Neue Analyse- und Visualisierungsmethoden in der Dialektometrie

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie
der Philosophischen Fakultät der
Eberhard Karls Universität Tübingen
vorgelegt von Thomas Zastrow
aus Boppard am Rhein

2011

Gedruckt mit Genehmigung der
Philosophischen Fakultät der Eberhard Karls Universität Tübingen

Erstgutachter: Prof. Dr. Erhard Hinrichs
Zweitgutachter: Prof. Dr. Gerhard Jäger
Tag der mündlichen Prüfung: 5.7.2011
Dekan: Prof. Dr. Jürgen Leonhardt
Verlag: TOBIAS-lib, Tübingen

Danksagung

Für meine Doktorarbeit gebührt vielen Menschen mein herzlicher Dank. Besonders möchte ich mich bei meinem Doktorvater Herrn Prof. Dr. Erhard Hinrichs bedanken, er brachte mir stets sehr viel Geduld entgegen und sorgte mit wertvollen Ratschlägen für das Gelingen der Arbeit.

Ich danke Herrn Prof. Dr. Gerhard Jäger für die Erstellung des Zweitgutachtens.

Ein großer Dank geht an die Kolleginnen und Kollegen des Buldialects Projekts - ohne ihre stetige und wertvolle Unterstützung wäre mir die Bearbeitung eines für mich völlig neuen Themas nicht möglich gewesen: Georgi Kolev, Prof. Dr. John Nerbonne, Dr. Petya Osenova, Dr. Jelena Prokic, Dr. Kiril Simov und Prof. Dr. Vladimir Zhobov.

Stets anregend und inspirierend waren die Diskussionen mit Prof. Dr. Hans Goebl. Sie erweiterten meine Sichtweise und viele der dieser Arbeit zugrundeliegenden Ideen sind in diesen Gesprächen entstanden. Ohne die Salzburger Dialektometrie-Software VDM wären viele der hier gezeigten Visualisierungen nicht möglich gewesen.

Ich danke Kathrin Beck für aufmerksames und intensives Korrekturlesen.

Inhaltsverzeichnis

Kapitel 1

Einführung

In der Dialektometrie werden sprachliche Unterschiede und Ähnlichkeiten zwischen den Dialekten einer Sprache mit mathematischen Methoden gemessen und quantifiziert. Hierzu kommen verschiedene Methoden der Statistik, der numerischen Klassifikation und der Mustererkennung zum Einsatz. Die Ergebnisse dieser Algorithmen können auf verschiedene Art und Weise analysiert werden: Algorithmen aus dem Bereich Maschinelles Lernen und der Dimensionsreduktion ermitteln Gruppierungen und Trennlinien innerhalb der Gesamtheit der betrachteten Dialektdaten. Anschließend können die Ergebnisse visualisiert werden, wobei der graphischen Darstellung in Form von Landkarten besondere Bedeutung zukommt.

In der vorliegenden Arbeit werden neue Analyse- und Visualisierungs Methoden im Bereich der Dialektometrie entwickelt. Die einzelnen Methoden unterscheiden sich unter anderem in der Behandlung der untersuchten Daten: Aggregierende Methoden beziehen den gesamten Datensatz in die Analyse mit ein, extrahierende Methoden hingegen legen das Augenmerk auf die Betrachtung einzelner Elemente. Es ergeben sich verschiedene Blickwinkel auf die Daten, die wiederum zu unterschiedlichen Ergebnissen und Interpretationen eines Datensatzes führen können.

Die beschriebenen Methoden werden auf einen umfangreichen Datensatz der bulgarischen Sprache angewandt und die Ergebnisse anschließend analysiert. In den darauffolgenden Visualisierungen liegt der Schwerpunkt der Darstellung auf verschiedenen Arten topographischer Karten. Um die Ver-

gleichbarkeit der Ergebnissen der unterschiedlichen Methoden erhalten zu können, erfolgen Analyse und Visualisierung der Ergebnisse jeweils unter Beibehaltung eines fixen Satzes an Parametern. Es folgt die Gegenüberstellung zweier etablierter dialektometrischer Methoden anhand eines italienischen Dialektdatensatzes.

Nach einer allgemeinen Einleitung in das Thema Dialekte (Kapitel 2) folgt ein Abschnitt über Dialektologie, die sich mit wissenschaftlicher Methodik dem Thema Dialekte widmet (Kapitel 3). In zeitlicher Reihenfolge gesehen, stellt die Dialektologie einen direkten Vorgänger der Dialektometrie dar: Letztere wird anschließend behandelt, es werden die benutzte Terminologie und bereits bekannte und erprobte Methoden der Dialektometrie dargestellt (Kapitel 4). Hierauf aufbauend werden drei neue im Rahmen dieser Arbeit entwickelte dialektometrische Methoden vorgestellt. Dabei handelt es sich um eine vektorbasierte Herangehensweise (Kapitel 5), eine N-Gramm-Analyse (Kapitel 6) und die Implementierung verschiedener informationstheoretischer Ansätze (Kapitel 7). Nach Anwendung dieser Methoden können die Ergebnisse auf verschiedene Weise analysiert und zum Zweck der Interpretation visualisiert werden (Kapitel 8). Hierzu werden Methoden aus dem Bereich des Maschinellen Lernens wie hierarchisches Clustering, Multidimensional Scaling und Intervallalgorithmen angewendet. Einige dieser Methoden sind in der Anwendung *Visual Dialectometry* (VDM) implementiert. Hierbei handelt es sich um eine Anwendung, die an der Universität Salzburg entwickelt wurde und die unter anderem zur Analyse und Visualisierung von dialektometrischen Ergebnissen eingesetzt werden kann. VDM wird auch in dieser Arbeit angewandt und in Kapitel 9 näher beschrieben.

Nach diesen einführenden Ausführungen wird im Kapitel 10 das *Buldialects-Projekt* und der in diesem Projekt entstandene Dialektdatensatz vorgestellt. Auf den phonetischen Teil dieses Datensatzes werden die zuvor vorgestellten dialektometrischen Methoden angewendet (Kapitel 11). Die Ergebnisse werden auf verschiedene Art und Weise analysiert und visualisiert, unter anderem werden hierarchisches Clustering und Intervallalgorithmen angewendet. Es folgt eine Analyse der Verteilung des Wortakzents in der bulgarischen Sprache. Die dreidimensionale Reliefdarstellung als Möglichkeit, eine weitere Dimension in die Visualisierung einzubeziehen, wird dargestellt

und auf die Ergebnisse der N-Gramm-Analyse angewendet. Schließlich werden die Ergebnisse der neu entwickelten dialektometrischen Methoden mit den Aussagen der traditionellen bulgarischen Dialektologie verglichen. Da der Datensatz des Projekts Buldialects auch einen lexikalischen Teil enthält, wird dieser in Form eines Sprachatlas aufbereitet dargestellt und mittels der dialektometrischen Methode *Relativer Identitätswert* analysiert (Kapitel 12).

Abschließend findet sich die Analyse und Gegenüberstellung italienischer Dialektdaten aus dem "Sprachatlas des Dolomitenladinischen und angrenzender Dialekte". Hierzu werden zwei etablierte dialektometrische Methoden (Relativer Identitätswert und Edit-Distance-Methode) auf die italienischen Dialektdaten angewandt und die Ergebnisse anschließend miteinander verglichen (Kapitel 13). In den Schlußbetrachtungen werden die Ergebnisse dieser Arbeit zusammengefasst und einander gegenübergestellt (Kapitel 14).

Kapitel 2

Dialekte

Der Begriff "Dialekt" existierte bereits in den Sprachen der Antike: Dialekt - Ursprünglich griechisch, *dia-lektos*, aber auch lateinisch: "Redeweise". In der heutigen Sprachwissenschaft werden Dialekte in der Dialektologie und in der darauf aufbauenden Dialektometrie erforscht. Dabei kann der Begriff "Dialekt", je nach Forschungsausrichtung unterschiedlich definiert werden. Gemeinsam sind den verschiedenen Verwendungen des Begriffs zwei Eigenschaften, die somit den kleinsten Nenner darstellen (siehe hierzu bspws. Bußmann 2002, S. 162 ff.):

- Dialekte werden als Gegensatz zur *Hochsprache* (häufig auch als *Standardsprache* bezeichnet) aufgefasst.
- Sie sind auf ein geographisches Gebiet begrenzt, welches einen Teil des Verbreitungsgebietes der jeweiligen Sprache umfasst.

Dialekte stellen somit einerseits die Ausdifferenzierung einer Sprache in Bezug auf ihre jeweilige geographische Ausdehnung dar. Sie entstehen unter anderem durch fehlende Sprachkontakte zwischen geographisch voneinander getrennten Gruppen einer Sprache (siehe Kapitel 2.1). Dialekte sind abgegrenzt zu anderen sprachlichen Varianten, wie sie beispielsweise in der Soziolinguistik untersucht werden, können diese aber unter bestimmten Voraussetzungen ergänzen. Sie unterscheiden sich von den individuellen Eigenschaften einzelner Sprecher, den Idiolekten (Bußmann, 2002, S. 289) und

sind immer einer Gruppe von Sprechern zugeordnet. Zur Definition von Dialekten spielt das gegenseitige Verstehen eine untergeordnete Rolle. Es gibt im Deutschen Sprecher unterschiedlicher Dialekte, die sich nur schlecht miteinander verständigen können, so dass in entsprechenden Situationen auf die Hochsprache ausgewichen wird. Im skandinavischen Raum herrscht die entgegengesetzte Situation: Hier können sich in manchen Situationen sogar Sprecher unterschiedlicher Sprachen gegenseitig verstehen, beispielsweise Dänen und Norweger (siehe zur sprachübergreifenden Kommunikation in den skandinavischen Sprachen: Gooskens 2007).

Andererseits steht den Dialekten die Hochsprache gegenüber. Bei der Hochsprache handelt es sich meistens ebenfalls um einen Dialekt, der aber durch kulturelle, politische oder ähnliche Gründe im gesamten Verbreitungsgebiet einer Sprache als gültig erklärt worden ist und dessen Verwendung nicht (mehr) nur an eine geographische Region gebunden ist. Die Standard- oder Hochsprache wird in den Medien verwendet und in den Schulen im Sprachunterricht vermittelt (siehe hierzu auch Bußmann 2002, S. 648). Oft existiert für die Hochsprache als einziger Dialekt einer Sprache eine verbindliche Schriftsprache, die anderen Dialekte werden in diesem Falle lediglich mündlich übermittelt oder in der Schriftsprache der Hochsprache niedergeschrieben. Dialekt und Hochsprache unterscheiden sich in verschiedenen Eigenschaften (Auflistung nach Löffler 2003):

- Vollständigkeit: Während Dialekte häufig nur die hochfrequent benutzten Bereiche einer Sprache abdecken, ist die Hochsprache vollständig auf allen sprachlichen Ebenen entwickelt.

- Verwendungssituation: Dialekte finden meist nur im engeren Kreis von Familie, Verwandten und Freunden Verwendung, in über die Familie hinausgehenden Situationen wird eher die Hochsprache eingesetzt. Letztere hat dementsprechend eine weitere kommunikative Wirkung als der begrenzt verwendete Dialekt.

- Soziologisches Milieu: Unterschiedliche soziologische Schichten verwenden die entsprechenden Dialekte in anderer Art und Weise.

- Diachrone Ausdifferenzierung: Dialekte werden als Vorgänger der sich erst später etablierenden Hochsprache angesehen.

Im Gegensatz zu regional begrenzten Dialekten unterliegt die Entwicklung der Hochsprache meistens einer institutionellen, staatlichen Kontrolle[1]. Dass diese staatliche Kontrolle häufig von breiten Schichten der jeweiligen Sprecher abgelehnt wird, hat die jahrelange Debatte um die Reform der deutschen Sprache gezeigt. Zusätzlich ergibt sich bei Sprachen, die über Ländergrenzen hinweg gesprochen werden, das Problem der supranationalen Zusammenarbeit: Es müssen unter Umständen mehrere, national geprägte Reformbewegungen harmonisiert werden[2]. Durch die Eingriffe der staatlichen Kontrollinstanzen kommt es zu Abweichungen der Hochsprache vom zugrunde liegenden Dialekt. Die Hochsprache entwickelt ein Eigenleben, sie wird zu einem neuen, selbständigen Dialekt der nicht mehr ortsgebunden ist. In der Dialektforschung kann die Hochsprache als *Goldstandard* verwendet werden: Die quantifizierten sprachlichen Unterschiede der einzelnen Dialekte werden hierbei jeweils in Relation zu den der Hochsprache entsprechenden dialektalen Kennzahlen gesetzt.

Durch die breite Anwendung der Hochsprache im 'offiziellen' Rahmen (Schule, Behörden etc.) und die stark zugenommene Verbreitung der (elektronischen) Medien in den letzten Jahrzehnten ist die Verwendung von Dialekten im deutschen Sprachraum zurückgegangen. Im öffentlichen Raum wird der Dialekt nur noch in einigen wenigen Bereichen bewusst eingesetzt - beispielsweise in der Volksmusik, der Mundartliteratur oder in politischen Reden. Ungeachtet dessen erfreut er sich im privaten Bereich auch weiterhin hoher Beliebtheit und Akzeptanz[3]. Die meisten Dialektsprecher sind heute in der Lage, auch in der Hochsprache zu kommunizieren.

Zwischen den Dialekten und der Hochsprache findet sich die *Umgangssprache*. Sie lässt sich nur schwer definieren bzw. von Dialekt und Hoch-

[1]Für die Orthographie der deutschen Hochsprache übernimmt diese Aufgabe der "Rat für deutsche Rechtschreibung". (http://rechtschreibrat.ids-mannheim.de/)

[2]Im Falle der deutschen Rechtschreibreform waren dies Deutschland, Österreich und die Schweiz.

[3]Zu diesem Themenkomplex siehe Frahm (2003) und hier besonders den Artikel 'Renaissance des Dialekts? - Eine empirische Studie' von Eva-Maria Walker und Felicitas Hartmann, S. 149-154.

sprache abtrennen und wird teilweise sogar regional unterschiedlich definiert (König u. Renn, 2007, page 30 ff.). Der Begriff "Mundart" wird meistens synonym zu Dialekt verwendet.

Dialekte können sich auf allen Ebenen der Sprache (Phonetik, Lexik, Morphologie und Syntax) unterschiedlich stark voneinander unterscheiden. Hauptsächlich in den Bereichen Phonetik und Lexik sind Dialekte Gegenstand der Forschung, einige wenige Ansätze beschäftigen sich auch mit syntaktisch bedingten Unterschieden zwischen Dialekten (für das Holländische siehe Spruit 2006).

2.1 Sprachkontakt und Dialekt

Aktuell existieren auf der Erde circa 7.000 verschiedene lebendige und aktiv gesprochene Sprachen[4]. Diese lassen sich zu *Sprachfamilien* zusammenfassen: Innerhalb einer Sprachfamilie weisen die enthaltenen Sprachen einen gewissen Grad an Ähnlichkeit zueinander auf, sie gehen alle auf einen gemeinsamen Vorgänger zurück. Neben den Sprachfamilien existieren *isolierte Sprachen*, die sich keiner Sprachfamilie zuordnen lassen[5].

Unabhängig ihres Verwandschaftsgrades kommt es zwischen den Sprachen zu vielfältigem Austausch: *Sprachkontakte* zwischen den Sprechern verschiedener Sprachen führen zu Veränderungen in den beteiligten Sprachen. Die sichtbarste Auswirkung vielfältiger Sprachkontakte ist das Vorhandensein vieler Lehn- und Fremdwörter.

Aber auch die dem Sprachkontakt entgegengesetzte Entwicklung ist allgegenwärtig: Gruppen von Sprechern innerhalb einer Sprache haben immer weniger sprachlichen Austausch miteinander. Auch hierfür können viele Gründe ausschlaggebend sein: Im selben Maße, wie sie zu stärkerem Sprachkontakt zwischen Sprachen beitragen können, können Faktoren wie geänderte politische Verhältnisse oder kulturelle Umorientierungen auch das Gegenteil bewirken. Fehlender Sprachkontakt über einen längeren Zeitraum resultiert schließlich in einer allmählichen Ausdifferenzierung der in den jeweiligen

[4] In Gordon (2005) wird die Zahl der aktuell lebendigen Sprachen mit 6.909 beziffert.

[5] Hier darf der Term "isolierte Sprache" nicht im Sinne geringer Sprachkontakte, wie unten beschrieben, verstanden werden.

Gruppen gesprochenen Sprachen. Hieraus können sich verschiedene Dialekte der entsprechenden Sprache entwickeln. Aus diesen Dialekten können dann wiederum bei immer weiter voranschreitender Auseinanderentwicklung eigene, neue Sprachen entstehen. Die Grenze zwischen einer dialektaler Variante einer bestehenden Sprache und einer gänzlich neuen Sprache ist fließend und häufig nicht eindeutig feststellbar. Im Falle von neu entstandenen Sprachen können diese dann wiederum zu einer neuen Sprachfamilie zusammengefasst werden.

Kapitel 3

Dialektologie

Bei der Dialektologie handelt es sich um den Bereich der Sprachwissenschaft, der sich unter Zuhilfenahme wissenschaftlicher Methoden mit Dialekten befasst (Bußmann, 2002, S 163 ff.). Ihre Urspünge gehen bis in das 17. Jahrhundert zurück. In Deutschland waren es unter anderen die Brüder Grimm, die ein wissenschaftliches Interesse an der dialektalen Ausdifferenzierung ihrer Muttersprache zeigten (Haas, 1990)[1]. In der zweiten Hälfte des 19. Jahrhunderts begann Georg Wenker mit der systematischen Erhebung von Dialektdaten im deutschsprachigen Raum (Veith u. a., 1984).

Ausgangspunkt aller dialektologischen Untersuchungen sind empirisch gewonnene *Dialektdaten*. Aus diesen werden Sprachatlanten und Dialektwörterbücher erstellt. In ihnen werden die Dialektdaten nach unterschiedlichen Kriterien geordnet; sie bieten so einen systematischen Zugang zu den zuvor erhobenen Dialektdaten. In diesen lassen sich dann Dialektstrukturen erkennen, die, einzeln oder gebündelt, als *Isoglossen* oder *Dialektkontinua* auftreten können (siehe Kapitel 3.3.1 und folgende).

[1] Im 19. Jahrhundert in Deutschland betriebene Dialektanalysen hatten nicht selten die Motivation, über gemeinsame sprachliche Wurzeln eine einheitliche "deutsche Nation" postulieren zu können.

3.1 Dialektdaten

3.1.1 Erhebung

Grundlage jeder dialektologischen Untersuchung sind die Dialektdaten. Bevor diese erhoben werden können, muss ein *Fragenkatalog* erstellt werden. Der Fragenkatalog enthält in standardisierter Form Fragen, die den Sprechern des jeweiligen Dialekts vorgelegt oder - je nach Methode - mündlich abgefragt werden. Aus dem Fragenkatalog können anschließend individuelle Fragebögen erstellt werden. Zusätzlich zu den Fragen des Fragenkatalogs enthält der Fragebogen Metadaten über den jeweiligen Sprecher (Alter, Geschlecht) sowie Ort und Zeit der Erhebung, eventuell auch Angaben über den Interviewer bzw. den oder die Transkribierenden. Er nimmt während der Datenerhebung oder in dessen Nachbereitung die erste verschriftlichte Form der Dialektdaten auf. Die zusätzliche Konservierung der Audio-Aufnahmen auf Tonband - oder heute digital - ermöglichen es, auch im Nachhinein auf die originalen Sprachdaten zugreifen zu können und weitere Transkriptionen oder andere Analysen durchführen zu können.

Der Zusammenstellung des Fragenkatalogs kommt große Bedeutung zu, besonders auch im Hinblick auf die hohen formalen Anforderungen der Dialektometrie (siehe Kapitel 4). Im Fragenkatalog wird festgelegt, welche linguistischen Elemente der untersuchten Sprache die Autoren einer Untersuchung in dialektologischer Hinsicht für relevant befinden: Die Auswahl der zu untersuchenden Elemente ist abhängig vom jeweiligen Forschungsinteresse. Fragenkataloge sind von Sprache zu Sprache verschieden: Sprachliche Elemente, die in einer Sprache dialektal ausdifferenziert sind, weisen in einer anderen Sprache unter Umständen keine oder nur geringe dialektale Unterschiede auf.

Für dialektometrische Untersuchungen besonders geeignet sind Fragenkataloge auf Wortbasis. Einen ersten Anlaufpunkt zur Erstellung eines solchen Fragenkatalogs bilden die sprachspezifischen *Swadesh-Listen*. Sie gehen auf den amerikanischen Linguisten Morris Swadesh zurück und enthalten pro Sprache ca. 200 Wörter, die als allgemeingültig und in jeder Sprache vor-

handen angesehen werden können[2]. Aufgrund ihrer Allgemeingültigkeit sind die Wörter der Swadesh-Listen eher kurz, viele bestehen nur aus einer Silbe. Da Swadesh-Listen zum (lexikologischen) Vergleich verschiedener Sprachen entwickelt wurden, müssen sie in der Regel zur Verwendung *innerhalb* einer Sprache und ihrer Dialekte angepasst werden.

Ist der Fragenkatalog definiert, beginnt anschließend die eigentliche Erhebung der Dialektdaten. Auch diese kann wieder in verschiedener Form erfolgen. In der *Interviewmethode* werden die einzelnen Dialekt-Regionen von linguistisch geschulten *Fieldworkern* besucht. Die Daten werden in persönlichen Gesprächen gesammelt[3]. Die Fieldworker stellen die Fragen des Fragenkatalogs authentischen Sprechern des regionalen Dialektes und notieren die Antworten auf den Fragebögen.

Ein Nachteil der Interviewmethode besteht darin, dass bereits das Interview an sich eine für die Probanden ungewohnte Situation darstellt. Die Antworten können so unbewusst verfälscht werden. Eine weitere, heute allerdings weniger gebräuchliche Methode, ist die *Korrespondenzmethode*: Hierbei wird der Fragebogen per Post an Lehrer oder andere versierte Sprecher aus den jeweiligen Dialektregionen verschickt. Diese füllen den Fragebogen aus und schicken ihre Antworten wiederum per Post zurück an den Fragesteller (Goebl, 2004, S. 250-251). Hierbei ergibt sich das Problem, dass nicht unbedingt authentische Sprecher der jeweiligen Dialekte als Gewährsleute herangezogen werden und auch dadurch die Daten unabsichtlich verfälscht werden können.

Ebenso hat die Auswahl der Sprecher einen hohen Einfluss auf die endgültigen Dialektdaten. Faktoren wie Alter, Geschlecht und soziale Herkunft sollen bei der Erhebung von Dialektdaten möglichst ausgeblendet werden. Als besonders geeignet haben sich Personen, die dem sogenannten *Norm*-Schema entsprechen, erwiesen:

- Non-educated

[2]Eine Gegenüberstellung der Swadesh-Listen für mehrere europäische Sprachen findet sich hier: http://de.wiktionary.org/wiki/Wiktionary:Swadesh-Liste (eingesehen am 5.2.2008).

[3]Es gibt bezüglich der Formalität mehrere Abstufungen in der Interviewmethode, für weitere Informationen, siehe Chambers u. Trudgill (1980), S. 24-33.

- Old
- Rural
- Male

Die Erhebung von Dialektdaten muss sich aber nicht auf “Norm”-Sprecher beschränken. So werden beispielsweise im Projekt SiN[4] ausnahmslos Frauen und keine Männer interviewt.

Zusammengefasst und auf (Land-) Karten annotiert, werden Dialektdaten in Form von Sprachatlanten oder Dialektwörterbüchern herausgegeben (zu Sprachatlanten, siehe 3.2).

3.1.2 Arten von Dialektdaten

Wie bereits erwähnt, können sich Dialekte in allen Ebenen der Sprache unterscheiden. Dies führt dazu, dass auch empirisch erhobene Dialektdaten unterschiedlich sein können und ein Vergleich der Daten erschwert wird.

In der Dialektologie und Dialektometrie sind bislang hauptsächlich drei unterschiedliche Arten von Dialektdaten analysiert worden: lexikalische, phonetische und syntaktisch variierende Dialektdaten.

