis r¨otlich-violett, in Richtung des rheinischen F¨achers zunehmend r¨otlich-bunt) relativ deutlich erkennen. Tats¨achlich erweist sich nur die Abgrenzung der Dialektgebiete im Westen als wirklich unm¨oglich. Innerhalb der großen Gebiete sind weitere Farbabstufungen auszumachen, die Grenzen verlaufen hier allerdings noch um einiges fließender. So ist in dem von uns als 'Oberdeutsch' definierten gr¨unen Gebiet ein etwas heller gef¨arbter Bereich im Westen zu erkennen, der mit Hessisch, Mosel- und Rheinfr¨ankisch korrespondiert sowie ein tendenziell dunklerer Bereich im Osten, der in etwa Ostfr¨ankisch und Bairisch entspricht. Zudem sind einige leicht anders gef¨arbte 'Ausreißer' zu erkennen, die meist in der Nachbarschaft gr¨oßerer St¨adte oder im unmittelbaren Grenzbereich zu finden sind. Dies m¨usste allerdings noch n¨aher untersucht werden. Im niederdeutschen Dialektraum l¨asst sich ein eher dunkel-lila gef¨arbter ostniederdeutscher von dem ins R¨otlich-Violette tendierenden westniederdeutschen Bereich abgrenzen. Letzterer weist allerdings vor allem in Richtung S¨uden zum rheinischen F¨acher hin eine so große Heterogenit¨at auf, dass in der gesamten Region keine klaren Grenzen auszumachen sein.

Wie wir zeigen konnten, sind die mit dem Levenshteinschen Sequenzabstandsmaß gewonnenen Ergebnisse nicht nur geographisch koh¨arent, sie decken sich alles in allem auch mit der Einteilung der traditionellen Dialektologie. Nat¨urlich ist Letzteres kein Qualit¨atskriterium an sich, es f¨ordert jedoch das Vertrauen in eine neue Methode, dass man eine Kongruenz mit bekannten Techniken und Resultaten sehen kann. Insgesamt gelangen wir daher zu dem Schluss, dass ein systematischer oder gar mechanischer Vergleich aggregierter Ausspracheunterschiede eine zufrieden stellende alternative Basis f¨ur die Dialektgeographie darstellt, wodurch nicht nur Resultate erzielt werden k¨onnen, die den mit traditionellen Ans¨atzen gewonnenen nahezu ebenb¨urtig sind, sondern zus¨atzlich, wie mit der Darstellung des Dialektkontinuums mittels multidimensionaler Skalierung demonstriert, bislang unzug¨angliche M¨oglichkeiten er¨offnet werden.

5 Ausblick

Diese Arbeit ist ein Bericht aus einem laufenden Projekt. Eine detaillierte Betrachtung vieler Alternativen innerhalb der Messtechnik ist sinnvoll, wie auch n¨ahere Untersuchungen zum Einfluss der phonetischen Merkmale, der Auswahl der W¨orter sowie der Auswahl der Erhebungsorte. Weiterhin w¨are es zweckm¨aßig, die Konsistenz und Validit¨at der Messungen genauer zu betrachten (Heeringa, Nerbonne & Kleiweg 2002), und auch die traditionelle Einteilung der Dialektgebiete aus diesem Blickwinkel nochmals zu inspizieren.

Es w¨are jedoch falsch zu vermuten, dass die bloße Best¨atigung der zeitgen¨ossischen Sicht die Ziele dieser Forschungsrichtung ersch¨opft. Wie Nerbonne & Heeringa (1998) und Heeringa & Nerbonne (2002) in Ans¨atzen zeigen, er¨offnet die Verwendung des Levenshteinschen Sequenzabstandsmaßes eine Reihe neuer M¨oglichkeiten. So ist es offensichtlich, dass diese Methode bei geeignetem Material auch auf Fragen der Variet¨atslinguistik außerhalb der Dialektgeographie angewendet werden k¨onnte. Vielleicht noch interessanter d¨urfte allerdings der Aspekt sein, dass sich die Dialektologie mit diesem quantitativen Maß nun die Techniken der numerischen Analyse erschließt, welche entscheidend dazu beitragen k¨onnten, Antworten auf Fragen nach z.B. dem Einfluss außersprachlicher Variablen auf die sprachliche Variet¨at zu finden.

Literatur

Aldenderfer, Mark S. & Roger K. Blashfield (1984), *Cluster Analysis*, Quantitative Applications in the Social Sciences, Sage, Beverly Hills.

Almeida, Almerindo & Angelika Braun (1986), ‘‘Richtig’ und ’Falsch’ in phonetischer Transkription: Vorschl¨age zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten’, *Zeitschrift für Dialektologie und Linguistik* **LIII**(2), 158–172.

Bloomfield, Leonard (1933), *Language*, Holt, Rhinehart and Winston, New York.