Lexikalische Daten

Lexikalische Dialektdaten beruhen auf unterschiedlichen, lexikalischen Realisierungen ein und desselben Konzepts in den verschiedenen Dialekten einer Sprache.

Abbildung 3.1 zeigt zwei Karten aus König (2005), S. 166. Dargestellt sind jeweils die Konzepte ”Junge” und ”Mädchen”. Im Norden Deutschlands und in der Hochsprache wird der Begriff “Junge” für einen männlichen jungen Menschen gebraucht. Im Süden ist hierfür der Begriff “Bua” gebräuchlich und in einigen Gebieten kommen auch gänzlich andere Begriffe wie beispielsweise ”Kerl” zur Anwendung. Ebenfalls unterschiedlich ist das Lexem für einen jungen weiblichen Menschen: “Deern” im Norden, “Mädchen” in der Mitte

[4]Projekt ”Sprachvariationen in Norddeutschland”:
http://sin.sign-lang.uni-hamburg.de/drupal/

und “Diandl” im Südosten Deutschlands (König, 2005, S. 166-167). Innerhalb des Verbreitungsgebiets eines Lexems kann es dann wiederum zu phonetisch motivierten Variationen des Lexems kommen (“Jong”, “Jung”, “Junge” oder “Mädchen”, “Mäken”, “Medche(n)“).

Lexikalisch variierende Dialektdaten können auf topographischen Karten eingezeichnet werden. In der Regel werden sich unterschiedliche lexikalische Realisierungen eines Konzepts auch in ihrer geographischen Verteilung unterscheiden. So kennzeichnen benachbarte Gebiete, in denen die gleiche Variante des entsprechenden Begriffs verwendet wird, auch ein zusammenhängendes *Dialektareal*. Dabei müssen die identifizierten Dialektareale zweier semantischer Konzepte nicht unbedingt nahtlos aneinander anschließen: So verläuft die geographische Grenze zwischen “Junge” und “Bua” quer in West-Ost-Richtung durch das Verbreitungegebiet von “Mädchen” und teilt dieses faktisch in zwei Teile.

Im Gegensatz zu phonetischen Dialektdaten sind bei lexikalischen Daten die Wörter als atomare Einheiten anzusehen. Sie lassen sich nicht weiter aufsplitten und können nur als Ganzes analysiert werden. Weiterhin können lexikalische Variationen eines Konzepts nicht quantitativ in Relation zueinander gesetzt werden. So lässt sich bei den Begriffen “Deern”, “Mädchen” und “Diandl” nur noch die Verschiedenheit, nicht aber ein feiner granulierter *Grad an Verschiedenheit* angeben. Dies führt dazu, dass die Palette der zur Verfügung stehenden dialektometrischen Methoden in Hinblick auf lexikalische Daten eingeschränkt ist[5].

Phonetische Daten

Phonetische Dialektdaten repräsentieren Unterschiede zwischen Dialekten mittels *phonetisch* verschiedener Realisierungen ein und desselben Lexems. Hier stellt das einzelne Phon die atomare Einheit dar. Im Gegensatz zu lexikalisch basierten Dialektdaten kann ein Vergleich phonetisch basierter Dialektdaten nur dann erfolgen, wenn in allen zu untersuchenden Dialekten ein und dasselbe Lexem für ein Konzept verwendet wird und die Unterschiede sich

[5]Der *Relative Identitätswert* und seine Ableitungen sind geeignet, binäre Variationen (identisch / nicht identisch) von lexikalischen Daten zu aggregieren, siehe Kapitel 4.3.1.

Die Bezeichnungen für *Junge* in den Mundarten des ehemaligen deutschen Sprachgebiets

Die Bezeichnungen für *Mädchen* in den Mundarten des ehemaligen deutschen Sprachgebiets

Abbildung 3.1: Zwei Karten aus König (2005), S. 166: Verschiedene lexikalische und phonetische Variationen der Begriffe ”Junge“ und ”Mädchen“

auf phonetische Variationen beschränken. Um auf das Beispiel aus dem vorangegangenen Kapitel zurückzukommen: Auf phonetischer Basis lassen sich nun Variationen wie “Jong”, “Jung” und “Junge” analysieren. Ein weitergehender Vergleich dieser phonetischen Variationen mit lexikalischen Variationen wie “Bua“ ist nicht mehr möglich. Auf dieser direkten Vergleichbarkeit verschiedener phonetischer Realisierungen eines Lexems beruhen einige dialektometrische Methoden, wie zum Beispiel die *Edit Distance-Algorithmen* sowie die hier vorgestellte *Vektoranalyse*.

Um phonetische Eigenschaften erfassen zu können, müssen die Dialektdaten in eine *Lautschrift* transkribiert werden. Hierzu wurden in der Vergangenheit verschiedene Lautsysteme entwickelt[6]. Eines der am weitesten verbreiteten Systeme ist das “International Phonetic Alphabet”, kurz IPA der “International Phonetic Association“[7].

Das IPA erhebt den Anspruch, sprachübergreifend gültig zu sein und fast alle Laute, die der menschliche Sprechapparat erzeugen kann, abzubilden. Derzeit enthält das IPA 107 eigenständige Symbole sowie 56 *diakritische Zeichen* und *Suprasegementale*[8]. Dabei ist festzuhalten, dass die IPA-Symbole von Sprache zu Sprache unterschiedlich interpretiert werden können. Das ”Handbook of the International Phonetic Association“ (IPA, 1999) enthält aktuell Beschreibungen für 29 Sprachen, wie die IPA-Symbole sprachspezifisch zu interpretieren sind.

Wenn möglich, benutzt das IPA einzelne Zeichen des lateinischen Alphabets. Da diese nicht ausreichend sind, um die Anzahl an möglichen Phonen adäquat abzubilden, kommen griechische Buchstaben sowie einige weitere Zeichen hinzu. Mittlerweile ist das IPA Bestandteil von Unicode und findet sich hier im Bereich U+0250 bis U+02AF[9].

Da viele, auch moderne, Computersysteme noch nicht in der Lage sind, Unicode zu verarbeiten, wurde *X-Sampa* (Extended Speech Assessment Methods Phonetic Alphabet) entwickelt. X-Sampa ist eine Weiterentwicklung von Sampa, welches bereits den Ansatz verfolgte, die Zeichen des IPA elek-

[6]Speziell für die deutsche Dialektologie wurde die Lautschrift *Teuthonista* entwickelt.

[7]Homepage der IPA: http://www.langsci.ucl.ac.uk/ipa/

[8]Stand: 6.2.2008, http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[9]LinuxLibertine und DejaVu sind zwei moderne Fonts, die die IPA Symbole an entsprechender Stelle der Unicode-Tabelle enthalten.

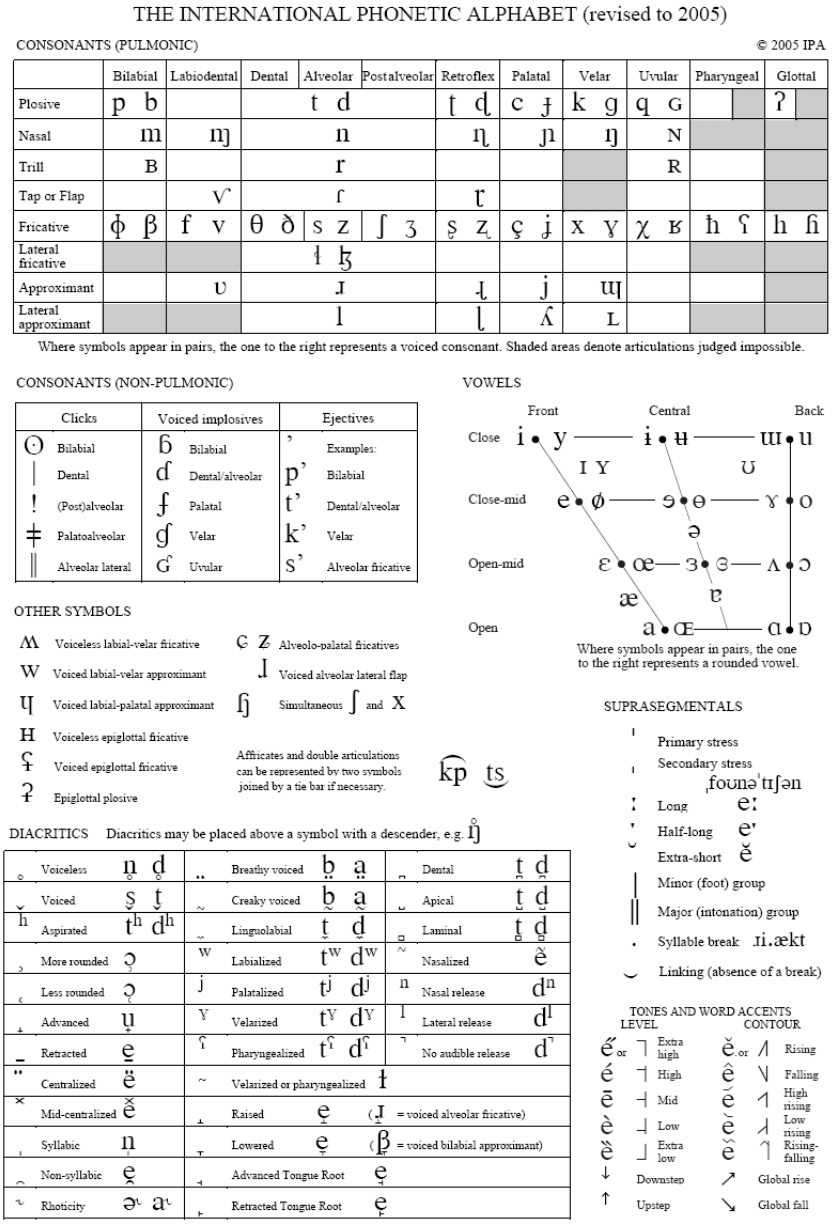


Abbildung 3.2: Das "International Phonetic Alphabet" (aus "Handbook of the International Phonetic Association", 1999)

tronisch verarbeitbar darzustellen. X-Sampa ist im Gegensatz zu seinem Vorgänger Sampa allerdings sprachunabhängig konzipiert (siehe Wells 1995).

Bei X-Sampa handelt es sich um eine eins-zu-eins Umschreibung der IPA-Symbole in Zeichenkombinationen des 128er ASCII-Satzes. Um auf diese Weise alle im IPA enthaltenen Zeichen darstellen zu können, werden für einige Symbole zwei oder drei ASCII-Zeichen benötigt. Abbildung 3.3 zeigt einige IPA-Symbole und ihre Entsprechungen in X-Sampa.

IPA	X-Sampa	Beschreibung
ɓ	b_<	voiced bilabial implosive
j	j\	voiced palatal fricative
ɳ	n`	retroflex nasal
ɸ	p\	voiceless bilabial fricative

Abbildung 3.3: Einige IPA-Symbole und ihre X-Sampa Entsprechungen

Grammatikalische Strukturen

Neben lexikalischen und phonetischen Variationen können sich Dialekte auch im Bereich der Syntax voneinander unterscheiden:

- weil er schon nicht mehr reden können hat
- weil er schon nicht mehr reden hat können
- weil er schon nicht mehr hat reden können
- weil er schon nicht mehr hat können reden

Dieses Beispiel aus Weiss (2003) verdeutlicht, dass insbesondere die flexiblen Wortstellungskombinationen des Deutschen auch eine Fülle an dialektalen Ausdifferenzierungen erlauben.

Die systematische Analyse von dialektalen Syntaxstrukturen kann noch nicht auf eine lange Tradition zurückblicken. Bis jetzt sind erst für wenige Sprachen syntaktische Dialektdaten in Form von Sprachatlanten erhoben worden, so zum Beispiel für das Holländische Barbiers u. a. (2005) und das Schweizerdeutsche Glaser u. Bucheli Berger (2000). Zu dialektometrischen Analysen der holländischen Syntax, siehe auch Spruit (2006).

3.2 Sprachatlanten und Dialektwörterbücher

Unabhängig von ihrer Erhebungsweise und Art können Dialektdaten nach ihrer Verschriftlichung in Form von *Dialektwörterbüchern* oder *Sprachatlanten* festgehalten werden. Hierzu werden die Daten der Fragebögen einer Dialekterhebung entweder in Wörterbuchform zusammengefasst oder auf topographischen Karten markiert. Dementsprechend sind Sprachatlanten eine erste geographische Visualisierung von Dialektdaten.

Als Grundlage für Sprachatlanten dienen *stumme Karten*. Dies sind topographische Landkarten ohne intensive Beschriftung: Nur grobe geographische Merkmale wie z.B. Flüsse, Gebirge oder Staatsgrenzen sind verzeichnet. Auf jeweils einer separaten stummen Karte wird nun ein Aspekt der Dialektdaten den geographischen Orten zugeordnet und dargestellt. So lässt sich für jeden dialektalen Aspekt der untersuchten Sprache eine *Arbeitskarte* erstellen, die in ihrer Gesamtheit den Sprachatlas darstellen. Durch die Verwendung unterschiedlicher Symbole oder Farben kann ein erster Eindruck von der Verteilung der dialektalen Variationen gewonnen werden.

Sprachatlanten werden seit den 80er Jahren des 19 Jahrhunderts systematisch erstellt. Heute stehen für die meisten europäischen Sprachen umfangreiche Sprachatlanten zur Verfügung. In Form von gedruckten Sprachatlanten sind Dialektdaten einer weitergehenden elektronischen Verarbeitung nicht zugänglich. Hierfür müssen die Daten erst manuell digitalisiert und in eine für elektronische Weiterverarbeitung geeignete Form gebracht werden.

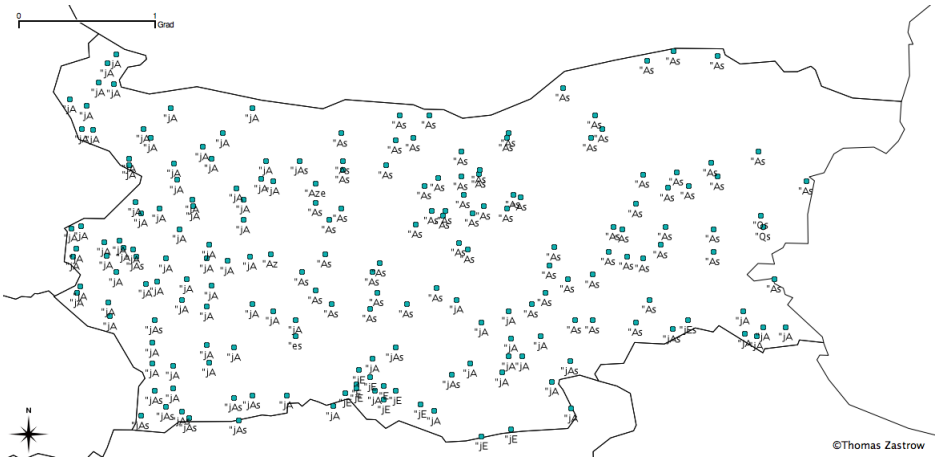


Abbildung 3.4: Die Karte zeigt die Verteilung der unterschiedlichen phonetischen Varianten des bulgarischen Wortes für “ich” im Datensatz Buldialects

3.2.1 Europäische Sprachatlanten

Im Folgenden werden einige europäische Sprachatlanten aufgeführt und näher beschrieben. Dabei kann diese Auflistung aufgrund der Vielzahl der verfügbaren Sprachatlanten nicht vollständig sein: Für die meisten europäischen Sprachen wurden in der Vergangenheit Sprachatlanten erstellt, für einige Sprachen auch mehrere. Hier werden lediglich die Sprachatlanten, an denen dialektometrische Analysen mit Bezug auf die in dieser Arbeit vorgestellten Methoden durchgeführt wurden, aufgeführt. Daneben wurden an skandinavischen, slawischen sowie niederländischen Sprachatlanten dialektometrische Untersuchungen durchgeführt.

Auch wenn die Sprachatlanten zum Teil über 100 Jahre alt sind, die enthaltenen Dialektdaten sind immer noch aktuell und werden auch heute noch eingesetzt. Neben diesen als “historisch” anzusehenden Dialektdaten werden auch heute noch neue, die aktuelle Sprachsituation darstellende Dialektdaten gesammelt[10].

[10]Stellvertretend sei hier das Projekt *Sprachvariationen in Norddeutschland* (SiN) genannt: Im Rahmen von SiN sammeln sechs norddeutsche Universitäten Sprach- und Dialektdaten im norddeutschen Raum.

Romanische Sprachatlanten

Jules Gilliéron und Edmond Edmont sammelten in Frankreich von 1896 an dialektale Daten nach der Interviewmethode. Diese erschienen bis 1910 in insgesamt 13 Bänden als Sprachatlas ("Atlas linguistique de la France").

Zwei Schüler Gilliérons, Karl Jaberg und Jakob Jud, erstellten anschließend einen Sprachatlas des Italienischen. Dieser wurde zwischen 1931 und 1940 publiziert ("Sprach- und Sachatlas Italiens und der Südschweiz").

Der aktuelle "Sprachatlas des Dolomitenladinischen und angrenzender Dialekte" (ALD) wird von Hans Goebl in Salzburg zusammengestellt und herausgegeben (Goebl 1998). Erschienen ist bereits der erste Band, der 217 Orte und 884 Karten umfasst; er enthält hauptsächlich phonetische und morphologische Dialektdaten. Ergänzt wird der ALD durch einen *sprechenden Sprachatlas*: Hierbei handelt es sich um eine multimediale Version des ALD, so dass Benutzer neben den transkribierten Daten auch direkten Zugriff auf die aufgenommenen Audiodaten haben[11].

Es existieren Sprachatlanten auch für weitere romanische Sprachen, so zum Beispiel für das Rumänische (Weigand, 1909).

Slawische Sprachatlanten

Auch in der Slawistik existieren Sprachatlanten für die meisten Sprachen, beispielsweise der bulgarische Sprachatlas (Stojkov, 1964)[12]. Neuere Daten für das Ostserbische und Westbulgarische finden sich in Sobolev (1998).

Der Wenker Atlas

Georg Wenker begann 1876 mit der Sammlung dialektaler Daten in Deutschland nach der Korrespondenz-Methode. Finanziert von der *Preußischen Akademie der Wissenschaften*, wurden bis 1887 ca. 50.000 Lehrer im gesamten deutschsprachigen Raum angeschrieben und gebeten, einen Fragebogen mit

[11] Der "Sprechende Sprachatlas" des ALD-1 im Internet: http://ald.sbg.ac.at/ald/ald-i/index.php?lang=de&id=0013

[12] Eine umfangreiche Liste von slawischen Sprachatlanten findet sich in: http://www.sbg.ac.at/rom/ag/variation/slaw%20sa.pdf

40 bis 49 Sätzen im jeweiligen regionalen Dialekt auszufüllen. Eine Teilmenge von 40 dieser Sätze sind als "Wenker-Sätze" in der Dialektologie bekannt geworden. Obwohl sie teilweise in für heutige Verhältnisse altertümlichen Deutsch verfasst sind[13], finden sie auch heute noch in der Dialektologie Anwendung.

Die entstandene Datenmenge wurde bis 1923 auf insgesamt 1.668 von Hand gezeichneten, farbigen Karten eingetragen. In ihrer Gesamtheit sind die von Wenker und seinen Kollegen Emil Maurmann und Ferdinand Wrede erstellten Karten nie in Buchform publiziert worden. Ab 1984 erschien eine Auswahl der Karten als "Kleine(r) Deutsche Sprachatlas" (Veith u. a., 1984). Der gesamte Datenbestand ist heute online als "Digitaler Wenkeratlas" (DIWA) im Internet zugänglich[14].

3.3 Dialektareale

Ziel der Dialektforschung ist es, im Raum ausdifferenzierte sprachliche Ähnlichkeiten oder Differenzen aufzufinden und sichtbar zu machen. Sprachatlanten stellen Sammlungen von auf Basis topographischer Karten erstellter *Arbeitskarten* dar. Unabhängig von der Art der erhobenen Dialektdaten[15] können diese Arbeitskarten auf zweierlei Weise hin untersucht werden:

- **Qualitativ:** Die *phänomenologische* Betrachtung der Daten ohne weitergehende Analyse kann bereits zusammenhängende Dialektareale erkennen lassen. Hierzu gehört auch das Zusammenfassen der Daten mehrerer Arbeitskarten zu sich ergänzenden oder überlappenden Dialektarealen. Qualitative Methoden werden häufig in manueller Form angewendet.

- **Quantitativ:** Mit quantitativen Methoden können die Daten der Arbeitskarten weitergehend analysiert und neue Karten erstellt werden. Durch vielfältige Methoden und weiterführende Analysemöglichkeiten

[13] Satz 8: *Die Füße tun mir (so sehr) weh, ich glaube, ich habe sie (mir) durchgelaufen.*
[14] http://www.diwa.info/
[15] Allerdings können nicht alle Methoden der Dialektforschung gleichermaßen auf alle Arten von Dialektdaten angewandt werden.

> können auf Basis eines einzigen Datensatzes verschiedene Aufteilungen des untersuchten Gebietes in Dialektareale zu Tage treten. Dabei werden graduell stärkere Grenzen zwischen den Arealen häufiger sichtbar sein als schwächere. Quantitative Methoden werden zumeist unter Zuhilfenahme elektronischer Datenverarbeitungsanlagen automatisiert angewendet.

Sind die Grenzen zwischen den Dialektarealen erarbeitet worden, können sie anschließend mit extralinguistischen Faktoren abgeglichen werden. Hierzu gehören markante Geländeverläufe wie z.B. Flüsse oder Gebirge. Ebenso können politische / soziologische Einflüsse wie auch regionale Gebietsgrenzen oder Wanderungsbewegungen in der Vergangenheit Einfluss auf die Einteilung von Dialektgebieten haben.

Dialektareale bzw. deren Abgrenzungen zueinander können im allgemeinen zwei verschiedene Formen annehmen: Als Isoglosse oder als Dialektkontinuum. Beide Formen schließen sich nicht gegenseitig aus, sie können innerhalb einer Sprache durchaus gemeinsam auftreten. In diesem Fall werden die verschiedenen Dialektkontinua sowohl als auch die sprachlich homogenen Bereiche mittels Isoglossen voneinander abgetrennt.

3.3.1 Isoglossen

Isoglossen (griechisch, zusammengesetzt aus “iso“ (gleich) und “glossa“ (Zunge, Sprache)) repräsentieren den *diskreten* Übergang eines Dialektareals zu einem anderen. Isoglossen entsprechen klaren Grenzen zwischen einzelnen Dialektgebieten und werden auf Karten als Linien dargestellt (Abbildung 3.5, Graphik A). Dabei markieren Isoglossen, die ein Gebiet flächig umfassen, eine dialektale Enklave. Diese sind nicht zu verwechseln mit Gebieten, die aufgrund der Verwendung einer anderen Sprache vom Rest des ansonsten sprachlich einheitlichen Gebietes abgetrennt sind.

Verlaufen mehrere Isoglossen in ähnlicher Weise, so lassen sie sich zu Bündeln zusammenfassen und verstärken somit die Grenze zwischen den Dialekten (Abbildung 3.5, Graphik B). Oft laufen Isoglossen allerdings nicht parallel, sondern entgegengesetzt oder überkreuzen sich (beispielsweise im “Rheinischen Fächer”: Wintgens 1982, S. 37 und Bußmann 2002, S. 163 ff.). In

diesem Fall ist die eindeutige Bestimmung der Grenze zwischen den Dialekten mittels Isoglossen schwierig bzw. nicht möglich (Abbildung 3.5, Graphik C). Abbildung 3.7 zeigt Isoglossen im Schwäbischen, die nicht nur parallel verlaufen, sondern sich zum Teil auch überlappen oder kreuzen.

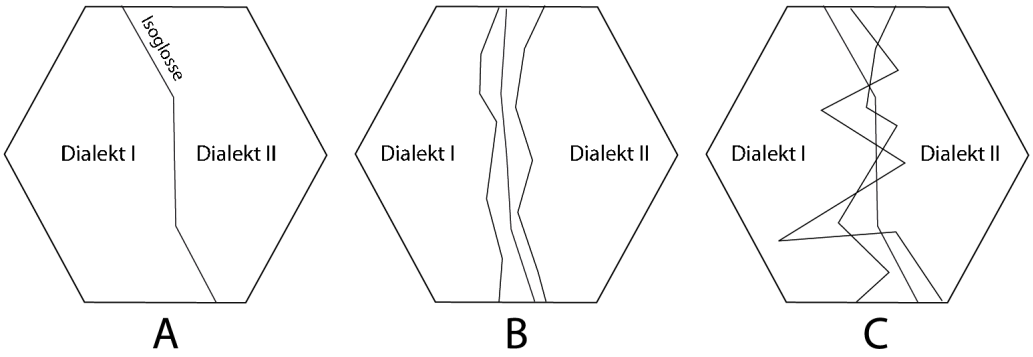


Abbildung 3.5: Schematische Beispiele für den möglichen Verlauf von Isoglossen: Abbildung A zeigt eine einfache Isoglosse, in B sind mehrere parallel verlaufende Isoglossen gebündelt und in C verlaufen Isoglossen entgegengesetzt bzw. überschneiden sich

3.3.2 Dialektkontinuum

Im Gegensatz zu den diskreten Isoglossen verläuft der Übergang von einem Dialekt zum nächsten im Dialektkontinuum *stetig* und nicht abrupt. Von Meßpunkt zu Meßpunkt lassen sich nur kleine Änderungen feststellen, die erst kumuliert über mehrere Meßpunkte einen eindeutigen Wechsel zwischen den Dialekten erkennen lassen. Mit der geographischen Entfernung innerhalb eines Dialektkontinuums sinkt die Ähnlichkeit zwischen den betrachteten Dialekten. Durch die Bündelung von ähnlich verlaufenden Isoglossen lassen sich Dialektkontinua ansatzweise modellieren.

Dialektdaten sind in der Regel an geographisch fixen Punkten erhoben worden. Diese Meßpunkte an sich stellen somit eine diskrete Einheit dar. Eine durchgehende Struktur ergibt sich durch graduelle Änderungen von Punkt zu Punkt, die mit einer geeigneten Methode bestimmt werden müssen. Vor-

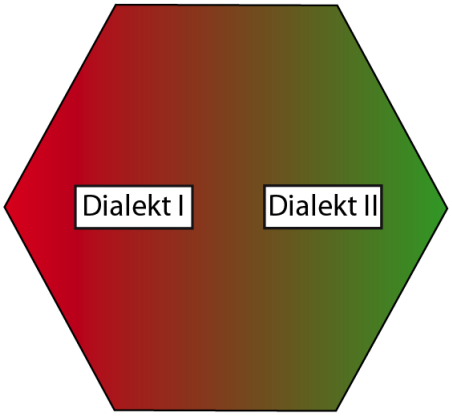


Abbildung 3.6: Dialektkontinuum

aussetzung für eine kontinuierliche Dialektstruktur sind ausreichend dicht gesetzte Meßpunkte.

Ein Dialektkontinuum findet sich im Norden Deutschlands und der Niederlande. Hier sind die Übergänge vom Niederdeutschen zum Niederländischen fließend und können nicht klar voneinander getrennt werden.

Graphisch lassen sich Dialektkontinua als Schraffur oder durch die Verwendung ähnlicher Farben von Punkt zu Punkt (Farbverlauf) auf den Arbeitskarten darstellen (Abbildung 3.6).

ARNO RUOFF: Alltagstexte
(Idiomatica Band 10 und 11)

Karte 2
Hauptsprachgrenzen im Untersuchungsgebiet

Planquadratnetz des Deutschen Spracharchivs
Graphik: Elke Schwedt

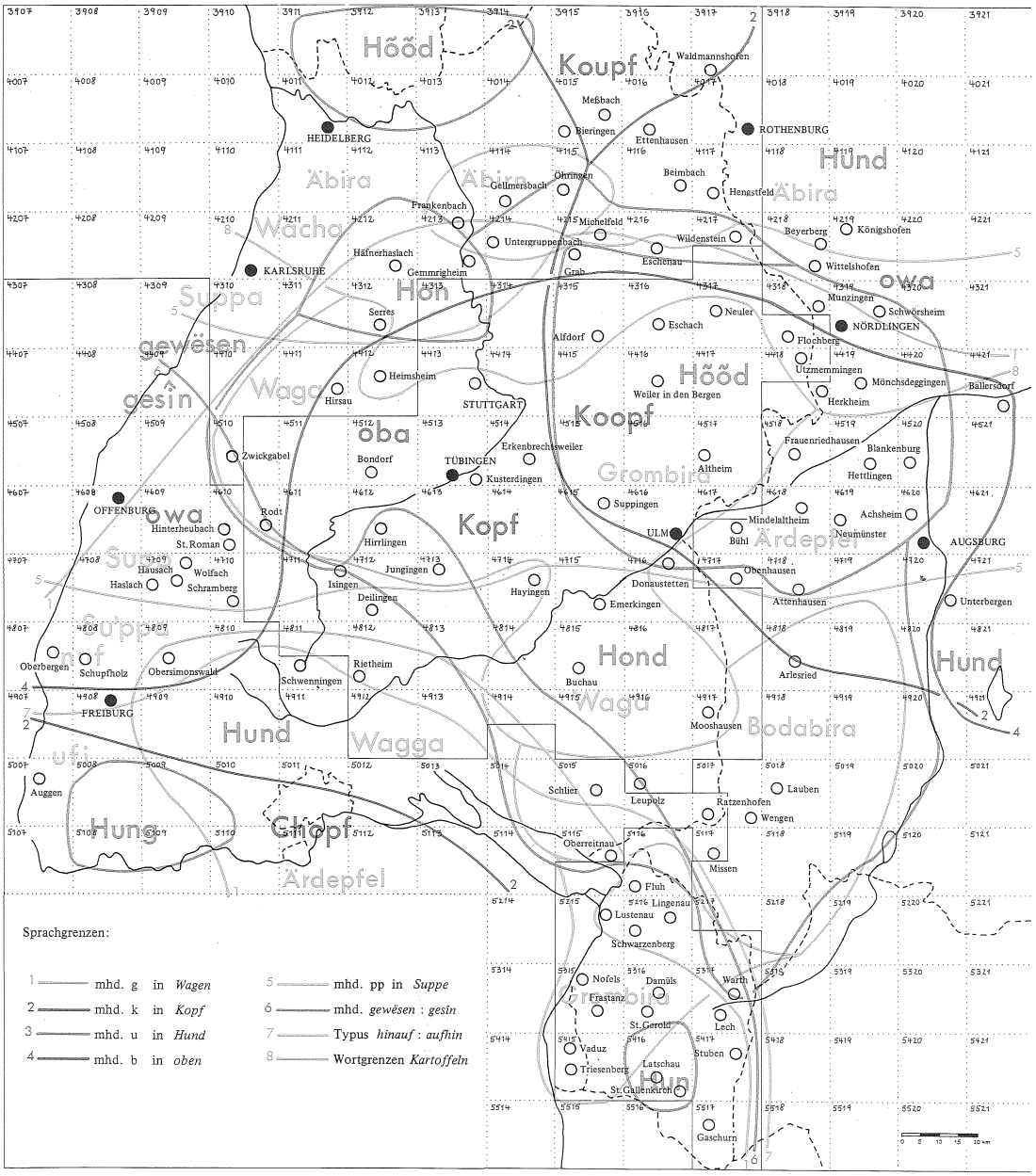


Abbildung 3.7: Isoglossen im Schwäbisch-Alemannischen Raum, Karte im Anhang zu Ruoff (1984)

Kapitel 4

Dialektometrie

In der Dialektometrie werden die zumeist in Sprachatlanten oder Dialektwörterbüchern gesammelten Dialektdaten mit Hilfe quantitativer Methoden (Statistik, Informationstheorie, etc.) unter Zuhilfenahme elektronischer Datenverarbeitungsanlagen und Verfahren analysiert. Ziel ist es, die sprachlichen Strukturen zwischen den einzelnen Dialekten einer Sprache sichtbar zu machen (Bußmann, 2002, S. 165). Diese sprachlichen Strukturen können Eigenschaften auf verschiedenen linguistischen Ebenen (Phonetik, Lexik etc.) sein (Goebl, 2004, S. 249)[1].

Um in der Dialektometrie verwendet werden zu können, müssen die zugrundeliegenden Dialektdaten einem hohen Maß an Formalität genügen. Dies bedeutet konkret, dass für alle zu untersuchenden Dialekte Listen mit denselben Lexemen *und* denselben morphologischen Eigenschaften zur Verfügung stehen müssen[2]. Diese Voraussetzung ist für viele momentan verfügbaren Dialektdatensätze nicht gegeben. Vor allem mittels Interviewmethode erhobene Dialektdaten enthalten über die ganze Bandbreite der untersuchten Orte häufig nicht genügend Lexeme gleicher Ausprägung.

In der Dialektometrie wird auf die *a priori* Einbeziehung extralinguistischer Informationen verzichtet. Hierzu gehören topographische Landschaftselemente wie Gebirge, Flüsse sowohl als auch soziologische Kriterien wie politische Gebietsstrukturen oder Gegensätze zwischen urbanen und ruralen

[1] Die meisten der in dieser Arbeit vorgestellten Methoden arbeiten auf der Basis phonetischer Daten.

[2] Für höhere Strukturanalysen, bspws. im Bereich der Syntax, gilt dies analog.

Arealen. Im Idealfall passen die so gefundenen dialektometrischen Strukturen mit den gegebenen extralinguistischen Strukturen überein. Auch lassen sich diachrone Veränderungen in einem Sprachgebiet (Wanderungsbewegungen etc.) mit dialektometrischen Methoden nachvollziehen (siehe hierzu unter anderem Goebl 2004, S. 270 und Alewijnse u. a. 2007).

Abbildung 4.1 zeigt den in vier Hauptteile[3] gegliederten dialektometrischen Prozess. Die erste Säule, entsprechend dem ersten Hauptteil, umfasst Akquisition, Transkription und Aufbereitung der Dialektdaten. Sie ist weitgehend identisch zu dem weiter oben bereits besprochenen Bereich der Dialektologie. Bei der Erfassung der Daten ist auf den benötigten hohen Formalitätsgrad zu achten. Die erhobenen Dialektdaten müssen anschließend in einer maschinenlesbaren Form transkribiert und digitalisiert werden. In der zweiten Säule werden anschließend mathematische, statistische oder ähnliche Methoden auf die Daten angewendet. Die so ermittelten, dialektspezifischen Kennzahlen können in Form einer *Ähnlichkeits-* oder *Distanzmatrix* dargestellt werden. Die erstellten Matrizen stellen die Ergebnisse der verschiedenen dialektometrischen Ansätze dar. Sie dienen als Grundlage für weitere Analysen. In der dritten Säule werden die so gewonnenen Matrizen mit Verfahren beispielsweise aus dem Bereich des *Machine Learning* (Clustering, Multi Dimensional Scaling etc.) weitergehend analysiert. Ziel ist es hier, in den Daten inhärente Strukturen aufzufinden.

Die vierte Säule umfasst die graphische Visualisierung der in Säule drei gewonnenen Datenstrukturen. Sind mehrere Matrizen erstellt worden, so können diese wiederum mit statistischen Mitteln miteinander verglichen werden (Korrelationsanalyse). Visualisierungen unterschiedlicher Matrizen können visuell miteinander verglichen werden.

Die Ergebnisse werden abhängig von der angewandten Methode, dem Analyseverfahren und der verwendeten Visualisierungstechnik voneinander abweichen und verschiedene Dialektareale zu Tage fördern. Einige dieser Gebiete bzw. die Begrenzungen zwischen ihnen werden sich dabei als stabiler als andere erweisen und immer wieder auftauchen. Wiederum andere

[3] Es sind auch feinere Gliederungen möglich. Goebl wendet in Goebl (2004) eine sechsteilige Gliederung an.

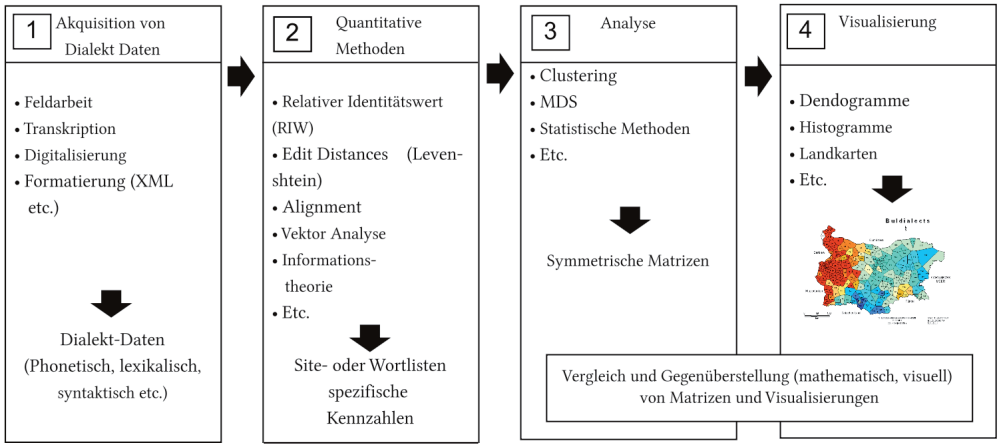


Abbildung 4.1: Vierstufiger Ablauf dialektometrischer Untersuchungen

Strukturen werden nur in wenigen oder vielleicht sogar nur in einer Kombination aus Methode, Analyseverfahren und Visualisierung sichtbar sein. Ziel dialektometrischer Untersuchungen ist es nun, diese immer wiederkehrenden Gebietsgrenzen aufzufinden und zu interpretieren, wobei letzteres auch unter Einbeziehung extralinguistischer Faktoren geschehen kann. Aber auch nur vereinzelt auftretende Dialektareale können von linguistischer Bedeutung sein: Beispielsweise lässt sich so das exponierte Auftreten einer bestimmten linguistischen Eigenschaft in einer Region auffinden (extrahierende Methoden).

Entwicklung der Dialektometrie

Jean Séguy begann in den 1970er Jahren, die in romanischen Sprachatlanten enthaltenen Dialektdaten mit quantitativen Methoden zu analysieren (Seguy 1971 und Seguy 1973). Fortgeführt und wesentlich erweitert wurden seine Ansätze durch das Institut für Romanistik an der Universität Salzburg (Hans Goebl, Roland Bauer). Als mathematische Methode wird hier der "Relative Identitätswert" (RIW_{jk}) und hiervon abgeleitete Varianten verwendet[4]. Zur Analyse wird die von Edgar Haimerl entwickelte Software

[4]Siehe Goebl 2007b, S. 169, sowie online
http://ald.sbg.ac.at/dm/germ/Theorie/Aehnlichkeitsmasse.htm

VDM ("Visual DialectoMetry", siehe Kapitel 9) verwendet. In Salzburg wurden Arbeiten im Bereich der Dialektometrie hauptsächlich an den romanischen Sprachatlanten Italiens (Goebl 2007a) und Frankreichs (Goebl 2004) durchgeführt.

Als einer der ersten wendete Brett Kessler 1995 *Edit Distance Algorithmen* auf irische Dialektdaten an. Es folgten Untersuchungen von John Nerbonne, Wilbert Heeringa und Charlotte Gooskens, die Edit Distance Algorithmen (hauptsächlich den Levenshtein Algorithmus) erweiterten und auf eine Vielzahl weiterer Sprachen anwandten (unter anderem Niederländisch, US-amerikanisch, Bulgarisch, Norwegisch, Gabonesisch und Deutsch). Die L04 Software von Peter Kleiweg implementiert Edit Distance Algorithmen, RIW sowie weitere Analysemethoden der Dialektometrie.

Die verwendeten Methoden unterscheiden sich in ihrer Herangehensweise bzw. den verfügbaren Daten in allen vier Säulen der obigen Abbildung. Allerdings existieren auch Berührungspunkte: So kommt z.B. das Clusteringverfahren im Bereich der Analyse überall zur Anwendung.

An den oben genannten europäischen bzw. afrikanischen Sprachen konnten die etablierten Methoden der Dialektometrie erfolgreich angewandt werden. In dieser Arbeit sollen nun dem Ziel des Projekt Buldialects entsprechend (Hinrichs u. a. 2005) dialektometrische Methoden auf den Kontext der südslawischen Sprachen in Form des Bulgarischen angewandt werden.

4.1 Terminologie

In der Dialektometrie werden unterschiedliche Terminologien benutzt. In dieser Arbeit sollen, hauptsächlich zurückgehend auf die in Groningen verwendeten Fachbegriffe, folgende Begriffe verwendet werden:

- **Meßpunkt:** Bei einem Meßpunkt (in englischen Publikationen häufig **Site**) handelt es sich um einen geographisch definierten Punkt, an dem Dialektdaten erhoben worden sind. Meistens ist ein Meßpunkt identisch mit einer Stadt, einem Dorf oder einer anderen menschlichen Ansiedlung. Durch den punktuellen Charakter der Meßpunkte wird die demographische Größe des Meßpunktes bzw. der umgebenden Ortschaft

	Wort 1	Wort 2	Wort 3
Messpunkt A	Variante A1	Variante A2	Variante A3
Messpunkt B	Variante B1	Variante B2	Variante B3
Messpunkt C	Variante C1	Variante C2	Variante C3

Tabelle 4.1: Einteilung der Terminologie

nicht berücksichtigt.

- **Wort:** Die Dialektdaten des Buldialect Datensatzes wurden in Form einzelner, gebräuchlicher Wörter erhoben. Wurden die Daten in Form der Interviewmethode erhoben, muss eine entsprechende Wortliste aus der Schnittmenge der erhobenen Daten erstellt werden.

- **Variante:** Dialektale Ausprägung eines Wortes an einem bestimmten Meßpunkt.

- **Element:** Ein Element ist der kleinste Bestandteil einer Variante. Liegen die Daten in IPA transkribiert vor, ist ein Element im Normalfall identisch mit einem IPA- bzw. X-Sampa Code, bei Untersuchungen mit *N-Grammen* werden zwei oder mehr X-Sampa Codes zu einem Element zusammengefasst.

4.2 Dialektometrische Datenstrukturen

Dialektometrische Methoden quantifizieren Unterschiede zwischen den einzelnen Dialekten einer Sprache. Zu diesem Zweck wird die jeweilige Methode auf die Daten angewandt und messpunktspezifische *Kennzahlen* errechnet. Diese abstrahieren die dialektalen Eigenschaften eines Messpunktes hin zu einem einzigen metrischen Wert, der den Dialekt des entsprechenden Messpunktes charakterisiert. In der Erstellung dieser Kennzahlen unterscheiden sich die jeweiligen dialektometrischen Methoden:

- **Individuelle Kennzahlen:** Hier wird für jeden Messpunkt ein Wert *unabhängig* von anderen Messpunkten oder dem gesamten Datensatz errechnet. Für sich alleine genommen hat ein solcher Wert faktisch keine Bedeutung: Erst die Gegenüberstellung mit anderen messpunktspezifischen Kennzahlen ergibt ein Gesamtbild der miteinander verglichenen Dialekte (Vektoranalyse, einige informationstheoretische Methoden).

- **Paarweise Kennzahlen:** Dialektale Eigenschaften jeweils zweier Messpunkte werden einander gegenübergestellt. Anschließend können die für jedes Feature individuell gewonnenen Kennzahlen miteinander zu einer messpunktspezifischen Kennzahl aggregiert werden (RIW, Edit Distance, Alignment Algorithmen, auf Gold Standard beruhende Methoden).

- **Datensatzbasierte Kennzahlen:** Unter Einbeziehung der Daten aller Messpunkte werden messpunktspezfische Kennwerte ermittelt (Information).

Unabhängig von der angewandten Methode ist die Art des Ergebnisses bei allen Methoden gleich: Messpunktspezifische Kennzahlen, deren Relationen zueinander den Grad von Ähnlichkeit oder Unähnlichkeit zwischen den einzelnen Dialekten angeben. Diese Kennzahlen können sich in beliebigen Zahlenräumen bewegen. Die Relation zwischen den beiden Messpunkten S_1 und S_2 ist der Betrag zwischen den zugehörigen Kennzahlen k_1 und k_2 (Formel 4.1).

$$D(S_1, S_2) = |k_1 - k_2| \tag{4.1}$$

Diese Relationen, bezogen auf die Gesamtheit aller untersuchten Messpunkte, können in Form einer *Ähnlichkeitsmatrix* dargestellt werden. In einer zweidimensionalen Matrix werden die *prozentualen* Ähnlichkeiten von Messpunkt zu Messpunkt angegeben. Dabei entspricht die Ähnlichkeit eines Messpunktes zu sich selbst immer 100% (oder, abhängig von der verwendeten Software, 1). Die Ähnlichkeiten zwischen zwei Messpunkten sind symmetrisch, so dass die Ähnlichkeit von Messpunkt 1 zu Messpunkt 2 dieselbe ist wie von Messpunkt 2 zu Messpunkt 1. Dies führt dazu, dass die

	S1	S2	S3	S4		S1	S2	S3	S4
S1	0	0,3	0,8	1,8		100	91	80	64
S2		0	0,5	1,5			100	87	70
S3			0	1,0				100	80
S4				0					100

Tabelle 4.2: Ähnlichkeitsmatrizen, links mit absoluten Werten, rechts in Prozentangaben der Messpunkte S1 bis S4

gesamte Matrix symmetrisch ist und an der Mitteldiagonalen (die jeweils die Ähnlichkeiten eines Messpunktes zu sich selbst, also 100%, enthält) gespiegelt werden kann.

Beispiel: Für die vier Messpunkte S1 bis S4 wurden die Kennzahlen <3.2, 3.5, 4.0, 5.0> ermittelt. Tabelle 4.2 zeigt die zugehörige Ähnlichkeitsmatrix auf der linken Seite in absoluten Werten und auf der rechten Seite in Prozentangaben. Die unteren, linken Hälften der Matrizen sind jeweils symmetrisch und enthalten die gleichen Werte wie die oberen Hälften (in der Tabelle 4.2 wurden die symmetrischen Teile der Matrizen leer dargestellt).

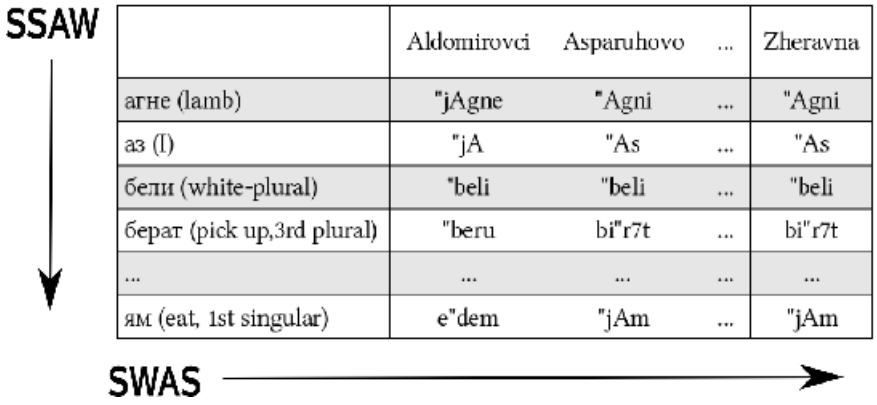
Das logische Gegenteil zur Ähnlichkeitsmatrix ist die *Distanzmatrix*: Hier werden nicht die Ähnlichkeiten, sondern die Unähnlichkeiten bzw. Distanzen zwischen Dialekten festgehalten. Distanz- und Ähnlichkeitsmatrix enthalten jeweils die gleichen Relationen zwischen den einzelnen Messpunkten, nur jeweils in entgegengesetzten Richtungen. Die Salzburger VDM-Software arbeitet beispielsweise mit Ähnlichkeitsmatrizen, wohingegen die Groninger L04-Software Distanzmatrizen erwartet.

Für einen Datensatz mit n Messpunkten ergibt sich eine Matrix mit

$$\frac{n^2 - n}{2} \tag{4.2}$$

einzelnen Werten, ohne die Mitteldiagonale, die die Relationen der Dialekte zu sich selber enthält. Im Falle des phonetischen Teils des

Buldialects-Datensatzes (197 Messpunkte) sind dies 19306 einzelne Werte.