Bolognesi, Roberto & Wilbert Heeringa (2002), ‘De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten’, *Gramma/TTT: Tijdschrift voor Taalwetenschap* **9**(1), 45–84.

Chomsky, Noam A. & Morris Halle (1968), *The Sound Pattern of English*, Harper and Row, New York.

Coseriu, Eugenio (1975), *Die Sprachgeographie*, Gunter Narr, T¨ubingen. [1]1956.

Goebl, Hans (1982), *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichische Akademie der Wissenschaften, Wien.

Goebl, Hans (1984), *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol.*, Max Niemeyer, T¨ubingen.

G¨oschel, Joachim (1992), Das Forschungsinstitut f¨ur Deutsche Sprache “Deutscher Sprachatlas´´, Wissenschaftlicher Bericht, Das Forschungsinstitut f¨ur Deutsche Sprache, Marburg.

Heeringa, Wilbert (2004), Measuring Dialect Pronunciation Differences using Levenshtein Distance, PhD thesis, Rijksuniversiteit Groningen.

Heeringa, Wilbert & Angelika Braun (2003), ‘The use of the Almeida-Braun system in the measurement of Dutch dialect distances’, *Computers and the Humanities* **37**(3), 257–271. Special Issue on Computational Techniques in Dialectometry, ed. by John Nerbonne and William Kretzschmar.

Heeringa, Wilbert & Charlotte Gooskens (2003), ‘Norwegian dialect examined perceptually and acoustically’, *Computers and the Humanities* **37**(3), 293–315.

Heeringa, Wilbert & John Nerbonne (2002), ‘Dialect areas and dialect continua’, *Language Variation and Change* **13**, 375–400.

Heeringa, Wilbert, John Nerbonne & Peter Kleiweg (2002), Validating dialect comparison methods, *in* W.Gaul & G.Ritter, eds, ‘Proceedings of the 24th Annual Meeting of the Gesellschaft f¨ur Klassifikation’, Springer, Heidelberg, pp. 445–452.

Herrgen, Joachim & J¨urgen Erich Schmidt (1989), Kontrastive Dialektgeographie, *in* W.Putsche, W.Veith & P.Wiesinger, eds, ‘Dialektgeographie und Dialektologie’, Vol. 90 of *Deutsche Dialektgeographie*, N.G.Elwert Verlag, Marburg, pp. 304–346.

Hummel, Lutz (1993), *Dialektometrische Analysen zum kleinen deutschen Sprachatlas (KDSA)*, Max Niemeyer Verlag, T¨ubingen.

IPA (1949), *The Principles of the International Phonetic Association*.

Jain, K. & R. C. Dubes (1988), *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey.

K¨onig, Werner (1994), *DTV Atlas zur deutschen Sprache*, Deutscher Taschenbuch Verlag, M¨unchen. [1]1978.

Kruskal, Joseph (1999), An overview of sequence comparison, *in* D.Sankoff & J.Kruskal, eds, 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison', CSLI, Stanford, pp. 1–44. [1]1983.

Lameli, Alfred (2004), *Standard und Substandard: Regionlismen im diachronen Längschnitt*, Steiner, Stuttgart.

Lenz, Alexandra (2002), Struktur und Dynammik des Substandards. Eine Studie zum Westmitteldeutschen (Wittlich/Eifel), PhD thesis, Marburg.

M¨oller, Robert (2003), 'Zur diatopischen Gliederung des alltagsprachlichen Wortgebrauchs. eine dialektometrische Auswertung von J¨urgen Eichhoff: Wortatlas der Deutschen Umgangsprache, Bd. 1/4 (Bern/M¨unchen)', *Zeitschrift für Dialektologie und Linguistik* **LXX**(3), 259–297.

Nerbonne, John & Wilbert Heeringa (1998), 'Computationele vergelijking en classificatie van dialecten', *Taal en Tongval* **50**(2), 164–193.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg (1999), Edit distance and dialect proximity, *in* D.Sankoff & J.Kruskal, eds, 'Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.', CSLI, Stanford, CA, pp. v–xv.

Niebaum, Hermann & J¨urgen Macha (1999), *Einführung in die Dialektologie des Deutschen*, Niemeyer, Tuebingen. [1]1983.

Séguy, Jean (1971), 'La relation entre la distance spatiale et la distance lexicale', *Revue de Linguistique Romane* **35**, 335–357.

Tourgerson, Warren S. (1952), 'Multidimensional scaling I', *Psychometrika* **17**, 401–419.

Vieregge, Wilhelm H., A.C.M. Rietveld & Carel Jansen (1984), A distinctive feature based system for the evaluation of segmental transcription in Dutch, *in* M. P.van den Broecke & A.Cohen, eds, 'Proc. of the 10th International Congress of Phonetic Sciences', Dordrecht, pp. 654–659.