	Aldomirovci	Asparuhovo	...	Zheravna
агне (lamb)	"jAgne	"Agni	...	"Agni
аз (I)	"jA	"As	...	"As
бели (white-plural)	"beli	"beli	...	"beli
берат (pick up,3rd plural)	"beru	bi"r7t	...	bi"r7t
...	...	...	...	...
ям (eat, 1st singular)	e"dem	"jAm	...	"jAm

Abbildung 4.2: Untersuchungsrichtungen SSAW und SWAS. Die Matrix stellt einen Ausschnitt aus dem Buldialects Datensatz dar

Eine symmetrische Matrix kann nun weitergehend in zwei Richtungen analysiert werden (Abbildung 4.2):

- **SSAW:** *Single Site, All Words* (Einzelner Meßpunkt, alle Wörter) - Hierbei handelt es sich um die Untersuchungsrichtung, bei der alle Varianten eines Messpunktes zusammengefasst untersucht werden. Dies ist die hauptsächliche Untersuchungsrichtung sowohl in der Vektoranalyse als auch in den informationstheoretischen Ansätzen.

- **SWAS:** *Single Word, All Sites* (Einzelnes Wort, alle Meßpunkte) - Hier werden alle dialektalen Varianten eines Wortes untersucht. Edit Distance und Alignment Algorithmen arbeiten in dieser Richtung.

4.3 Methoden der Dialektometrie

Jean Seguy (Seguy, 1971) in den 70ern, vor allem aber Hans Goebl (Goebl, 2007b) in den 80ern des vergangenen Jahrhunderts entwickelten quantitative

Methoden zur Analyse von Dialektdaten. Mittlerweile haben sich neben diesen einige weitere Methoden etablieren können, die grob in zwei Kategorien eingeteilt werden können:

- *Extrahierende Methoden* richten das Augenmerk auf eine oder wenige sprachliche Entitäten innerhalb des Datensatzes. Die zu untersuchende Entität wird aus der Gesamtheit der Daten extrahiert und anschließend isoliert vom Rest der Daten betrachtet. Zu den extrahierenden Methoden gehören die dialektologische Analyse der Daten in Form von Isoglossen bzw. Bündeln von Isoglossen sowie die weiter unten beschriebene Vektoranalyse.

- *Aggregierende Methoden* betrachten den Datensatz als Ganzes: Alle vorhandenen Daten fließen in die Analyse mit ein. Hierzu gehören der Relative Identitätswert und die von ihm abgeleiteten Varianten, die Edit Distance Algorithmen und die informationstheoretischen Ansätze.

Der Vorteil der aggregierenden Methoden besteht darin, dass der gesamte Datensatz als Einheit betrachtet und als solche analysiert wird. Demgegenüber stellt sich bei Anwendung extrahierender Methoden jedesmal die Frage, welche Entitäten für eine Dialektanalyse geeignet sind und welche nicht. Andererseits können die extrahierenden Methoden Strukturen, die nur auf Grund weniger Elemente existieren, besser sichtbar machen und aus der Masse der Daten hervorheben.

Im folgenden sollen zwei Methoden, die für die europäische Dialektometrie konstituierend sind, vorgestellt werden: Der Relative Identitätswert und die Edit Distance Algorithmen. Daran schließt sich ein Kapitel über Alignment-Algorithmen an.

4.3.1 Relativer Identitätswert

Die Methode *Relativer Identitätswert* (RIW) wurde von Hans Goebl bereits in den 80er Jahren des vergangenen Jahrhunderts in die Dialektometrie eingeführt (siehe Goebl 1982, Goebl 2006 und Goebl 2007b). Sie wurde erfolgreich an mehreren Datensätzen, hauptsächlich romanischen Sprachatlanten, angewandt (Goebl 2004).

Der RIW stellt die Relation zwischen gleichen (Koidentität, KOI) und ungleichen Entitäten (Kodifferenz, KOD) zweier Messreihen dar. Auf die Dialektometrie übertragen, bedeutet dies: Wie ist das Verhältnis von gleichen sprachlichen Identitäten zu der Gesamtheit der sprachlichen Identitäten zweier Dialekte? Sind zwei sprachliche Entitäten nicht identisch, wird keine weitere quantitative Unterscheidung durchgeführt: Auf atomarer Ebene ist der RIW somit ein binäres Maß, das Entitäten nach *identisch* oder *nicht identitisch* unterscheidet. Bereits in Goebl (1982), S. 81 ff. schlägt Goebl Verfeinerungen des RIW dahingehend vor, dass einzelne Entitäten gemäß ihrer Frequenz gewichtet werden (*Gewichteter Identitätswert*, GIW).

Es ergeben sich folgende Eigenschaften des RIW:

- **Wertebereich**: Der Wertebereich des RIW liegt zwischen 0 (keine Koidentitäten) und 1 (keine Kodifferenzen) bzw. entsprechend zwischen 0% und 100%. Ein RIW von 100% liegt auch immer dann vor, wenn der RIW eines Messpunktes zu sich selbst bestimmt wird ($RIW_{j,j}$)
- Der RIW ist symmetrisch, so dass $RIW_{j,k} = RIW_{k,j}$

Sind Koidentitäten und Kodifferenzen zweier Messpunkte j und k mit jeweils i Entitäten bestimmt, kann anschließend der RIW zwischen ihnen berechnet werden (Formel 4.3)[5]:

$$RIW_{j,k} = 100 \cdot \frac{\sum_1^i KOI_{j,k}^i}{\sum_1^i KOI_{j,k}^i + \sum_1^i KOD_{j,k}^i} \tag{4.3}$$

Beispiel: Die beiden Messpunkte j und k in Tabelle 4.3 zeigen vier Entitäten, davon drei Koidentitäten und eine Kodifferenz. Als RIW lässt sich ermitteln:

$$RIW_{j,k} = 100 \cdot \frac{3}{3+1} = 75\% \tag{4.4}$$

[5]Bei fehlenden Daten, sogenannten Nullstellen, schlägt Gobel vor, die entsprechende sprachliche Entität überhaupt nicht in die Berechnung einfließen zu lassen.

j	A	B	C	D
k	A	B	C	E

Tabelle 4.3: Zwei Messpunkte j und k mit jeweils 4 Entitäten, davon drei Koidentitäten und 1 Kodifferenz

Wird der RIW von jedem Messpunkt zu jedem anderen Messpunkt berechnet, lassen sich die Ergebnisse in Form einer symmetrischen Matrix darstellen und weitergehend analysieren. Da bei dieser Vorgehensweise alle Daten des Datensatzes miteingerechnet werden, handelt es sich beim RIW um eine *aggregierende* Methode. Im Gegensatz zu den meisten anderen Methoden in der Dialektometrie kann der RIW nicht nur auf phonetische, sondern ebenfalls auf andere Arten dialektaler Daten, bspws. lexikalischer Natur, angewandt werden. Diese hohe Flexibilität wird erkauft durch die eingeschränkte Ausdifferenzierung der sprachlichen Entitäten in der Definition von Koidentität und Kodifferenz.

Die in Salzburg von Edgar Haimerl entwickelte Software VDM enthält Routinen zur Berechnung verschiedener Ähnlichkeitsmaße, unter anderem auch RIW und die gewichtende Variante GIW[6]. Die Groninger Software L04 implementiert ebenfalls den GIW[7].

4.3.2 Edit Distance Algorithmen

Edit Distance Algorithmen[8] berechnen die *Kosten*, die entstehen, wenn eine Zeichenkette in eine andere überführt werden soll. Je größer die berechneten Kosten sind, desto größer ist auch der Unterschied zwischen den beiden Zeichenketten. Einige Algorithmen, wie beispielsweise die Hamming Distance, können nur auf Strings gleicher Länge angewendet werden, andere können

[6] http://www.sbg.ac.at/rom/people/proj/dm/vdm/features.html

[7] http://www.let.rug.nl/kleiweg/L04/Manuals/giw.html

[8] Die meisten gängigen Programmiersprachen enthalten Edit Distance Algorithmen als Funktionen. Besonders umfangreich fällt die Java-Bibliothek SimMetrics (http://www.dcs.shef.ac.uk/s̃am/simmetrics.html) aus: Sie implementiert neben den gängigen Edit Distance Algorithmen eine ganze Reihe weiterer, beispielsweise die L2-Distanz oder die Jaccard-Distanz.

Distanzen auch zwischen Strings unterschiedlicher Länge berechnen.

Der Levenshtein Algorithmus (Levenshtein, 1965) wurde 1965 von Vladimir Levenstein entwickelt. Er stellt eine Erweiterung der Hamming-Distanz dar und berechnet den Abstand zweier Zeichenketten mit Hilfe der drei Operationen[9] *Einfügen*, *Löschen* und *Ersetzen* einzelner Charaktere. Dabei können den drei Operationen verschiedene Kosten zugewiesen werden, üblicherweise bekommen Einfügen und Löschen einen Kostenwert von 1 und das Ersetzen einen Wert von 2. Ersetzen stellt somit faktisch eine Kombination aus Löschen und Einfügen dar. Mit Hilfe der drei genannten Operationen ist es auch möglich, Zeichenketten unterschiedlicher Längen miteinander zu vergleichen. Der Levenshtein-Algorithmus ist symmetrisch, was bedeutet, dass die Kosten der Umwandlung von Zeichenkette A nach Zeichenkette B dieselben sind wie umgekehrt von B nach A.

Dem Paradigma der dynamischen Programmierung folgend, wird die Kostenberechnung in kleinere Teilaufgaben eingeteilt und diese dann nacheinander abgearbeitet. Die Kosten der einzelnen Teilaufgaben werden aufaddiert: Die Summe stellt die kleinstmöglichen Kosten zur Überführung der einen Zeichenkette in die andere dar. Der Prozess lässt sich in Form einer zweidimensionalen Matrix darstellen. Die einzelnen Zellen der Matrix stehen dabei jeweils für eine der zum jeweiligen Zeitpunkt möglichen Operationen. Tabelle 4.4 zeigt den Levenshtein-Algorithmus anhand der beiden Zeichenketten “Tasche” und “Taste”. Gestartet wird in der oberen linken Ecke, die als Startwert in beide Richtungen 0 bekommt und den beiden Zeichenketten vorgeordnet ist. Anschließend wird die Tabelle von oben links nach unten rechts aufgefüllt, wobei immer der kleinste Wert der drei möglichen Operationen in die Summe einfliesst. In der rechten unteren Zelle befindet sich anschließend als Ergebnis der gesuchte Kostenwert (hier: 3).

Die wichtigsten Eigenschaften der ungewichteten Levenshtein-Distanz sind wie folgt:

- Sind beide Zeichenketten identisch, beträgt die Levenshtein-Distanz 0.
- Sie beträgt maximal die Länge der größeren der beiden Zeichenketten.

[9] Die Variante Damerua-Levenshtein-Distanz erweitert die Levenshtein-Distanz um eine vierte Operation, das Vertauschen zweier Charaktere.

```
int LevenshteinDistanz(char s[1..m], char t[1..n])
{
  // leere Matrix deklarieren
  declare int d[0..m, 0..n]

 //Erste Zeile bzw. Spalte füllen
  for i = 0 to m  d[i, 0] = i

  for j = 0 to n d[0, j] = j

 //Über die gesamte Matrix iterieren
  for j = 1 to n
  {
    for i = 1 to m
    {
      if s[i] = t[j] then
        d[i, j] = d[i-1, j-1]
      else
       //Minimum zuweisen
        d[i, j] = minimum (d[i-1, j] + 1,
         d[i, j-1] + 1,   d[i-1, j-1] + 1)
    }
  }
  //Zelle unten rechts enthält das Resultat
  return d[m,n]
}
```

Programm 1: Der Levenshtein-Algorithmus als Pseudo-Code

		T	A	S	C	H	E
	0	1	2	3	4	5	6
T	1	0	1	2	3	4	5
A	2	1	0	1	2	3	4
S	3	2	1	0	1	2	3
T	4	3	2	1	2	3	4
E	5	4	3	2	3	4	3

Tabelle 4.4: Der Levenshtein-Algorithmus berechnet die Kosten, die eine Umwandlung der Zeichenkette “Tasche” in “Taste” beanspruchen würde. Die endgültigen Kosten finden sich in der Zelle rechts unten (hier: 3)

- Sind beide Zeichenketten unterschiedlich lang, beträgt die Levenshtein-Distanz mindestens die Differenz der Länge der beiden Zeichenketten.

Wird bei der Art der Charaktere keine weitere Kostenunterscheidung vorgenommen, wird jede Operation gleich bewertet, unabhängig davon, welche Charaktere beteiligt sind. So beträgt die Levenshtein-Distanz von ABC zu $ABCDE$ 2, ebenso wie die Distanz zu $ABCFG$. Programm 1 zeigt den Levenshtein-Algorithmus als Pseudo-Code realisiert.

In der Dialektometrie können Edit Distance Algorithmen auf phonetische Dialektdaten angewendet werden[10]. Hier bestehen die Zeichenketten aus phonetischen Symbolen, meistens in IPA kodiert. Abhängig von der Art der jeweiligen Symbole können die Kosten der oben genannten Operationen verfeinert angesetzt werden. Die Ersetzung eines Konsonanten durch einen Vokal wäre so beipielsweise “teurer“, als die Ersetzung durch einen anderen Konsonanten. Paarweise werden nun die Überführungskosten für die Wörter zweier Messpunkte berechnet. Anschließend ergeben die einzelnen Kosten aufsummiert die Ähnlichkeit bzw. Distanz zwischen den beiden verglichenen Messpunkten. Wird dies von jedem Messpunkt zu jedem anderen Messpunkt getan, ergibt sich eine symmetrische Distanz-Matrix. Letztere kann anschließend mit weiteren Analysemethoden (Clustering, Multi Dimensional Scaling

[10]Bei lexikalischen Dialektdaten sind die Unterschiede zwischen den Wörtern zu groß und die Edit Distance Algorithmen würden keine nutzbaren Werte mehr liefern.

etc.) untersucht und die Ergebnisse auf Karten eingezeichnet werden. In der hier dargestellten Weise angewandt, *aggregieren* Edit Distance Algorithmen die Unterschiede zwischen den Wörtern aller Dialekte eines Datensatzes.

Mit dem Levenshtein-Algorithmus wurden bereits mehrere dialektometrische Analysen erfolgreich durchgeführt: So zum Beispiel durch Bret Kessler an irischen Gälisch (Kessler, 1995), vor allem aber auch durch John Nerbonne und Wilbert Heeringa. Letztere untersuchten ausführlich das Niederländische (Heeringa, 2004), die Bantu-Sprachen in Gabun (Alewijnse u. a., 2007) und die LAMSAS-Staaten in den USA (Nerbonne, 2005).

4.3.3 Alignment Algorithmen

Alignment Algorithmen richten korrespondierende Teile zweier Zeichenketten aneinander aus. Dies ermöglicht es, beispielsweise die einzelnen Bestandteile von Komposita phonetisch korrekt miteinander zu assoziieren. So würde ein direkter Vergleich mittels Levenshtein-Algorithmus die beiden Komposita *Rückhand* und *Handrücken* als weitestgehend verschieden voneinander ansehen. Erst eine vorbereitende Alinierung der beiden Zeichenketten würde die beiden sich phonetisch entsprechenden Teilketten zusammenfassen.

Alignment Algorithmen wurden unter anderem von Michael A. Covington und George Kondrak entwickelt (siehe hierzu Covington 1996 bzw. Kondrak 2000). Abbildung 4.3 zeigt anhand von Beispielen aus Kondrak (2000) das Verhalten der beiden Algorithmen.

	Covington's alignments							*ALINE's alignments*						
three : trēs	θ	r	i	y				||	θ	r	iy	||	
	t	r	ē	s				||	t	r	ē	||	s
blow : flāre	b	l	-	-	o	w		||	b	l	o	||	w
	f	l	ā	r	e	-		||	f	l	ā	||	re
full : plēnus	f	-	-	-	u	l		||	f	u	l	||	
	p	l	ē	n	u	s		||	p	-	l	||	ēnus
fish : piscis	f	-	-	-	i	š		||	f	i	š	||	
	p	i	s	k	i	s		||	p	i	s	||	kis
I : ego	-	-	a	y				||	ay	||			
	e	g	o	-				||	e	||	go		
tooth : dentis	-	-	-	t	u	w	θ		||	t	uw	θ	||	
	d	e	n	t	i	-	s	den	||	t	i	s	||	

Abbildung 4.3: Beispiele für das Verhalten der Alignment Algorithmen von Covington (links) und Kondrak (ALINE, rechts). Aliniert werden englische Wörter mit ihren lateinischen Entsprechungen. Tabelle aus Kondrak 2000, S. 7

Kapitel 5

Vektoranalyse

In der Dialektometrie kann die Vektoranalyse als extrahierende Methode eingesetzt werden: Mittels Ketten von aufeinanderfolgenden Vektoren werden die Vorkommen des zu extrahierenden Elements identifiziert. Anschließend können die so erstellten Vektorketten miteinander verglichen werden.

In der Mathematik sind *Skalare* Entitäten, die durch einen einzelnen numerischen Wert definiert werden. Hierzu gehören beispielsweise Temperatur- oder Längenangaben. Im Gegensatz zu den Skalaren besitzen *Vektoren* zusätzlich zu ihrem Wert eine Richtung, die im umgebenden Raum beziehungsweise auf der anzeigenden Fläche definiert ist. Ein Vektor ist zweidimensional, wenn seine Richtung Positionsangaben in der flachen (euklidischen) Ebene enthält. Dreidimensionale Vektoren beschreiben zusätzlich die dritte Raumdimension. Vektoren können allerdings auch in höher dimensionalen Bezugssystemen konstruiert werden. Die Länge des Vektors repräsentiert seinen nummerischen Betrag.

Beispiele für Vektoren aus der Physik sind Geschwindigkeiten oder Verformungen[1] (Schwartz u. a., 1960, S. 1ff.).

Ein Vektor mit der Länge 0 heisst *Nullvektor* und kann jede beliebige Richtung annehmen.

In Formel 5.1 wird der *Vektor a* durch das Symbol $\vec{a}$ gekennzeichnet[2]. Definiert wird ein Vektor durch Angabe der Positionsänderungen entlang

[1] Sowohl Geschwindigkeiten als auch Verformungen sind zwei- als auch dreidimensional möglich.

[2] In der Literatur werden teilweise auch andere Schreibweisen verwendet.

der Achsen des jeweiligen Koordinatensystems[3]:

$$\vec{a} = \begin{pmatrix} 4 \\ 3 \end{pmatrix} \tag{5.1}$$

Der Vektor $\vec{a}$ in Formel 5.1 ist zweidimensional, wird also durch zwei Koordinatenänderungen definiert. Bei drei- oder mehrdimensionalen Vektoren muss die Matrix auf der rechten Seite um die entsprechenden Werte ergänzt werden. Da ein Vektor nur durch die Änderungen entlang der jeweiligen Achsen definiert ist, beschreibt er eine unendliche Anzahl an realisierbaren gerichteten Pfeilen im gegebenen Koordinatensystem.

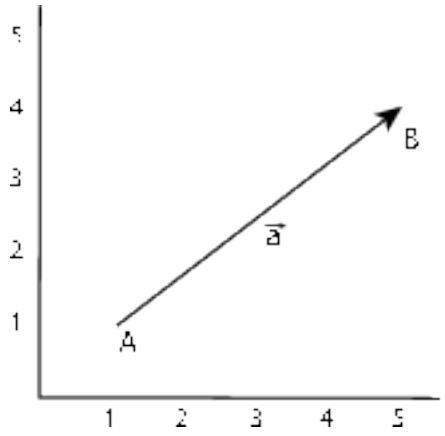


Abbildung 5.1: Vector $\vec{a}$, mit Startpunkt A (1,1) und Endpunkt B (5,4)

Vektoren sind nicht an Anfangs- oder Endpunkte gebunden. Allerdings können Vektoren auch durch Punkte definiert werden. In Formel 5.2 wird der Vektor $\vec{a}$ durch die beiden Punkte A und B definiert (siehe auch Abbildung 5.1):

$$\vec{a} = \overrightarrow{AB} \tag{5.2}$$

Zwei Vektoren $\vec{a}$ und $\vec{b}$ sind identisch, wenn sie dieselbe Länge und Richtung besitzen: Beides ist dann der Fall, wenn die Koordinatenänderungen entlang der jeweiligen Achsen bei $\vec{a}$ und $\vec{b}$ gleich sind. Somit sind ebenfalls alle parallel verlaufenden Vektoren mit derselben Richtung identisch.

[3]Ohne weitere Angabe zum Koordinatensystem ist hier immer das kartesische gemeint.

Die Länge eines Vektors im zweidimensionalen Raum berechnet sich nach Pythagoras:

$$|\vec{a}| = \sqrt{\Delta x^2 + \Delta y^2} \tag{5.3}$$

wobei Δx und Δy den relativen Koordinatenänderungen auf den jeweiligen Achsen entsprechen. Gekennzeichnet wird die Länge eines Vektors durch senkrechte Striche links und rechts. Die Länge eines Vektors stellt wiederum einen Skalar dar.

Der Vektor $\vec{a}$ aus Formel 5.1 hat somit eine Länge von:

$$|\vec{a}| = \sqrt{4^2 + 3^2} = 5 \tag{5.4}$$

Zwei Vektoren lassen sich addieren, indem der Startpunkt des zweiten Vektors an den Endpunkt des ersten Vektors gehängt wird. Zusammengefasst ergeben die beiden addierten Vektoren einen dritten Vektor, der vom Anfangspunkt des ersten zum Endpunkt des zweiten Vektors reicht (Abbildung 5.2). Analog dazu ergibt die Subtraktion zweier Vektoren einen neuen dritten Vektor, der die Endpunkte der beiden voneinander subtrahierten Vektoren miteinander verbindet.

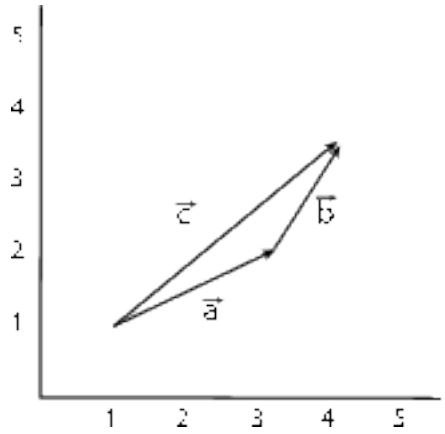


Abbildung 5.2: Die Vektoren $\vec{a}$ und $\vec{b}$ ergeben addiert den Vektor $\vec{c}$

Das *Skalarprodukt* zweier Vektoren ist wie folgt definiert:

$$\vec{a} \cdot \vec{b} = \begin{pmatrix} \Delta ax \\ \Delta ay \end{pmatrix} \cdot \begin{pmatrix} \Delta bx \\ \Delta by \end{pmatrix} = \Delta ax \Delta bx + \Delta ay \Delta by \tag{5.5}$$

Das Skalarprodukt ergibt sich als Summe aus den Produkten der jeweiligen Koordinatenänderungen zweier Vektoren und ist ein Skalar. Zwischen den beiden Vektoren $\vec{a}$ und $\vec{b}$ in Abbildung 5.3 berechnet sich das Skalarprodukt *sp* wie folgt:

$$sp(\vec{a}, \vec{b}) = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3 \end{pmatrix} = 3 + 3 = 6 \tag{5.6}$$

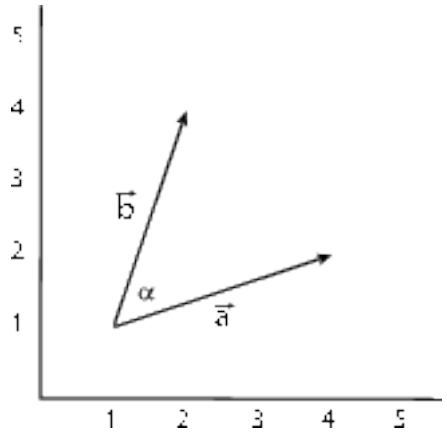


Abbildung 5.3: Skalarprodukt und Winkel zwischen zwei Vektoren

Der Kosinus des Winkels α zwischen zwei Vektoren lässt sich mit Hilfe des Skalarprodukts und der Längen der beiden Vektoren (Formel 5.3) wie folgt berechnen:

$$cos(\alpha) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|} \tag{5.7}$$

Für die beiden Vektoren in Abbildung 5.3 ergibt sich somit ein (gerundeter) Kosinus:

$$cos(\alpha) = \frac{sp(\vec{a}, \vec{b})}{\sqrt{10}^2} = \frac{6}{10} = 0,6 \tag{5.8}$$

Der *Arkuskosinus* als Umkehrfunktion des Kosinus errechnet anschließend den Winkel zwischen den beiden Vektoren, im Falle der beiden Vektoren in Abbildung 5.3 ergibt sich ein Winkel von ca. 53°.

5.1 Vektoren in der Dialektometrie

5.1.1 Vektorketten

Die Vektoranalyse ist ein geometrischer Ansatz, der dazu genutzt werden kann, um dialektale Unterschiede zwischen verschiedenen Datensätzen von Sprachdaten sichtbar zu machen. Bei der in dieser Arbeit beschriebenen Vorgehensweise liegen die Vorteile der Vektoranalyse darin, dass sie a) die *Anzahl* der Vorkommen eines bestimmten Elements und gleichzeitig b) dessen *Positionsänderungen* innerhalb des Datensatzes berücksichtigt. Letzteres unterscheidet den hier vorgestellten vektorbasierten Ansatz von rein frequenzbasierten Analysen. Des Weiteren lassen sich mit der Vektoranalyse Untersuchungen sowohl in der SSAW- als auch in der SWAS-Richtung durchführen. In der hier dargelegten Weise kann die Vektoranalyse nur auf phonetische, nicht aber auf lexikalische Daten angewendet werden.

Eine Eigenschaft der Vektoranalyse besteht darin, dass sie immer nur ein Element gleichzeitig betrachten kann. Somit handelt es sich um eine extrahierende und nicht um eine aggregierende Methode. Allerdings können so elementspezifische Strukturen, die bei aggregierenden Methoden in der Masse der Daten untergehen, hervorgehoben und sichtbar gemacht werden.

Vektoren sind Elemente eines mehrdimensionalen Raumes oder einer zweidimensionalen Ebene. In der Dialektometrie kann die *lineare* Abfolge der einzelnen Varianten innerhalb eines Datensatzes als eine solche zweidimensionale Ebene aufgefasst werden. Dementsprechend spannen sich an der X-Achse entlang die einzelnen Elemente der aktuellen Variante, an der Y-Achse hingegen die Varianten des Messpunktes einer nach der anderen auf.

Abbildung 5.4 zeigt ein Beispiel aus dem Buldialects-Datensatz. Die Daten werden als eine zweidimensional aufgespannte Ebene dargestellt: Der Ur-

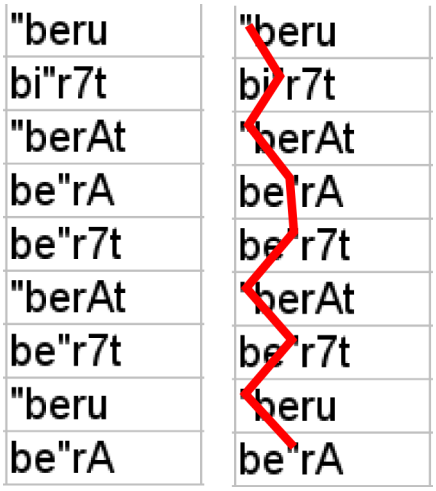


Abbildung 5.4: Vektorketten, ein Beispiel aus dem Buldialects-Datensatz. Die Daten werden als zweidimensional aufgespannte Ebene dargestellt (links). Anschließend kann eine Vektorkette die Vorkommen eines Elements markieren (rechts, hier der Wortakzent)

sprung (0,0) liegt per Definitionem *außerhalb* des Datensatzes, die X-Achse läuft von links nach rechts über die einzelnen Elemente der jeweiligen Variante und die Y-Achse von oben nach unten entlang der einzelnen Varianten des Datensatzes (links)[4]. Anschließend kann eine Vektorkette, die die Vorkommen eines Elements markiert durch die Daten gezogen werden (rechts).

Dabei kann der Dialektdatensatz sowohl in der SSAW- als auch in der SWAS-Richtung aufgespannt sein. Im ersten Fall ergibt sich eine auf den aktuellen Messpunkt, im zweiten eine auf das entsprechende Wort und seine Varianten bezogene Analyse.

Ausgehend vom Ursprung können nun Vektoren, die auf einzelne Elemente zeigen, erstellt werden. Nimmt man den Endpunkt eines solchen Vektors als Startpunkt für einen weiteren Vektor, der auf das nächste Vorkommen des Elements zeigt, so ergibt sich eine Kette von zwei Vektoren. Führt man diese Kette weiter bis zum letzten Vorkommen des Elements, ergibt sich eine *Vektorkette*, die aus derselben Anzahl Vektoren besteht wie das entsprechen-

[4]Diese Richtung der Y-Achse wurde aus praktischen Gründen gewählt. Eine von unten nach oben verlaufende Y-Achse würde dasselbe Ergebnis, nur spiegelverkehrt produzieren.

de Element Vorkommen im Datensatz hat.

Die einzelnen Vektoren einer Kette repräsentieren so einerseits die Anzahl der vorkommenden Elemente und andererseits die *relativen* Positionsänderungen der Elemente zueinander. Dabei gilt, dass:

- *X-Achse:* Ein positiver Wert in der X-Koordinate bedeutet eine Positionsänderung nach rechts, zum Ende des Wortes hin. Bei einem negativen Wert findet die Positionsänderung nach links, zum Anfang einer weiter unten stehenden Variante hin statt. Ist der X-Wert 0, findet keine Positionsänderung statt

- *Y-Achse:* Definitionsgemäß können Positionsänderungen entlang der Y-Achse nur positive Werte oder 0 annehmen. Ein Wert von 1 bedeutet, dass das gesuchte Element auch in der direkt darunter liegenden Variante vorkommt. Ein Wert von 0 bedeutet, dass das Element mehr als einmal in der aktuellen Variante vorkommt. Letzteres gilt nicht im Falle des ersten Vektors, da dieser, vom Ursprung ausgehend, in der Y-Koordinate immer den Wert 0 annimmt, falls die erste Variante das gesuchte Element enthält

Einige besonderen Vektoren bzw. Vektorketten, aus denen sich Aussagen über die Struktur der Daten ziehen lassen:

- Hat der Vektor die Form $X = 1$ und $Y = 0$, dann liegt eine Verdoppelung des Elements vor (beispielweise “f” in “Schiff”).

- Ist der erste (und damit auch letzte) Vektor der Kette der Nullvektor, kommt das gesuchte Element im Datensatz nicht vor

- Haben alle Vektoren einer Kette den X-Wert 0, dann nimmt das gesuchte Element keinerlei Positionsänderungen im Datensatz vor

- In der SWAS Richtung gilt: Die Anzahl der verschiedenen X-Werte entspricht der Anzahl der verschiedenen Varianten in Bezug auf das gesuchte Element

- Wenn die Anzahl der Vektoren in einer Vektorkette gleich der Anzahl der Varianten ist und alle Y-Werte der Vektoren bis auf den ersten den Wert 1 haben, dann kommt das gesuchte Element in jeder Variante genau einmal vor

- Eine Vektorkette hat ihre maximale Länge dann, wenn jede Variante an erster und an letzter Position das gesuchte Element enthält

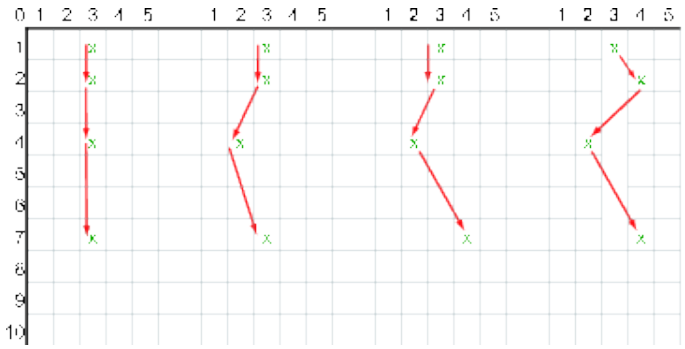


Abbildung 5.5: Beispiel: Vier Vektorketten vc_a, vc_b, vcd_c, vc_d (von links nach rechts)

Abbildung 5.5 zeigt Ausschnitte von vier Vektorketten, die jeweils denselben Startpunkt und die gleiche Anzahl Vektoren, aber jeweils einen anderen Verlauf haben. Dies drückt sich in der Länge der Vektorketten aus, die in den Formeln 5.9 bis 5.12 berechnet werden:

$$|vc_a| = \sqrt{1^2 + 0^2} + \sqrt{2^2 + 0^2} + \sqrt{3^2 + 0^2} = 6 \tag{5.9}$$

$$|vc_b| = \sqrt{1^2 + 1^2} + \sqrt{1^2 + 2^2} + \sqrt{1^2 + 3^2} \approx 6,522 \tag{5.10}$$

$$|vc_c| = \sqrt{1^2 + 1^2} + \sqrt{1^2 + 2^2} + \sqrt{2^2 + 3^2} \approx 6,841 \tag{5.11}$$

$$|vc_d| = \sqrt{1^2 + 1^2} + \sqrt{2^2 + 2^2} + \sqrt{2^2 + 3^2} \approx 7,847 \tag{5.12}$$

Hier wird die Länge der Vektorketten durch Aufsummieren der Längen der einzelnen Vektoren errechnet, was nicht der oben beschriebenen Addition zweier oder mehrerer Vektoren entspricht. Es ergeben sich so individuelle Kennzahlen für den jeweils untersuchten Datensatz, die anschließend miteinander verglichen werden können. Eine rein frequenzbasierte Analyse hätte aufgrund der gleichen Anzahl untersuchter Elemente in allen Datensätzen keine Unterschiede zwischen den Datensätzen aufdecken können.

5.1.2 Auswahl des Elements und Erstellung der Vektorketten

Prinzipiell kann für jedes in einem Datensatz vorkommende Element eine Vektorkette erstellt werden. Aus dialektometrischer Sicht ist dies sinnvoll für Elemente, die

- häufig auftreten. Im Gegensatz dazu können auch sehr seltene Elemente, die nur in wenigen Messpunkten überhaupt vorkommen, auf dialektale Besonderheiten hinweisen. In diesem Fall können allerdings in der SSAW-Richtung nur für diese Messpunkte überhaupt Vektorketten erstellt werden und Messpunkte ohne das entsprechende Element bleiben außen vor.

- häufige Positionsänderungen innerhalb der Varianten vorweisen. Wenn nur selten oder keine Positionsänderungen des Elements vorliegen, dann enthält die Vektorkette ähnlich einer Frequenzanalyse lediglich quantitative Informationen.

Elemente, die die oben genannten Anforderungen erfüllen, lassen sich über Vektorketten in der SWAS-Richtung ermitteln:

1. Erstellung von Vektorketten für jedes auftretende Element in jedem Wort.

2. Aneinanderhängen aller Vektorketten eines Elements zu einer großen Vektorkette.

3. Die Elemente mit den längsten Vektorketten kommen a) am häufigsten vor und / oder haben b) die meisten Positionsänderungen innerhalb der gesamten Datenmenge.

Anschließend können für die so ermittelten Elemente Vektorketten in der SSAW-Richtung erstellt werden. Die Längen der resultierenden Vektorketten repräsentieren nun messpunktspezifische Kennzahlen aus positiven natürlichen Zahlen. Aus diesen lässt sich wiederum durch Bildung der Differenzen eine symmetrische Ähnlichkeits- bzw. Distanzmatrix erstellen.

Die Auswahl der Elemente ist abhängig von der untersuchten Sprache und lässt sich nicht pauschal angeben. Voraussichtlich werden dies aber die linguistisch relevanteren Elemente wie beispielsweise Vokale sein.

5.1.3 Reihenfolge der Varianten

Der Raum, in dem die Vektorketten erstellt werden, wird durch die Varianten aufgespannt: In SSAW-Richtung jeweils für alle Worte eines Messpunktes, in SWAS-Richtung jeweils für alle Varianten eines Wortes. In beiden Fällen darf die Reihenfolge der Varianten von Messpunkt zu Messpunkt bzw. von Wort zu Wort nicht verändert werden.

Die Reihenfolge der Varianten an sich ist beliebig: Wird sie allerdings geändert, so muss dies für alle Datensätze gleichermaßen erfolgen, da die resultierenden Vektorketten sonst nicht mehr vergleichbar wären. Einige Extremfälle sind denkbar: Werden z.B. alle Varianten, die das gesuchte Element enthalten, an den Anfang geschoben, so ergibt sich eine zusammengestauchte Vektorkette. Diese enthält allerdings im Vergleich zu den anderen Vektorketten immer noch alle dialektalen Besonderheiten (Anzahl und Positionsänderungen des Elements).

5.1.4 Interpretation der Vektorketten

Vektorketten können als messpunkt- bzw. wortspezifische *Fingerabdrücke* angesehen werden. Sie sind spezifisch und individuell für den entsprechenden Datensatz, haben alleine für sich genommen aber keine linguistische bzw. dialektale Bedeutung. Werden Vektorketten in der SWAS-Richtung erstellt, so

können die Ergebnisse zur Identifizierung linguistisch relevanter Elemente im Datensatz verwendet werden. In der SSAW-Richtung hingegen werden durch die Vektorketten verschiedene Wörter in Beziehung zueinander gesetzt: Eine linguistisch relevante Aussage lässt sich so nicht direkt treffen. Diese entsteht erst aus dem Vergleich zweier oder mehrerer Vektorketten in SSAW-Richtung miteinander. Da die Varianten in allen miteinander zu vergleichenden Datensätzen stets die gleiche Reihenfolge haben müssen, bedeutet ein Vergleich von zwei oder mehr Vektorketten indirekt einen Vergleich der sich jeweils entsprechenden Varianten. Dieser Vergleich kann, wie oben beschrieben, mittels der absoluten Längen der Vektorketten durchgeführt werden. Je länger eine Kette ist, desto häufiger kommt das untersuchte Element im Datensatz vor bzw. desto häufiger liegen Positionsänderungen innerhalb der Wörter vor. Anstelle der Länge der Vektorketten wäre aber auch die Verwendung der Winkel zwischen den einzelnen Vektoren einer Kette möglich.

Abbildung 5.6: Die Vektoren zweier Messpunkte vergleichen jeweils dieselben Varianten

Abbildung 5.6 zeigt einen Ausschnitt zweier Vektorketten in der SSAW-Richtung für zwei Messpunkte. Hier hat das Element in der Variante des Wortes 2 eine Positionsänderung vorgenommen. Dementsprechend ist der dargestellte Vektor in Messpunkt 2 verschieden zu dem in Messpunkt 1.

Dabei müssen Start- und Endpunkt eines Vektors in verschiedenen Datensätzen nicht zwingend auf die Varianten desselben Wortes zeigen. Fehlt

in einem Datensatz in der entsprechenden Variante das gesuchte Element, so wird die Variante von der Vektorkette übersprungen. Auch dies führt zu einer Veränderung der Vektorkette in ihrer Länge bzw. in der Anzahl ihrer einzelnen Vektoren.

Die hier beschriebene Methode der Vektoranalyse trifft somit eine Aussage über den *Grad der Verschiedenheit* der untersuchten Datensätze zueinander. Dies bezieht sich immer auf ein einzelnes Element im Fokus, die restlichen Elemente werden ausgeblendet.

Kapitel 6

N-Gramm Analysen von Dialektdaten

Als N-Gramm wird eine sequentielle Abfolge von N Elementen bezeichnet. N-Gramm Analysen sind in der Linguistik eine weit verbreitete Analysemethode und werden auf verschiedenen Ebenen der Sprache (Grapheme, Silben, Wörter etc.) durchgeführt. Die Analyse der entstandenen N-Gramme gibt Aufschluß über die möglichen Kombinationen der Elemente: Welche Kombinationen sind häufig, welche treten eher selten auf und welche existieren überhaupt nicht? Beispielsweise tritt in der deutschen Sprache auf Graphem-Ebene ein "q" immer nur in Kombination mit einem darauf folgendem "u" auf.

In der phonetischen Dialektometrie bilden die Phone[1] eines Wortes sequentielle Abfolgen, die mittels N-Gramm-Modellen analysiert werden können. Dabei bilden die Wörter innerhalb eines Datensatzes die Grenze, wortübergreifende Bigramme sollen nicht erstellt werden. Somit hat die hier vorgestellte Variante der N-Gramm Analyse unter anderem einen *wortbasierten* Charakter.

Mittels N-Grammen lassen sich in der Dialektometrie mehrere Fragestellungen untersuchen und beantworten:

- Gibt es Kombinationen aus Phonen, die ausschließlich oder überdurchschnittlich häufig zusammen auftreten?

[1] In den vorliegenden Datensätzen realisiert als IPA- bzw. X-Sampa Codes.

- Lassen sich bestimmte Kombinationen aus Phonen einzelnen Dialekten oder Dialektarealen zuordnen?
- Ist die Gesamtheit der Phon-Kombinationen eines Dialektes charakteristisch für diesen, oder treten die vorkommenden N-Gramme aus Phonen (nahezu) gleichverteilt im gesamten untersuchten Gebiet auf?

Die Untersuchungen in dieser Arbeit beschränken sich auf Untersuchungen von Bigrammen, sprich, sequentiellen Kombinationen aus jeweils zwei X-Sampa-Codes. Dies geschieht aus praktischen Gründen, da die im bulgarischen Datensatz vorkommenden Varianten der verwendeten Wörter größtenteils kurz sind. Kombinationen aus drei oder mehr X-Sampa-Codes fallen deswegen nur in geringer Anzahl an.

Beispiel: Gegeben ist die Zeichenkette *abxyabcz*. Um die in dieser Kette enthaltenen Bigramme zu erhalten, wird, ausgehend von den ersten zwei Elementen, ein zwei Elemente breites Fenster jeweils um ein Element weiter nach rechts bewegt. Dementsprechend enthält die oben genannte Kette die sieben Bigramme *ab, bx, xy, ya, ab, bc, cz*.

6.1 Bigramm-Matrizen

Die wie oben beschriebenen Bigramme können anschließend in einer zweidimensionalen quadratischen Matrix angeordnet werden[2]. Dabei steht per Definitionem die vertikale Dimension für das jeweils erste, die horizontale Dimension für das jeweils zweite Element eines Bigramms. Die Zellen der Matrix nehmen anschließend die absoluten Anzahlen der entsprechenden Bigramme auf (Tabelle 6.1). In einer solchen Matrix steht die Summe aller Zellen in einer *Zeile* für die Gesamtheit aller Bigramme, die mit dem entsprechenden Element auf der vertikalen Achse beginnen. Demgegenüber steht die Summe aller Zeilen einer *Spalte* für die Summe jener Bigramme, die mit dem entsprechenden Element auf der horizontalen Achse enden.

[2]Bei N-Grammen mit $N > 2$ würde sich die Anzahl der Dimensionen entsprechend vergrößern. Gleichzeitig würde, im Vergleich zu den Bigrammen, eine Abnahme der absoluten Anzahl der N-Gramme stattfinden. Dies resultiert in einer generell geringeren Datendichte.

	a	b	c	x	y	z
a		2				
b			1	1		
c						1
x					1	
y	1					
z						

Tabelle 6.1: Bigramme als Matrix

Bei n verschiedenen Elementen über den gesamten Datensatz ergibt sich eine quadratische Matrix mit $n \times n$ Zellen. Da Bigramme nicht symmetrisch sind (ab ist ein anderes Bigramm als ba), ist auch die Matrix im Gegensatz zu den in der Dialektometrie verwendeten Distanz- oder Ähnlichkeitsmatrizen nicht symmetrisch.

In der Dialektometrie lässt sich für jeden Messpunkt eine solche Matrix mit allen in dem jeweiligen Messpunkt vorkommenden Bigrammen berechnen. Damit die Matrizen vergleichbar bleiben, verzeichnet jede Matrix nicht nur die in dem jeweiligen Messpunkt vorkommenden, sondern in beiden Dimensionen alle im gesamten Datensatz vorkommenden X-Sampa-Codes. Diese Anordnung erlaubt so einen direkten Vergleich aller Matrizen bzw. Messpunkte. Zusätzlich lässt sich ein weiterer Wert berechnen: Die *Bigrammdichte* ϱ_{bg} innerhalb einer Matrix. Sie berechnet sich wie folgt:

$$\varrho_{bg} = \frac{Anzahl\ Zellen\ \neq\ 0}{Anzahl\ aller\ Zellen} \tag{6.1}$$

Mit Hilfe der Bigrammdichte lassen sich Messpunkte quantitativ miteinander vergleichen: Je höher die Bigrammdichte, desto mehr Varianz weisen die in dem jeweiligen Messpunkt verwendeten Bigramme auf.

Werden die Werte in der Matrix nicht durch Zahlen, sondern durch Graustufen verschiedener Intensität visualisiert (0: Weiß, maximaler Wert: Schwarz), so ergeben sich für jeden Messpunkt Muster. Diese können, ähnlich zu den Vektorketten in der Vektoranalyse, als individuelle "Fingerabdrücke"

des jeweiligen Messpunktes angesehen werden. Zwei in ihrer Bigrammstruktur ähnliche Messpunkte werden ebenfalls ähnliche "Fingerabdrücke" aufweisen.

Um nun die Matrizen zweier Messpunkte mathematisch miteinander zu vergleichen, wird für jede Zelle in den beiden Matrizen die absolute Differenz gebildet und diese Differenzen dann für die Gesamtheit aller Zellen der beiden Matrizen aufaddiert. Je mehr Bigramme in den beiden miteinander verglichenen Messpunkten ähnlich häufig auftreten, umso kleiner wird die Gesamtdifferenz zwischen den beiden Messpunkten ausfallen. Die Gesamtdifferenz zweier Matrizen ist somit ein positiver Wert mit den Eigenschaften:

- Er beträgt 0, wenn beide Matrizen identisch sind.
- Der Maximalwert beträgt die Summe aller in beiden Matrizen vorkommender Werte. In diesem Fall gibt es kein einziges übereinstimmendes Bigramm.

Berechnet man die Gesamtdifferenzen von jedem Messpunkt zu jedem anderen, so können diese wiederum in einer symmetrischen Matrix mit leerer Mitteldiagonalen angeordnet werden. Diese kann dann mit den herkömmlichen Mitteln der Dialektometrie analysiert bzw. visualisiert werden. Hierbei gilt, je kleiner der jeweilige Wert ist, desto ähnlicher sind sich die beiden Matrizen, es handelt sich also um eine Distanz- und nicht um eine Ähnlichkeits-Matrix. Letztendlich enthält die so entstandene Matrix die Distanzen der Summen aller im untersuchten Datensatz vorkommenden Bigramme. Damit handelt es sich bei dieser Art der Bigramm-Analyse um eine *aggregierende*, sprich den gesamten Datensatz einbeziehende Methode.

Das hier gezeigte Vorgehen ist nur zulässig, wenn die verwendeten Datensätze eine ähnliche Größe in Bezug auf die Anzahl der enthaltenen Bigramme aufweisen. Bei zu großen Abweichungen in dieser Größe würden die errechneten Differenzen lediglich die rein quantitativen- und nicht die qualitativen Verhältnisse der in den beteiligten Messpunkten verwendeten Bigramme aufweisen.

Beispiel: Die in Abbildung 6.1 gezeigten Bigramm-Matrizen enthalten

	a	b
a	1	0
b	0	3

Bigramm-Matrix 1

	a	b
a	2	0
b	0	5

Bigramm-Matrix 2

Abbildung 6.1: Zwei Bigramm-Matrizen mit einer Gesamtdifferenz von 3

in den Zellen *aa* und *bb* Unterschiede, die sich zusammen auf den Wert 3 addieren.

6.2 Analyse einzelner Bigramme

Sollen einzelne Bigramme und nicht die Gesamtheit aller Bigramme analysiert werden, so können die entsprechenden Zellen aus den oben erwähnten Matrizen herauskopiert und zu einer neuen Zeile angeordnet werden. Die zweite Dimension würde hierbei aus den untersuchten Messpunkten bestehen. Alle so entstandenen Zeilen können anschließend in einer neuen, rechteckigen Matrix mit den Dimensionen $n \times m$ angeordnet werden, wobei diesmal n für die Anzahl der verschiedenen Bigramme und m für die Anzahl der Messpunkte steht. Die so entstandene Matrix aggregiert nun den gesamten Datensatz, nicht vorhandene Bigramme müssen deswegen nicht berücksichtigt werden.

Kapitel 7

Informationstheorie

Die Wurzeln des Begriffs *Information* stammen aus dem Lateinischen: *informare* - "Form geben, bilden". Heute wird das Konzept der Information mehrdeutig in den verschiedenensten Bereichen angewendet. Neben den gängigen, alltagssprachlichen Bedeutungen existieren in den Geistes- und Naturwissenschaften unterschiedliche, jeweils fachspezifische Definitionen des Konzepts *Information*. Dies umfasst so unterschiedliche Gebiete wie Philosophie oder Quantenphysik, in denen die Information an sich eine entscheidende Rolle einnimmt und jeweils ganz unterschiedliche Bedeutungen zugewiesen bekommt.

In dieser Arbeit wird der Begriff *Information* im Sinne der Shannon'schen Informationstheorie verwendet (Lyre, 2002, S. 23 ff). Die Information ist in dieser Definition quantifizierbar und erlaubt den direkten Vergleich mehrerer Datenmengen (hier: Dialekte) miteinander. Dies ist nur möglich, wenn sich die miteinander verglichenen Datenmengen jeweils aus Einheiten desselben Typs zusammensetzen (hier: XSampa-Codes). Dies gilt sowohl für die atomaren Einheiten der Daten als auch für alle höherstufigen, die atomaren Einheiten zusammenfassenden Einheiten (beispielsweise ganze Wörter). So können informationstheoretische Werte für jede Einheiteneinteilung der Daten separat berechnet und anschließend miteinander verglichen werden. Das Alphabet A einer Datenmenge ist definiert als die Menge an Zeichen, in der jedes in der Datenmenge vorkommende Zeichen genau einmal auftritt.

Die Anfänge der Informationstheorie gehen auf Claude Elwood Shannon[1] zurück. In den 1940er Jahren im Bereich der Nachrichtentechnik tätig, entwickelte Shannon ein theoretisches Maß zur Quantifizierung der Güte von technischen Informationsübertragungen wie Telefonleitungen, Funksprechgeräten und ähnlichem (Lyre, 2002, S. 23). Im Jahr 1947 veröffentlichte Shannon den Artikel *A Mathematical Theory of Communication*[2], der heute als grundlegende Arbeit der modernen Informationstheorie gilt (Shannon 1947 bzw. Lyre 2002, S. 23). Die Informationstheorie hat Berührungspunkte zu vielen anderen geistes- und naturwissenschaftlichen Feldern, beispielsweise der Informatik, Physik und Statistik.

Dass die Informationstheorie auch zur generellen Beschreibung natürlichsprachlicher Phänomene herangezogen werden kann, bewies Shannon selbst bereits wenige Jahre später. In seinem Aufsatz *Prediction and Entropy of Printed English* analysierte Shannon die Entropie der Buchstabenverwendung in der englischen Sprache (Shannon, 1951). Kurz darauf wandte Karl Küpfmüller die verwendete Methodik auch auf die deutsche Sprache an (Küpfmüller, 1954).

Eine Alternative zur Shannon'schen Informationstheorie stellt die *Kolmogorow-Komplexität* dar. Hier wird die Komplexität einer Nachricht beschrieben durch die kleinste mögliche Datenmenge, welche die entsprechende Nachricht darstellen kann. Die Kolmogorow-Komplexität findet heute hauptsächlich Anwendung in Algorithmen zur Datenkompression (Cover u. Thomas, 2006, S. 466 ff.).

Wie im folgenden näher erläutert, beruhen alle informationstheoretischen Maße auf der Wahrscheinlichkeit bzw. der Frequenz der unterteilenden Elemente einer Nachricht bzw. Datenmenge. Von einfachen Frequenzanalysen unterscheiden sich informationstheoretische Maße hauptsächlich dadurch, dass die Frequenzen der individuellen Elemente mittels verschiedener Algorithmen *aggregiert* werden können und die errechneten Kennzahlen den gesamten (Teil-) Datensatz beschreiben. Somit sind die informationstheore-

[1] Shannon baute unter anderem auf frühere Arbeiten von R. V. L. Hartley auf, siehe beispielsweise Hartley (1927).

[2] Im genannten Artikel (Shannon, 1947) beruft Shannon sich wiederum auf vorangegangene Arbeiten, u.a. werden H. Nyquist und R.V.L. Hartley zitiert.

tischen Methoden generell aggregierender Natur und beziehen sich immer auf den Datensatz als Ganzes und nicht nur auf einen Teil davon.

Im folgenden werden die Grundlagen der Shannon'schen Informationstheorie dargestellt. Anschließend werden einige informationstheoretische Verfahren für die Anwendung in der Dialektometrie adaptiert.

7.1 Der Begriff der Information

Wird eine Nachricht über einen technischen Kommunikationskanal wie bspws. eine Telefonleitung übertragen, dann setzt sie sich einerseits aus der gewollt übertragenen *Information* und andererseits dem nicht gewollten *(Daten-) Rauschen* zusammen. Um die in einer Nachricht enthaltene Information quantifizieren zu können, muss diese in kleinere *Einheiten* aufteilbar sein. Anschließend kann der Informationsgehalt eines jeden Zeichens in der gewählten Einheit berechnet werden[3]:

$$I(z) = -log_2 p(z) \tag{7.1}$$

Wobei $p(z)$ die Wahrscheinlichkeit des Auftretens des Zeichens z in der gesamten Nachricht bezeichnet.

Durch den Logarithmus zur Basis 2 wird die Information im Binärsystem gemessen. Hier können auch andere Basen verwendet werden, das Binärsystem ist allerdings weit verbreitet. Die Einheit der Information im Binärsystem ist *Shannon* (Sh)[4]. Wird die Information beispielsweise mit dem Logarithmus zur Basis 10 berechnet, heißt die Einheit Hartley. In dieser Arbeit wird einheitlich das Binärsystem und damit auch die Einheit Shannon verwendet.

Der Informationsgehalt einer Datenmenge in Shannon lässt sich veranschaulichen durch die Anzahl der "Ja/Nein"-Fragen, die ein Fragender durchschnittlich stellen muss, bis er zu dem korrekten Ergebnis kommt.

[3] Zur Ableitung der Informationsformel, siehe Klimant u. a. (2006) mit direktem Bezug auf Hartley (1927).

[4] Häufig wird auch der Begriff Bit für die im Binärsystem gemessene Information verwendet. Der ISO Standard 2382-16 legt sich allerdings auf die Bezeichnung Shannon fest. Zu beachten ist, dass die Einheit der Information kontinuierlich ist, im Gegensatz zum diskreten, in der Computertechnik verwendeten Bit.

Beispielsweise lässt sich das Ergebnis eines Münzwurfs mit einer einzigen "Ja/Nein"-Frage ermitteln: *War es Kopf?*. Ein Münzwurf enthält dementsprechend die Information von einem Shannon. Analog ergibt sich bei einem sechsseitigen Würfel die Information von ca. 2,58 Sh pro Wurf: Es sind durchschnittlich 2,58 Fragen notwendig, bis ein Ratender die korrekte gewürfelte Zahl ermittelt hat.

Ausgehend von dem Informationsgehalt der einzelnen Zeichen kann der Informationsgehalt der gesamten Nachricht durch Aufsummieren berechnet werden (mit einem Alphabet der Größe n):

$$I = \sum_{1}^{n} I(z_n) \tag{7.2}$$

Zwei Münzwürfe zusammen genommen tragen somit die Information von 2 Shannon.

Beispiel: Die Zeichenkette

$$abaabaab \tag{7.3}$$

enthält insgesamt 8 Zeichen, 5 mal das Zeichen a und 3 mal das Zeichen b. Es ergeben sich die Wahrscheinlichkeiten für a und b:

$$p(a) = \frac{5}{8} = 0.625 \; bzw. \; p(b) = \frac{3}{8} = 0.375 \tag{7.4}$$

Der Informationsgehalt von a bzw. b in Shannon beträgt demnach:

$$I(a) = -log_2 0.625 = 0.6780 Sh \; bzw. \; I(b) = -log_2 0.375 = 1.4150 Sh \tag{7.5}$$

Das seltener auftretende Zeichen b trägt demnach mehr Information als das häufiger auftretende Zeichen a.

Aufsummiert ergibt sich der Informationsgehalt der gesamten Zeichenkette aus 7.3:

$$I('abaabaab') = 5 \cdot 0.6780 + 3 \cdot 1.4150 = 3.39 + 4.245 = 7.635 \; Sh \tag{7.6}$$

Festzuhalten ist, dass die Information einer Datenmenge oder eines bestimmten Zeichens aus dieser Datenmenge nur bestimmt werden kann, wenn die Datenmenge als Ganzes vorliegt. Die Übertragung von Sender zu Empfänger muss also abgeschlossen sein, da sich die Wahrscheinlichkeiten der einzelnen Zeichen mit jedem weiteren übertragenen Zeichen abändern und sich die endgültigen Wahrscheinlichkeiten nicht während der Übertragung bestimmen lassen. Ausnahme sind hier Datensätze, deren Alphabet sich nur aus einem einzigen Zeichen zusammensetzt: Nachrichten dieser Art tragen überhaupt keine Information. Sind die Einheiten einer Nachricht annähernd gleichverteilt und liegt bereits eine erhebliche Menge an Daten vor, so haben weitere hinzukommende Daten ebenfalls nur noch einen geringen Einfluss auf die Wahrscheinlichkeitsverteilungen der einzelnen Elemente (dies ist zum Beispiel bei Münz- oder Würfelwürfen der Fall). Unberührt hiervon ist die Tatsache, dass der absolute Informationsgehalt der gesamten Nachricht mit jedem weiteren Zeichen steigt.

Interpretation der Information:

- Je kleiner die Wahrscheinlichkeit eines Zeichens ist, desto größer ist seine Information. Dies bedeutet, dass seltene Zeichen einen höheren Informationsgehalt besitzen als häufiger auftretende Zeichen. Dabei ist der Zusammenhang zwischen Wahrscheinlichkeit und Informationsgehalt aufgrund des verwendeten Logarithmus nicht linear: Bei gleichmäßig steigender Wahrscheinlichkeit steigt der Informationsgehalt der Zeichen mit höher werdender Wahrscheinlichkeit immer langsamer an (Abbildung 7.1).

- Die Information eines Zeichens ist 0, wenn dieses Zeichen das einzige in der Nachricht vorkommende Zeichen ist, das Alphabet also nur aus diesem einen Zeichen besteht und es eine 100%ige Wahrscheinlichkeit hat ($p = 1$). Folglich ist der Informationsgehalt der gesamten Datenmenge ebenfalls 0. Dabei ist es unerheblich, wie oft sich das Zeichen wiederholt:

$$I(z|p = 1) = -log_2 1 = 0 \tag{7.7}$$

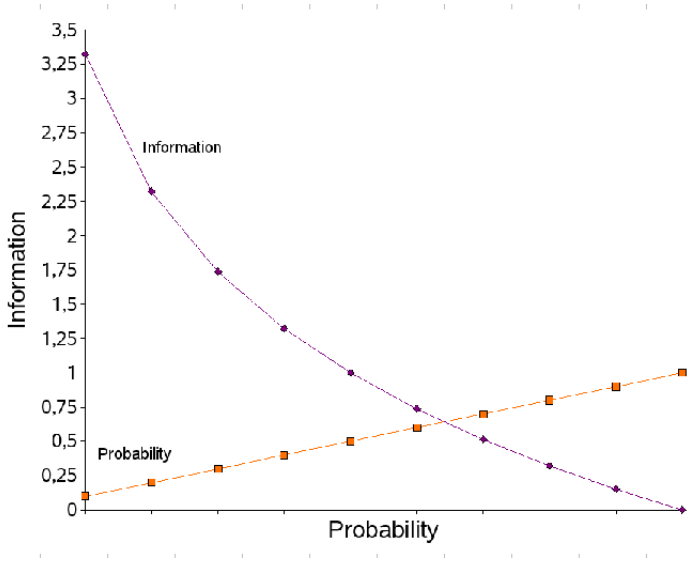


Abbildung 7.1: Zusammenhang zwischen steigender Wahrscheinlichkeit und Information: Mit höher werdender Wahrscheinlichkeit steigt der Informationsgehalt immer langsamer an

- Sind die Zeichen eines Datensatzes *gleichverteilt*, haben alle Zeichen also jeweils die gleiche Wahrscheinlichkeit, so haben sie auch den gleichen Informationsgehalt. Dies ist bei einem perfekten Münzwurf der Fall: Beide Seiten haben die gleiche Wahrscheinlichkeit (jeweils 0,5) und somit auch den gleichen Informationsgehalt von 1 Shannon. Analog gilt dies für die verschiedenen Seiten eines sechseitigen Spielwürfels. Jede Seite hat dieselbe Wahrscheinlichkeit von 1/6, und somit einen Informationsgehalt von ca. 2,58 Shannon.

- In der Informationstheorie ist negative Information nicht definiert[5].

- Werden keine weiteren Einteilungen der Nachricht definiert, haben weder die *absolute Position* eines atomaren Zeichens im Datensatz noch die *relative Position* eines Zeichens zu den anderen Zeichen der Nachricht Einfluss auf den Informationsgehalt der Zeichen. In diesem Fal-

[5]In der Quantenmechanik dagegen kann es sehr wohl zum Auftreten negativer Information kommen.

le wird die Nachricht wie eine ungeordnete Menge von Zeichen interpretiert. Dieser Umstand kann sich ändern, wenn die Nachricht als Ganzes in weitere, nicht-atomare Einheiten unterteilt wird und der Informationsgehalt nun auf Basis dieser größeren Einheiten berechnet wird. Dies geschieht beispielsweise bei der Miteinbeziehung der Wortgrenzen innerhalb der Dialektdaten.

- Das hier dargestellte Prinzip der Informationsquantifizierung lässt sich auf jede abgeschlossen und vollständig übertragen vorliegende Datenmenge anwenden, sofern diese sich in atomare Einheiten aufsplitten lässt.

7.2 Information in Teilmengen

Lässt sich eine Nachricht in mehrere Teilmengen aufteilen, so kann für jede dieser Teilmengen die enthaltene Information einzeln berechnet werden. Damit die gemessenen Informationsmengen in den einzelnen Teilmengen miteinander vergleichbar bleiben, werden die Wahrscheinlichkeiten der einzelnen Zeichen in Bezug auf die *gesamte* Datenmenge berechnet. Anschließend kann der Informationsgehalt der einzelnen Teilmengen in Relation zu der Information des gesamten Datensatzes gesetzt werden.

Beispiel: Die Zeichenkette aus dem oben genannten Beispiel (Kapitel 7.1) enthält 8 Zeichen. Aus diesen 8 Zeichen lassen sich 4 Teilmengen mit jeweils 2 Zeichen bilden[6]. Diese Teilmengen enthalten dann jeweils die in Tabelle 7.1, Zeile 3 angegebene Menge an Information in Shannon. Addiert man die Information der vier Teilmengen auf, so ergibt sich wieder der bereits in Kapitel 7.1 genannte Wert an Gesamtinformation von 7.635 Bit. Zeile 4 in Tabelle 7.1 enthält die Relation der jeweiligen Teilinformation zur Information des gesamten Datensatzes.

Da die Teilmengen 1, 3 und 4 jeweils aus denselben beiden Zeichen bestehen (a und b), ergibt sich in diesen drei Teilmengen derselbe Informations-

[6]Es wären natürlich auch andere Einteilungen denkbar.

	1	2	3	4
Teilmenge	ab	aa	ba	ab
Information Sh	2.093	1.356	2.093	2.093
Relation zum gesamten Datensatz	0,274	0,177	0,274	0,274

Tabelle 7.1: Vier Teilmengen

gehalt, unabhängig von der Reihenfolge der auftretenden Zeichen. Abweichend ist lediglich das zweite Teilstück.

Die Information ist umso höher, je kleiner die Wahrscheinlichkeit der verwendeten Zeichen ist. Dementsprechend haben Teilmengen mit seltener verwendeten Zeichen einen höheren Informationswert.

Damit sich partielle Informationen verschiedener Teilmengen miteinander vergleichen lassen, müssen die Teilmengen in etwa die gleiche Größe, sprich, annähernd dieselbe Anzahl an Zeichen, aufweisen. Die Einteilung der gesamten Datenmenge in weitere, die atomaren Einheiten bündelnde Einheiten ermöglicht es nun auch, den einzelnen Zeichen abhängig von ihrer Position innerhalb der Teilmenge einen anderen Wahrscheinlichkeitswert zuzuweisen. Hierdurch vergrößert sich generell das Alphabet, was im Endeffekt zu einem größeren Informationsgehalt des gesamten Datensatzes führt.

7.3 Von der Information zur Entropie

Auf das Konzept der Information aufbauend, entwickelte Shannon im bereits zitierten Artikel (Shannon, 1947) ein Konzept zur Berechnung der *Informationsdichte* einer Nachricht oder Datenmenge - die Entropie. Wie vorangehend beschrieben, summiert sich die Information der einzelnen Zeichen einer Nachricht auf, so dass generell gilt: Längere Nachrichten haben einen höheren Informationsgehalt als kürzere. Aus der Sicht der Information lassen sich somit nur Nachrichten von annähernd gleicher Länge miteinander vergleichen. Das Konzept der Entropie geht nun einen Schritt weiter und beschreibt die Informationsdichte einer Nachricht. Auf dieser Basis lassen sich dann auch Nachrichten unterschiedlicher Länge in Relation zueinander setzen.

In der Literatur finden sich neben dem Begriff Informationsdichte weitere Definitionen und Interpretationen des informationstheoretischen Entropiekonzepts. So lässt sich die Entropie auch als *Grad der Überraschung* beim Auftreten eines bestimmten Zeichens interpretieren.

Der Begriff *Entropie* stammt ursprünglich aus der Physik, genauer aus der Thermodynamik. Der zweite Hauptsatz der Thermodynamik lautet:

Energie ist nicht in beliebigem Maße in andere Arten umwandelbar, sondern nur bis zu Maximalwerten, die von einer weiteren Zustandsgröße, der sog. *Entropie*, abhängen (Baehr u. Kabelac, 2009).

Shannon erkannte, dass seine Formel zur Beschreibung der Informationsdichte einer Datenmenge eine fundamentale Ähnlichkeit zum Entropiebegriff der Thermodynamik aufwies. Ließ sich auf der einen Seite das Mischungsverhalten von Flüssigkeiten oder Gasen verschiedener Temperaturen beschreiben, so war es andererseits das Verhältnis von gewollter Information zu ungewolltem Datenrauschen[7].

Shannon definiert in Shannon (1947) die informationstheoretische Entropie wie folgt:

$$H(X) = -\sum_{1}^{n} p(x_n)\ log_2\ p(x_n) \tag{7.8}$$

Zu beachten ist, dass die Summe in Formel 7.8 nicht über jedes einzelne Zeichen des Datensatzes läuft, sondern lediglich für jedes vorkommende Zeichen bzw. dessen Wahrscheinlichkeit einmal. So stellt die Entropie, ähnlich den aufsummierten Informationswerten in Formel 7.2, einen Kennwert für den gesamten Datensatz dar. Allerdings liegen die Informationswerte aus Formel 7.2 in einem höheren Wertebereich als die Entropiewerte der einzelnen Datensätze.

[7]Später fand eine Rücktransformation des Informationsbegriffs in die theoretische Physik statt. So konnte durch das Berücksichtigen der Information in thermodynamischen Systemen das Paradoxon des sogenannten “Maxwellschen Dämonen” aufgelöst werden (Brillouin, 1962).

Ausgehend von der oben gezeigten Formel zur Berechnung der Entropie lassen sich weitere verwandte Konzepte definieren:

- Die **Blockentropie** (Joint Entropy) kombiniert die Entropien zweier Datenmengen X und Y, die jeweils aus einer Reihe von Elementen x bzw. y bestehen:

$$H(X,Y) = -\sum_{1}^{x,y} p_{x,y}\ log_2\ p_{x,y} \tag{7.9}$$

- Die **Bedingte Entropie** (Conditional Entropy) berechnet ähnlich wie die Blockentropie die Gesamtentropie zweier Datenmengen X und Y, allerdings mit bekannter Datenmenge X:

$$H(X,Y) = -\sum_{1}^{x,y} p_{x,y}\ log_2\ \frac{p_{x,y}}{p_x} \tag{7.10}$$

- Die **Kullback-Leibler Divergenz**[8] gibt an, wie unterschiedlich zwei Wahrscheinlichkeitsverteilungen $P = p(X)$ und $Q = p(Y)$ zueinander sind:

$$D(P||Q) = -\sum_{x \in X} p(x)\ log_2\ \frac{p(x)}{p(y)} \tag{7.11}$$

  Die Kullback-Leibler Divergenz ist nicht symmetrisch, so dass $D(P||Q) \neq D(Q||P)$. P kann eine ideale, vorgegebene Wahrscheinlichkeitsverteilung und Q eine gemessene darstellen. Die Kullback-Leibler Divergenz gibt in diesem Fall den Grad der Abweichung der gemessenen von der vorgegebenen Verteilung an. In der Dialektometrie angewandt, könnte die Kullback-Leibler Divergenz beispielsweise zwischen den jeweiligen Dialekten und einem als Gold-Standard definiertem Dialekt (oder der Hochsprache) gemessen werden.

[8] Für die Kullback-Leibler Divergenz existiert auch die Bezeichnung Relative Entropie, die allerdings auch für die Transinformation verwendet wird. Letztere lässt sich zwar über die Kullback-Leibler Divergenz definieren, ist aber nicht mit ihr identisch. Zur klaren Unterscheidung der beiden Konzepte sollte daher auf die Bezeichnung Relative Entropie verzichtet werden.

Ist in dieser Arbeit lediglich von *der Entropie* die Rede, so ist immer Bezug auf oben genannte (Formel 7.8), grundlegende Entropie-Definition von Shannon genommen.

Beispiel: Die Entropie für die oben genannte Zeichenkette *abaabaab* berechnet sich wie folgt:

$$H('abaabaab') = p(a) \cdot I(a) + p(b) \cdot I(b) = 0,954375 \tag{7.12}$$

Interpretation:

- Die Entropie bezeichnet die *Informationsdichte* eines Datensatzes. Sinkt in einem Datensatz bei gleichbleibender Zeichenanzahl die Entropie, so hat der mittlere Informationsgehalt der Zeichen abgenommen. Steigt die Entropie an, so steigt auch der mittlere Informationsgehalt aller Zeichen.

- Die *maximale Entropie* ist erreicht, wenn die Zeichen eines Datensatzes gleichverteilt sind. In diesem Fall hat jedes vorkommende Zeichen die gleiche Wahrscheinlichkeit und trägt somit auch den gleichen Gehalt an Information. Dies führt schussendlich zu dem bei der jeweiligen Anzahl an Zeichen maximal möglichen Entropiewert.

- Bestehen die Zeichen eines Datensatzes nur aus einem einzigen Zeichen, so hat dieses eine Wahrscheinlichkeit von 100%. Da in diesem Fall der Grad der Überraschung gleich null ist, ist auch der Informationsgehalt und somit ebenfalls die Entropie null.

7.3.1 Einbeziehung der Position

In ihrer ursprünglichen Form sind informationstheoretische Methoden quantitativer Natur und betrachten lediglich die unterschiedliche Anzahl der vorkommenden Elemente. Die *Reihenfolge* der Elemente zueinander und ihre Position innerhalb der Nachricht werden nicht eingerechnet.

Um auch die relativen Positionen der Elemente innerhalb der Nachricht in die Berechnung einzubeziehen, kann die Berechnung des Informationsgehalts wie oben beschrieben duch das Einfügen weiterer Einheiten abgeändert werden. Die Wahrscheinlichkeit eines Elements wird nicht auf Basis aller Elemente des Datensatzes sondern auf Basis der *Vorkommen des Elements an der enstprechenden Stelle in der höheren, nicht-atomaren Einheit* berechnet. Dabei wird unterschieden, an welcher Stelle in der Einheit das Element vorkommt und jede dieser Möglichkeiten wie ein eigenes, individuelles Element behandelt. Damit dies möglich wird, muss zwischen den atomaren Einheiten einer Nachricht und deren Gesamtheit eine weitere Einheit eingezogen werden. Im Rahmen der Dialektometrie kann diese Rolle das *Wort* übernehmen, welches somit als natürliche Gliederung der dialektalen Daten in die Berechnungen miteinbezogen wird.

Beispiel: Ein Datensatz enthält insgesamt 1000 Elemente. Das aktuelle Element z ist 100 Mal enthalten. Die Information von z berechnet sich demnach wie oben beschrieben:

$$I(z) = -log_2 p(z) = -log_2(0.1) \approx 3.32\ Sh \tag{7.13}$$

Für alle z ergibt sich bei dieser Art der Berechnung eine Gesamtinformation von ca. 332 Shannon.

Der Datensatz lässt sich nun in weitere Einheiten, beispielsweise Wörter, aufteilen. Von den insgesamt 100 Vorkommen von z kommen insgesamt 30 an zweiter Stelle im Wort, 50 an dritter und 20 an vierter Stelle im Wort vor. Die Information wird nun separat für alle diese Möglichkeiten berechnet. Für ein z an zweiter Stelle innerhalb eines Wortes berechnet sich die Information demnach:

$$I(z)_2 = -log_2 p(z)_2 = -log_2(0.03) \approx 5.05\ Sh \tag{7.14}$$

Analog errechnen sich 4.32 Shannon für z an dritter Stelle und 5.64 Shannon an vierter Stelle. Gemäß dem Paradigma der Informationstheorie tragen seltenere Elemente mehr Information als häufigere Elemente. Aufsummiert ergibt sich für alle Elemente z eine Gesamtinformation von

$$I(z)_{gesamt} = 30 * 5.05 + 50 * 4.32 + 20 * 5.64 = 480.3\ Sh \tag{7.15}$$

Bislang wurden alle Elemente eines Datensatzes wie eine ungeordnete Menge behandelt. Durch die hier gezeigte Methode steigt die Anzahl der verschiedenen Elemente innerhalb eines Datensatzes (das Alphabet wird größer), gleichzeitig ändert sich die absolute Anzahl der Elemente nicht. Folge ist ein generelles Ansteigen der Information wie im obigen Beispiel berechnet. Damit wird eine feinere Auflösung der Unterschiede zwischen den einzelnen Datensätzen möglich.

Anstelle jeder möglichen Position des Elements z innerhalb des Wortes eine eigene Wertigkeit zuzuordnen, kann das Wort auch in eine festgelegte Anzahl Abschnitte aufgeteilt werden. Die Worte könnten so beispielsweise in Präfix, Suffix und Stamm aufgeteilt werden. Hier sind weitere Aufteilungen in verschiedenen Formen von N-Grammen denkbar.

Kapitel 8

Analyse von dialektometrischen Ergebnissen

Sind mittels der in den vorangegangenen Kapiteln vorgestellten Methoden Dialektdaten quantitativ ausgwertet worden, so folgt als nächster Schritt die Analyse der Ergebnisse (dritter Schritt im dialektometrischen Workflow, siehe Abbildung 4.1). Hierdurch werden die gewonnenen Resultate automatisiert geordnet und - hauptsächlich in Form von Visualisierungen - der Interpretation zugänglich gemacht, entsprechend der vierten Säule in Abbildung 4.1.

8.1 Intervallagorithmen

Intervallalgorithmen wurden von Hans Goebl in die Dialektometrie eingeführt (Goebl, 1982, S. 93 ff.). Mit ihrer Hilfe können die Werte einer numerischen Messreihe in eine vorgegebene Anzahl von n Klassen eingeteilt werden. Goebl unterscheidet drei Varianten von Intervallalgorithmen, die ebenfalls in der VDM-Software implementiert sind.

Die Elemente der Messreihe werden ihrer Größe nach geordnet und anschließend anhand von einem der folgenden Algorithmen in n Klassen eingeteilt (Abbildung 8.1):

- **MINMWMAX**: Die zwei Bereiche zwischen dem Minimum und dem arithmetischen Mittel einer Messreihe einerseits, sowie dem arithme-

tischen Mittel und dem Maximum andererseits werden jeweils in $n/2$ Klassen eingeteilt. Anschließend werden die einzelnen Messwerte diesen Klassen zugeordnet. Bei diesem Intervallagorithmus können die einzelnen Klassen eine unterschiedliche Anzahl Elemente beinhalten.

- Auch der Algorithmus **MEDMW** teilt die Messreihe entlang des arithmetischen Mittels und dem Minimum bzw. Maximum in zwei Bereiche ein. Anschließend werden die in die jeweiligen Bereiche fallenden Werte möglichst gleichmäßig in $n/2$ Klassen aufgeteilt. Im Gegensatz zu MINMWMAX sind die Klassen annähernd gleich groß.

- Ohne Berücksichtigung des arithmetischen Mittels, teilt **MED** die numerisch geordneten Werte einer Messreihe in die vorgegebene Anzahl Klassen ein. Ähnlich wie bei MEDMW entstehen so annähernd gleich große Klassen.

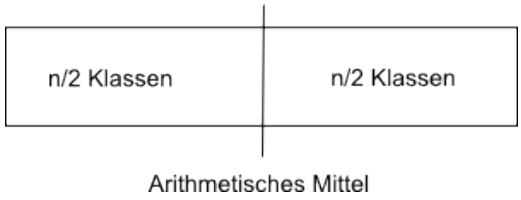


Abbildung 8.1: Einteilung der Messreihe in Klassen unter Zuhilfenahme von Intervallagorithmen

Die Algorithmen MEDMW und MED legen das Hauptaugenmerk auf die gleichmäßige Verteilung der Messwerte auf die vorgegebene Anzahl Klassen. Wird das Verfahren MINMWMAX angewandt, so können die entstandenen Klassen durchaus unterschiedlich groß sein, numerisch näher zusammenliegende Werte werden sich so aber auch eher in einer Klasse wiederfinden.

Nach erfolgter Einteilung in Klassen können diese auf einer topographischen Karte oder als Histogramm visualisiert werden.

8.2 Clusteranalyse

Die Clusteranalyse ist ein Verfahren aus dem Bereich des *Machine Learning*. Mit ihrer Hilfe können Messreihen analysiert und die den Daten inhärenten Strukturen sichtbar gemacht machen. Dabei werden Elemente mit hoher Ähnlichkeit zu größeren Gruppen, den Clustern, zusammengefasst. In weiteren Durchläufen werden anschließend die einzelnen Cluster zu immer größeren Clustern gebündelt. In der sich ergebenden Struktur finden sich innerhalb eines Clusters möglichst ähnliche Elemente, wohingegen die Menge der Cluster die größtmögliche Unähnlichkeit zueinander aufweisen.

Clusteranalyse kommt in vielen empirisch arbeitenden Wissenschaften zur Anwendung. Dies umfasst nicht nur naturwissenschaftliche und technische Bereiche, sondern auch beispielsweise empirische Sozial- und Wirtschaftswissenschaften. In der Dialektometrie kann Clustering zur Analyse der mit den verschiedenen Methoden gewonnenen messpunktspezifischen Kennzahlen verwendet werden. Dargestellt in Form einer Ähnlichkeits- oder Distanzmatrix, können sie als Input für die verschiedensten Clusteringalgorithmen dienen. Die einzelnen Dialekte stellen somit die atomaren Elemente dar, die im Verlauf des Clustering zu immer größeren Clustern zusmmengefasst werden.

Die sich so ergebenden Cluster können als Dialekt-Gruppen mit abgestufter Ähnlichkeit zueinander interpretiert werden. Je weiter zwei Cluster in der Hierarchie voneinander entfernt sind, desto unähnlicher sind sich die jeweils enthaltenen Dialekte. Dies gilt sowohl für Ähnlichkeits- als auch für Distanzmatrizen, wobei das eine Clustering jeweils das Negativ des anderen darstellt.

Der hierarchische Aufbau der Cluster ergibt sich aus dem Zusammenspiel zwischen der Beschaffenheit der Daten, dem verwendeten Distanzmaß und dem Clusteringalgorithmus. Er kann anschließend an den eigentlichen Clusteringprozess auf verschiedene Art und Weise visualisiert und unter Hinzuziehen extralinguistischer Informationen interpretiert werden (Abbildung 8.2).

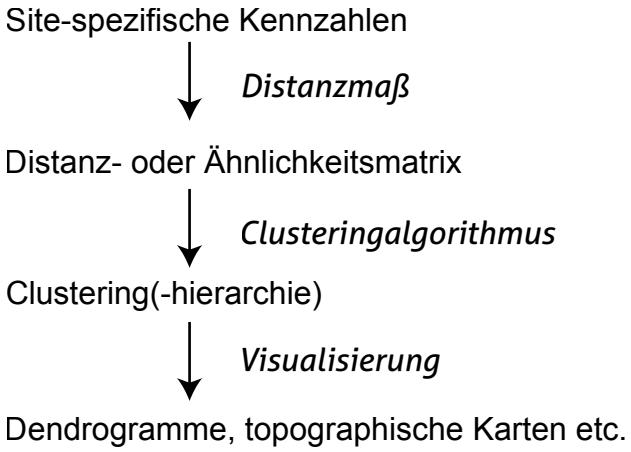


Abbildung 8.2: Ablauf des Clustering

8.2.1 Eigenschaften der Clusteranalyse

Die Anzahl der gebildeten Cluster m hat einen Wert zwischen 1 und n, wobei n die Anzahl der Elemente in der untersuchten Messreihe darstellt:

$$1 <= m <= n \tag{8.1}$$

Kann ein Element immer nur Mitglied eines Clusters sein, handelt es sich um *hartes Clustering*. Können Cluster sich überlappen, einzelne Elemente also Mitglied in mehreren Clustern sein, spricht man von *weichem Clustering*.

Die Methoden der Clusteranalyse sind *unüberwacht*, es werden keine Trainingsdaten oder vorhergehende manuell zu treffende Annahmen benötigt. Die Clusteranalyse ist ein multivariates Verfahren und kann als solche im Gegensatz zu univariaten Verfahren mehrere Aspekte bzw. Dimensionen der Elemente einer Messreihe in die Analyse einbeziehen.

Neben den in dieser Arbeit verwendeten *hierarchischen* Clusteringalgorithmen sind *partitionierende* Methoden weit verbreitet.

- **Hierachisches Clustering:** Hierbei bilden die Cluster in ihrer Gesamtheit eine hierarchische Baumstruktur aus. Der oberste, alle anderen umfassende Cluster repräsentiert die Wurzel des Hierarchiebaumes.

Auf der gegenüberliegenden Seite repräsentieren die atomaren Elemente die terminalen Knoten des Baumes. Je nachdem, ob *top-down* oder *bottom-up* vorgegangen wird, beginnt die Konstruktion der Hierarchie mit der Wurzel, sprich, dem größten, alle anderen umfassenden Cluster (top-down) oder an den Blättern, den atomaren Elementen des Datensatzes (bottom-up). Im ersten Fall werden alle Elemente anfangs in ihrer Gesamtheit als ein großer Cluster betrachtet, der dann im weiteren Verlauf rekursiv in mehrere kleinere Cluster aufgeteilt wird bis jedes Element einen eigenen Cluster bildet. Bei der bottom-up Vorgehensweise wird anfänglich jedes Element als einzelner Cluster betrachtet, die anschließend im weiteren Verlauf des Clusteringprozesses zu größeren Clustern zusammengefasst werden.

Im Gegensatz zu den Intervallalgorithmen sind beim hierarchischen Clustern die Kategorien bzw. Cluster nicht vorgegeben, ihre Einteilung und Anzahl ergeben sich durch automatisierte Berechnungen während des Clusteringprozesses. Die Einteilung der Daten in Cluster erfolgt iterativ: In jedem Durchgang werden die bereits vorhandenen Cluster entweder zu größeren Einheiten zusammengefaßt bzw. in weitere, kleinere Cluster aufgeteilt. Beispiele für hierarchisches Clustering sind die Linkage-Verfahren, die Pair-Group-Algorithmen und die WARD-Methode.

- **Partitionierendes Clustering:** Beim partitionierenden Clustering werden alle atomaren Elemente gleichzeitig in Cluster eingeteilt. Im Gegensatz zum hierarchischen Clustering liegt zu jeder Zeit des Clusteringprozesses die volle Anzahl der errechneten Cluster vor. In weiteren Iterationsschritten werden die einzelnen Elemente den ihnen am ähnlichsten Clustern neu zugeordnet, bis sich schließlich keine Änderungen in der Einteilung der Cluster mehr ergeben. Der *k-means Algorithmus* ist ein weit verbreitetes Beispiel für partitionierende Cluster Algorithmen.

Abbildung 8.3 zeigt einen vierstufigen, hierarchischen Clustering-Prozess. In der ersten Stufe A liegen die vier atomaren Elemente 1 bis 4 noch ungruppiert vor. In den weiteren Stufen B bis D werden zunächst mit Hilfe

einer Distanzfunktion die Elemente 1 und 2 (Cluster a) bzw. 3 und 4 (b) zu Clustern zusammengefasst. In der letzten Stufe D werden anschließend die beiden entstandenen Cluster zu einem dritten, übergreifenden Cluster (c) zusammengefasst. Erfolgt der Prozess von links nach rechts, handelt es sich um Bottom-up Clustering. Umgekehrt von rechts nach links findet top-down Clustering statt.

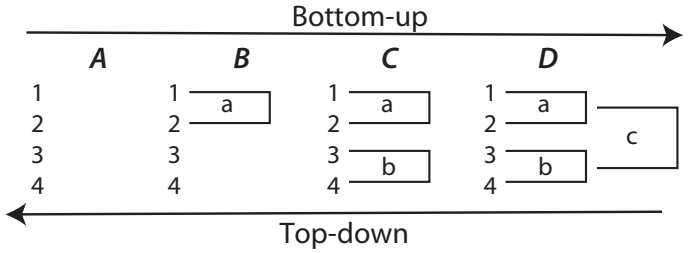


Abbildung 8.3: Vierstufiger Clustering-Prozess

8.2.2 Distanzfunktionen

Der Grad der Ähnlichkeit zwischen Elementen lässt sich mittels einer *Distanzfunktion* berechnen. Je nach Art der vorliegenden Daten, numerisch oder alphanumerisch, uni- oder multivariat, können verschiedene Distanzfunktionen angewendet werden. Distanzfunktionen sind nur im positiven Bereich definiert, sie sind symmetrisch ($\Delta x(A, B) = \Delta x(B, A)$) und die Distanz eines Elements zu sich selbst oder einem weiteren, identischen Element beträgt 0.

Die einfachste Distanzfunktion berechnet den Abstand Δx zwischen zwei numerischen, *univariaten* Elementen A und B als Betrag der Subtraktion:

$$\Delta x(A, B) = |A - B| \tag{8.2}$$

Bei *multivariaten* Daten können, abhängig von der Art der jeweiligen Daten, verschiedene Distanzfunktionen angewendet werden. Multivariate Elemente sind durch mehrere Eigenschaften - bzw. Dimensionen - definiert. Für die Elemente A und B mit n Merkmalen ergeben sich: $A = a_1...a_n$ bzw. $B = b_1...b_n$.

Die *Euklidische Distanz* berechnet den Abstand zwischen zwei Punkten im höher dimensionalen Raum. Abgeleitet vom *Satz des Pythagoras*, berechnet sie sich wie folgt:

$$\Delta x(A, B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{8.3}$$

Ähnlich zur Euklidischen Distanz, kann die *Manhatten-Distanz* zur Berechnung der Ähnlichkeit multivariater, numerischer Elemente herangezogen werden. Sie stellt die Summe der Distanzen zwischen den einzelnen Merkmalen dar:

$$\Delta x(A, B) = \sum_{i=1}^{n} |a_i - b_i| \tag{8.4}$$

Die *Jaccard-Distanzfunktion* berechnet den Abstand zwischen zwei Mengen an Elementen und ist definiert als der Quotient aus der Schnittmenge mit der Vereinigungsmenge[1]:

$$\Delta x(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{8.5}$$

Liegen alphanumerische und keine numerischen Daten vor, können *Edit-Distanz Algorithmen* zur Berechnung der Ähnlichkeit zwischen den Zeichenketten benutzt werden[2]. Die *Hamming-Distanz* berechnet die Distanz zwischen zwei gleich langen Strings:

$$\Delta x(A, B) = \sum_{i=1}^{n} 1 : \ a_i = b_i \mid 0 : \ a_i \neq b_i \tag{8.6}$$

[1]Die Jaccard-Distanzfunktion ist ähnlich zu der dialektometrischen Methode des *Relativen Identitätswerts*.

[2]Siehe auch Kapitel 4.3.2.

Andere Edit-Distanz Algorithmen wie beispielsweise die Levenshtein-Distanz können den Abstand auch bei unterschiedlicher Länge der beiden Zeichenketten bestimmen.

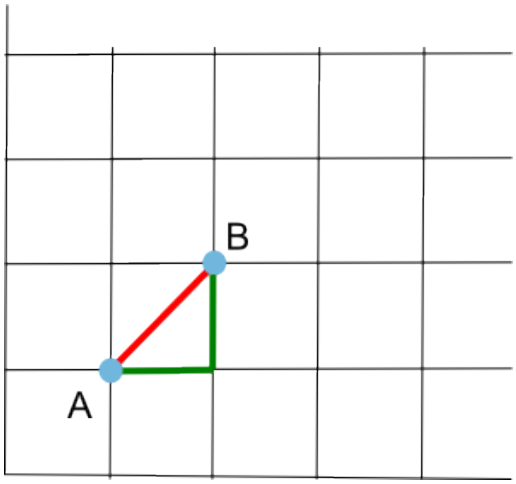


Abbildung 8.4: Numerische Distanzfunktionen. Die rote Linie markiert die euklidische Distanz, die grüne die Manhattan Distanz zwischen den Elementen A und B

8.2.3 Hierarchische Clusteringalgorithmen

Die oben genannten Distanzfunktionen berechnen Ähnlichkeiten zwischen den einzelnen Elementen eines Datensatzes. Die sich jeweils ähnlichsten Elemente werden zu Clustern gruppiert. Im weiteren Verlauf werden diese anfänglichen Cluster zu größeren Clustern zusammengefasst. Es werden wieder jeweils die Cluster, die die größte Ähnlichkeit zueinander aufweisen, zu einem größeren Cluster zusammengefasst. Um die Ähnlichkeit zwischen Clustern zu berechnen, können verschiedene *hierarchische Clusteringalgorithmen* eingesetzt werden. Je nach eingesetztem Verfahren werden sich unterschiedliche Strukturen in der resultierenden Cluster-Hirarchie ergeben.

Die beiden Cluster X und Y sind definiert durch ihre Elementmengen x bzw. y:

$$x = \{x_1 ... x_m\} \; bzw. \; y = \{y_1 ... y_n\} \tag{8.7}$$

Linkage Algorithmen

Linkage oder *Neighbor Joining Algorithmen* berechnen die Ähnlichkeit $D(X, Y)$ zwischen den Clustern X und Y auf Grundlage zweier in den jeweiligen Clustern enthaltener Elemente (siehe Abbildung 8.5).

- Single Linkage: Der Abstand zweier Cluster ist identisch zum Abstand der beiden ähnlichsten Elemente aus den jeweiligen Clustern:

$$D(X, Y) = min(\Delta x(x, y)) \tag{8.8}$$

- Average Linkage: Die beiden Elemente mit dem mittleren Abstand zueinander bestimmen den Abstand der Cluster:

$$D(X, Y) = \overline{\Delta x(x, y)} \tag{8.9}$$

- Complete Linkage: Die am weitesten voneinander entfernten Elemente aus beiden Clustern geben auch den Abstand der Cluster an:

$$D(X, Y) = max(\Delta x(x, y)) \tag{8.10}$$

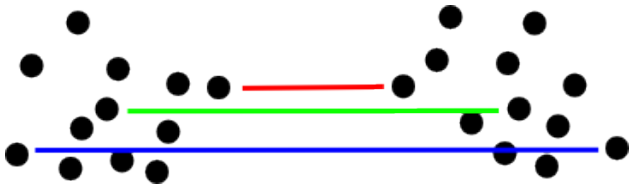


Abbildung 8.5: Verschiedene Linkage-Verfahren um den Abstand zweier Cluster zueinander zu bestimmen: Single Linkage (oberste, rote Linie), Average Linking (mittlere, grüne Linie) und Complete Linkage (untere, blaue Linie)

Pair Group Algorithmen

Grundlage der *Pair Group Methoden* ist die Gesamtheit der Distanzen zwischen allen möglichen Elementpaaren aus zwei Clustern.

- Unweighted Pair Group Method with Arithmetic Averages (UPGMA): Der Abstand zweier Cluster zueinander ist definiert als das arithmetische Mittel zwischen den Distanzen aller möglicher Elementpaaren aus beiden Clustern

- Weighted Pair Group Method with Arithmetic Averages (WPGMA): Ähnlich zu UPGMA, aber die Anzahl der Elemente in den jeweiligen Clustern fließt als Gewichtung in die Distanz-Berechnung zwischen den einzelnen Elementen zusätzlich mit ein

Zentroid Algorithmen

Das Zentroid *cent* eines Clusters ist ein zumeist künstlicher Datenpunkt, der sich berechnet aus dem Mittelwert aller im Cluster enthaltener Elemente (Formel 8.11):

$$cent(X) = \frac{\sum_{i=1}^{m} x_i}{m} \tag{8.11}$$

Der Abstand zweier Cluster ist definiert als der Abstand der beiden Zentroiden zueinander. Ähnlich zu den Pair Group Methoden, lassen sich ungewichtete und gewichtete Zentroid-Algorithmen unterscheiden.

- Unweighted Pair Group Centroid (UPGMC): Der Abstand zweier Cluster entspricht dem Abstand der beiden Zentroiden zueinander

$$D(X, Y) = |cent(X) - cent(Y)| \tag{8.12}$$

- Weighted Pair Group Centroid (WPGMC): Diese Methode ist identisch zu UPGMC, allerdings fließt die Anzahl der Elemente je Cluster als Gewichtung in die Erstellung der Zentroiden ein

Wards Methode

Die Distanz $D(X, Y)$ zwischen zwei Clustern X und Y kann nach der Methode Ward wie folgt berechnet werden (für eine detailliertere Erklärung, siehe Deichsel u. Trampisch (1985), S. 29-33):

$$D(X, Y) = \frac{n_1 \cdot n_2}{n_1 + n_2}(cent_1 - cent_2)^2 \tag{8.13}$$

wobei n_1 bzw. n_2 für die Anzahl der im jeweiligen Cluster enthaltenen Elemente steht und $cent_1$ bzw. $cent_2$ für das jeweilige Zentroid.

Weiterführende Erklärungen zur Funktionsweise der anderen oben aufgeführten Distanzfunktionen finden sich in (Schulte im Walde, 2003, S. 199 ff.).

8.2.4 Noisy Clustering

Hierarchische Clusteralgorithmen besitzen die Eigenschaft, dass während des Clusteringprozesses nicht ersichtlich ist, wie die jeweilige Clusteraufteilung zustande gekommen ist. Ob der numerische Abstand zweier Cluster im Vergleich zu den Abständen zu den anderen Clustern als eher hoch oder eher niedrig angesehen werden muss, lässt sich nicht alleine anhand des Ergebnisses des Clusteringprozesses entscheiden. So kann ein numerisch relativ geringer Wert ebenso zur Aufspaltung eines Clusters führen wie ein relativ hoher Wert. Hieraus ergibt sich die Frage, wie stabil die aktuelle Aufteilung der Cluster ist und welche Werte konkret zu ihrer Entstehung geführt haben.

Abbildung 8.6 zeigt die drei Cluster A, B und C mit den (hier willkürlich festgelegten) Distanzen $D(A, B) = 1$ und $D(B, C) = 9$. Hier spielt es keine Rolle, durch welche konkrete Clusteringmethode die jeweiligen Distanzen entstanden sind - durch den wesentlich höheren Distanzwert zwischen den Clustern B und C im Vergleich zu der Distanz zwischen A und B ergibt sich eine wesentlich stabilere Grenze zwischen diesen beiden Clustern.

Um einen Eindruck von der Stabilität der aktuellen Clusteraufteilung zu gewinnen, kann *Noisy Clustering*[3] eingesetzt werden. Hierbei wird zu den

[3] In Nerbonne u. a. (2008) beschreiben die Autoren einen Vergleich zwischen Noisy Clustering und Bootstrapping: Beide Methoden führten zu ähnlichen Ergebnissen.

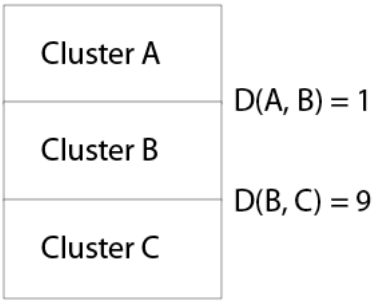


Abbildung 8.6: Drei Cluster A, B und C mit den Distanzen $D(A, B) = 1$ und $D(B, C) = 9$

Werten der Distanzmatrix ein per Zufallsgenerator errechneter Wert γ hinzugefügt. Dieser Wert darf einen bestimmten, vorher festgelegten Betrag nicht überschreiten und wird für jede Zelle der Distanzmatrix individuell berechnet. γ kann beispielsweise definiert werden als ein Zufallswert, der zwischen 0 und einem festzulegenden Anteil der Standardabweichung der originalen Daten[4] liegen muss:

$$\gamma = q \cdot \sigma_X \; mit \; 0 \leq q \leq 1 \tag{8.14}$$

Anschließend wird der Clusteringprozess bei gleichbleibenden Parametern erneut durchgeführt und die Aufteilung der Cluster bei einer bestimmten Anzahl von Clustern mit der Aufteilung bei Verwendung der Originaldaten verglichen. War die originale Clusteraufteilung instabil bzw. auf der Grundlage relativ kleiner Distanzen zwischen den Clustern zustandegekommen, so wird die neue Aufteilung mit den absichtlich verrauschten Daten ein anderes Bild der Clusteraufteilung produzieren.

Um die Stabilität des Clusteringprozesses deutlich zu machen, kann die Obergrenze, die γ nicht übersteigen darf, sukzessive erhöht werden und der Clusteringprozess neu ausgeführt werden. Dabei dürfen die sonstigen Pa-

[4]Wird Rauschen zu den originalen Daten hinzugefügt, so wird die Standardabweichung in den neu berechneten Daten höher sein als in den Ausgangsdaten. Zur besseren Vergleichbarkeit sollte γ deswegen immer in Bezug auf die Standardabweichung der originalen Daten berechnet werden.

rameter des Clusterings, wie Distanzmaß oder Clusteringalgorithmus, nicht verändert werden.

Ab einem bestimmten Wert wird das Rauschen so groß werden, dass es die in den Daten inhärenten Strukturen komplett überdeckt und sich keine sinnvolle Aussage aus der Aufteilung der Cluster mehr ziehen lässt.

8.2.5 Visualisierung von Clusterprozessen

Ausgangspunkt für jeden Clusteringprozess ist eine Distanz- oder Ähnlichkeitsmatrix. Diese lässt sich in Form einer *Heatmap*[5] visualisieren. Den einzelnen Distanzen der Matrix werden, abhängig von ihrem Wert, unterschiedliche Farbtöne zugewiesen. Diese entstammen dem Spektrum eines zuvor definierten Farbverlaufs, dessen Start- und Endpunkt jeweils dem Maximum bzw. dem Minimum der in der Matrix enthaltenen Werte entsprechen.

Abbildung 8.7 visualisiert eine Distanzmatrix, die aus einer künstlichen, univariaten Messreihe mit den Werten {1,2,3,5,7,9,12,15,18} gebildet wurde. Der Farbverlauf von Dunkelblau nach Hellgelb entspricht dem Verlauf der enthaltenen Werte von kleinster Distanz (0, Blau) hin zu größter Distanz (1.00, Gelb).

Wurde ein hierarchisches Clusteringverfahren auf eine Distanzmatrix angewandt, so kann die entstandene Struktur in Form eines *Dendrogramms*[6] visualisiert werden. Das Dendrogramm entpricht einer Baumstruktur, die einzelnen Elemente stellen die *terminal leafs* und die zusammenfassenden Cluster die *non-terminal leafs* dar. Die Wurzel des Baumes wird durch den größten, alle anderen umfassenden Cluster repräsentiert. Ausgehend von den konkreten Elementen hin zu der Wurzel steigt die Distanz zwischen den einzelnen Clustern. Ein Dendrogramm ist unabhängig von verwendetem Distanzmaß oder dem jeweiligen Clusteringalgorithmus und stellt den Clusteringprozess als Ganzes dar.

[5] Heatmaps können in gängigen Statistik-Anwendungen, bspws. Orange oder SAS JMP erstellt werden.

[6] Als Alternative zu Dedrogrammen können *Clustergramme* verwendet werden, siehe Schonlau (2002).

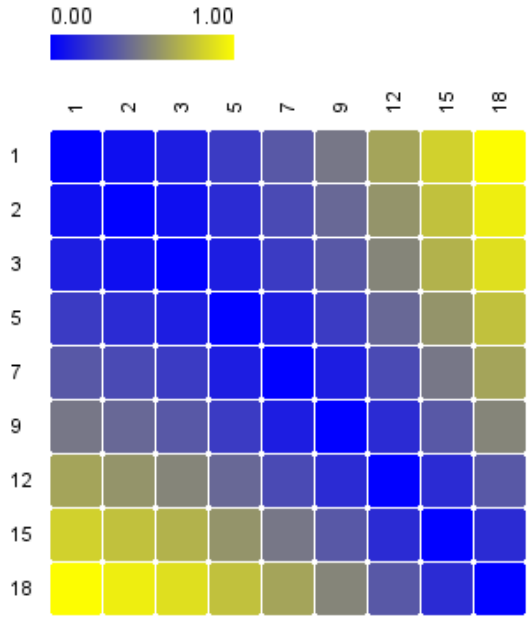


Abbildung 8.7: Heatmap, die einer aus den Werten {1,2,3,5,7,9,12,15,18} gebildeten Distanzmatrix entspricht

In der Dialektometrie kann ein Dendrogramm *diachron* interpretiert werden. Dabei entsprechen die *terminal leafs*, die an geographisch fixen Punkten empirisch festgehaltenen Dialektdaten, der aktuellen dialektalen Aufsplitterung der untersuchten Sprache. Die immer weitere Zusammenfassung der Cluster bis hin zu einem einzigen, alle Dialekte umfassenden Cluster entspricht einer rückwärts gewandten Bewegung in der Zeit. Bei dieser Interpretation des Dendrogramms wird davon ausgegangen, dass Dialekte sich durch immer feinere Aufsplitterungen übergeordneter Dialekte gebildet haben. Überkreuzende Einflüsse über die Cluster-Grenzen hinweg oder das Verschmelzen zweier oder mehrerer Dialekte miteinander wird in der diachronen Auffassung des Dendrogramms nicht berücksichtigt. Synchron interpretiert stellt das Dendrogramm ein Maß für die Ähnlichkeit bzw. Distanzen der einzelnen Dialekte zum Zeitpunkt der Datenerhebung dar.

Abbildung 8.8 zeigt vier Dendrogramme, die die hierarchischen Clusteringverfahren Single Linkage, Average Linking, Complete Linkage und Wards Methode visualisieren. Zugrunde liegt wiederum die bereits oben genannte Distanzmatrix aus den Werten {1,2,3,5,7,9,12,15,18}. Deutlich erkennbar ist, dass alle vier Clusteringalgorithmen die Daten in unterschiedliche Clusterstrukturen aufteilen.

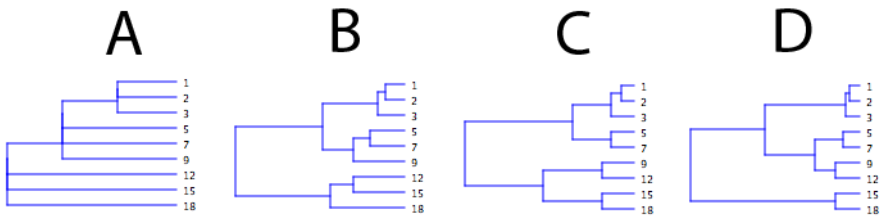


Abbildung 8.8: Ergebnis verschiedener Clustering-Verfahren anhand der Datenreihe {1,2,3,5,7,9,12,15,18 }: A - Single Linkage, B - Average Linking, C - Complete Linkage und D - Wards Methode

Besonders nützlich im Bereich der Dialektometrie sind Visualisierungen der Resultate auf topographischen Karten. Im Gegensatz zu Dendrogrammen können die Ergebnisse von Clusteringprozessen auf einer topographischen Karte nicht als Ganzes dargestellt werden. Es kann immer nur jeweils

ein Schritt während des Clusteringprozesses, entsprechend einer bestimmten Anzahl an Clustern, auf der Karte visualisiert werden. Dies kann beispielsweise durch unterschiedliche Formen der die einzelnen Messpunkte anzeigenden Punkte oder durch die Verwendung verschiedener Farben (Voronoi-Karte) erfolgen.

8.3 Multidimensional Scaling

Multidimensional Scaling (MDS) ist eine weitere Visualisierungstechnik, die Elemente aufgrund ihrer Ähnlichkeit bzw. Distanz in Relation zueinander anordnet. Dies geschieht in einer zweidimensionalen Ebene oder einem dreidimensionalen Raum: Je ähnlicher sich zwei Elemente sind, desto näher aneinander werden sie plaziert. Gruppen ähnlicher Elmente werden so zu wolkenähnlichen Gebilden zusammengefasst, Ausreißer in den Daten befinden sich abseits der gebildeten Wolken. Der Abstand der Elemente zueinander wird hierbei als ein kontinuierlicher Werteraum aufgefasst. Somit umgeht MDS den Nachteil des hierarchischen Clustering[7]: Die Einteilung der Daten erfolgt immer in eine diskrete Anzahl an Cluster. Dabei ist es anhand des Resultats nicht mehr ersichtlich, wie groß der Abstand der einzelnen Cluster zueinander ist. Durch den kontinuierlichen Werteraum sind die gebündelten Wolken des MDS unscharf, ohne klare Abgrenzungen voneinander. Ausreißer werden so nicht als eigenständige Cluster dargestellt, sondern räumlich zwischen den Wolken plaziert.

Ein Nachteil des MDS besteht darin, dass eine exakte Visualisierung von n Elementen mit $n > 1$ genau $n - 1$ Dimensionen benötigt. In der Praxis bedeutet dies, dass lediglich Datenmengen mit bis zu $n = 3$ in einer zweidimensionalen Ebene und Datenmengen mit bis zu $n = 4$ im dreidimensionalen Raum exakt visualisiert werden können (Abbildung 8.9). Für jedes weitere Element käme eine weitere Dimension hinzu, die aber nicht realisierbar ist.

Um die Beschränkung auf 2 oder 3 Dimensionen umgehen zu können, wurden verschiedene Vorgehensweisen entwickelt, um mit Hilfe von Näherungs-Algorithmen die höheren Dimensionen auf den zwei- oder dreidimensionalen Raum abbilden zu können. Dies resultiert wiederum in theoretisch endlos laufenden Algorithmen, die aus mathematischer Sicht nicht zu einem engültigen Ergebnis kommen können. Allerdings ergeben sich in der Praxis nach einer gewissen Anzahl an Durchläufen stabile Kombinationen aus Wolken und Ausreißern, deren relative Positionen zueinander sich im weiteren Verlauf des Algorithmus nicht mehr oder nur noch kaum verändern.

Bildlich gesprochen, üben die einzelnen Objekte eine anziehende bzw. ab-

[7] Das Noisy Clustering versucht auf eine andere Weise, denselben Nachteil zu umgehen.

stoßende Kraft auf alle anderen Objekte des Datensatzes aus. Diese Gravitationswirkung der einzelnen Elemente aufeinander kann mittels unterschiedlicher Algorithmen berechnet werden. Dabei überprüft der Algorithmus, inwieweit die aktuellen Entfernungen der Elemente zueinander den gegebenen Entfernungen in der Distanz-Matrix entspricht. Anschließend werden die Elemente neu zueinander ausgerichtet, so dass ihre Entfernungen nun eher den Vorgaben entsprechen. Hierzu berechnet sich die Differenz $diff$ zwischen der Angabe aus der Distanz-Matrix dm und der aktuellen Entfernung auf der Anzeigefläche af:

$$diff = af - dm \tag{8.15}$$

Hieraus ergeben sich die Näherungs-Algorithmen[8], für weitergehende Informationen zu den hier genannten Näherungs-Algorithmen, siehe Li (2005):

- Kruskal Stress: $diff^2$
- Sammon Stress: $\frac{diff^2}{af}$
- Signed Sammon Stress: $\frac{diff}{af}$
- Signed Relative Stress: $\frac{diff^2}{dm}$

MDS: Beispiel

Die Tabelle 8.1 enthält 7 Key-Value Paare. Als Distanz-Maß wurde die Euklidische Distanz gewählt. Abbildung 8.10 zeigt die einzelnen Stufen des MDS auf[9]. Als Einstieg werden die Daten per Zufallsgenerator auf der darstellenden Fläche verteilt, noch ohne Beachtung der Abstände zueinander (links oben). Aufgrund dieser, auf Zufallszahlen beruhenden Anfangsverteilung der Datenpunkte ist es nicht mehr möglich, ein MDS exakt zu wiederholen. Anschließend werden mit Hilfe des Kruskal-Algorithmus die Abstände zwischen den Punkten so berechnet, dass sie nun den Distanzen der jeweiligen

[8]Es sind auch weitere Algorithmen möglich, die hier genannten stehen in der OpenSource-Software *Orange* (http://www.ailab.si/orange/) zur Verfügung.

[9]Dieses und alle weiteren MDS-Untersuchungen wurden mit dem OpenSource-Programm Orange durchgeführt.

Key	Value
A	1
B	2
C	3
D	5
E	8
F	9
G	10

Tabelle 8.1: Sieben Key-Value-Paare

Key-Value-Paare entsprechen. Insgesamt wurden im Beispiel 800 Iterationen durchgeführt (rechts oben). Die Punkte werden annähernd in einer Linie angeordnet, wobei die beiden Gruppen <A, B, C> und <E, F, G> gebildet werden. Der Wert D befindet sich zwischen den beiden Gruppen, etwas näher an der Gruppe <A, B, C> gelegen. Die dritte Graphik (links unten) enthält zusätzlich Linien, die automatisch von der Software zwischen ähnlichen Elementen gezogen wurden. Die vierte Graphik (rechts unten) zeigt die Anwendung eines anderen Näherungsalgorithmus (Signed Relative Stress).

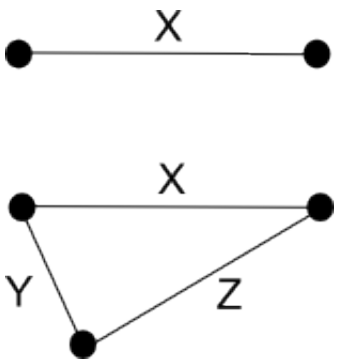


Abbildung 8.9: MDS im 2-dimensionalen Raum. Zwei Elemente mit der Distanz X können durch eine eindimensionale Linie visualisiert werden. Maximal können 3 Elemente mit den Abständen X, Y und Z zueinander in der zweidimensionalen Ebene exakt visualisiert werden. Bereits ein viertes Element benötigt einen dreidimensionalen Raum zur exakten Darstellung, jedes weitere Element würde eine weitere Dimension benötigen

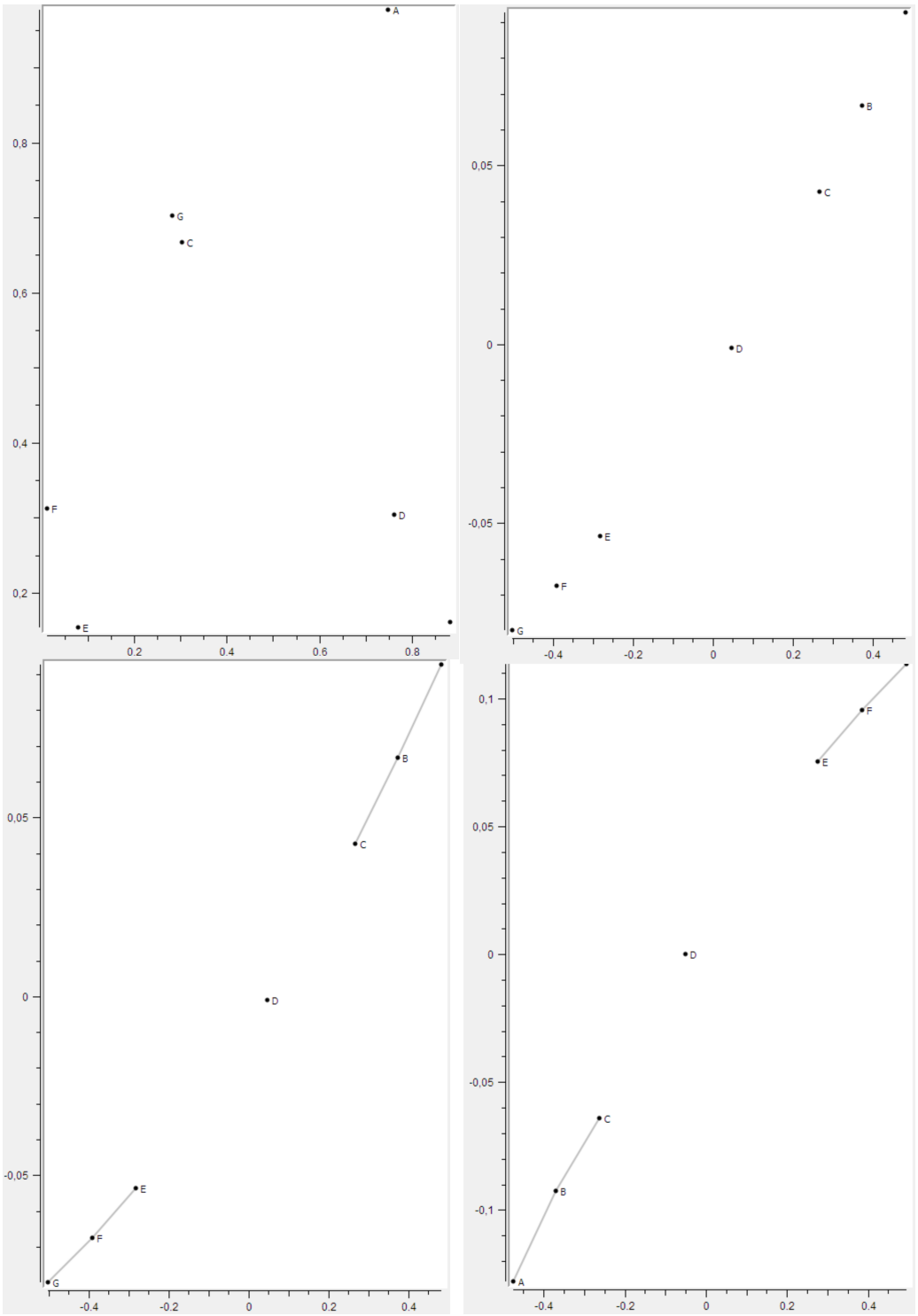


Abbildung 8.10: Multidimensional Scaling der Daten aus Tabelle 8.1

Kapitel 9

Visualisierung von dialektometrischen Ergebnissen mittels VDM

9.1 Visualisierung

9.1.1 Geographische Karten

Abhängig von der Anzahl der untersuchten Messpunkte können Ähnlichkeitsmatrizen wie oben beschrieben ein erhebliches Ausmaß annehmen. Eine Interpretation bzw. Analyse des umfangreichen Zahlenmaterials ist ohne (graphische) Hilfsmittel nicht oder nur noch schlecht möglich.

Eine Eigenschaft dialektaler Daten stellt die Tatsache dar, dass sie in Gestalt der Messpunkte geographisch fix distributiert sind. Anders gesagt, die Verteilung der Dialektdaten im Raum ist fest gegeben und als Parameter nicht veränderlich. Die räumliche, zweidimensionale Anordnung der Messpunkte zueinander stellt eine im Rahmen dialektometrischer Untersuchungen feste, nicht veränderbare Größe dar. Somit können geographische Karten als Grundlage weiterer Visualisierungen dienen.

Ausgangspunkt zur Erstellung dialektometrischer Karten ist eine *stille Karte*. Diese enthält lediglich die Grenzen des untersuchten Gebiets sowie optional weitere wichtige topographische Geländepunkte wie Gebirge, Flüsse

oder ähnliches.

Mit Hilfe der geographischen Koordinaten können Messpunkte auf einer stillen Karte auf zweierlei Weise visualisiert werden:

- **Punktkarte:** Auch wenn Messpunkte in Form von Städten ein erhebliches geographisches Ausmaß annehmen können, so werden sie in der Dialektometrie doch generell punktförmig behandelt. Abhängig von der ermittelten Kennzahl können diese Punkte nun in verschiedenen Farben oder Größen auf der stillen Karte eingetragen werden. Abbildung 9.1 zeigt eine solche Punktkarte für die Daten des Projekts Buldialects.

- **Voronoi-Karte**[1]**:** Hier dienen die Messpunkte jeweils als Zentrum für eine durch Polygone um den jeweiligen Messpunkt herum definierte Region. Diese kann anschließend in der der jeweiligen Kennzahl zugeordneten Farbe eingefärbt werden (siehe auch Abbildung 9.3 weiter unten).

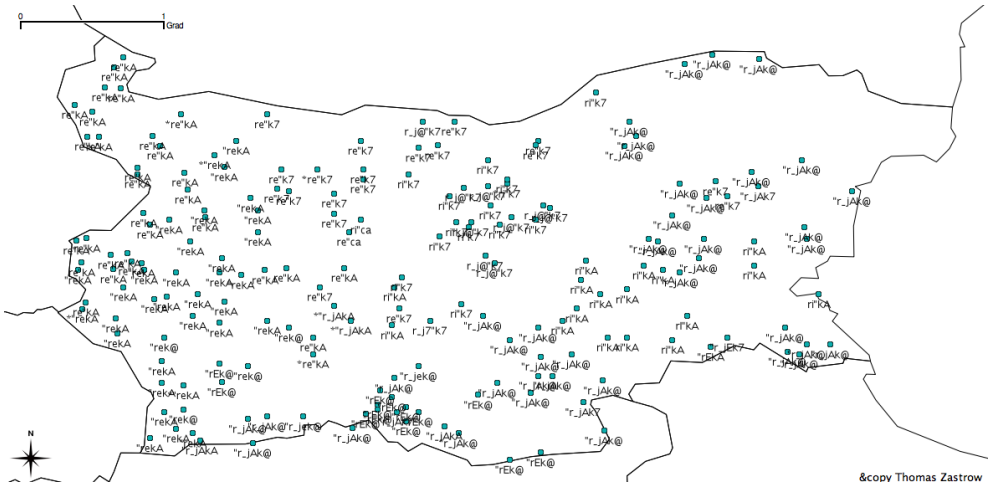


Abbildung 9.1: Eine Punktkarte

Gegenüber einer Punktkarte hat die Voronoi-Karte den Vorteil, die gesamte Fläche des jeweiligen Gebietes abzudecken und in den entsprechen-

[1]Nähere Informationen zur Mathematik der Voronoi-Diagramme finden sich unter http://de.wikipedia.org/wiki/Voronoi-Diagramm

den Farben einzufärben. Hierdurch lassen sich dialektale Strukturen wie Isoglossen oder Dialekt-Kontinua leichter visuell erkennen. Andererseits erfolgt die Einteilung des untersuchten Gebiets nach rein mathematischen Gesichtspunkten, ohne beispielsweise Geländemarken wie Gebirge, Flüsse etc. zu berücksichtigen. Hierdurch ensteht besonders in Gebieten mit einer geringen Anzahl an Messpunkten die Gefahr, dass Dialektareale über ihren eigentlichen Wirkungsgrad hinaus dargestellt werden. Auch unterschiedliche Abstände zwischen den Messpunkten können zu einer verzerrten Darstellung der Polygone führen.

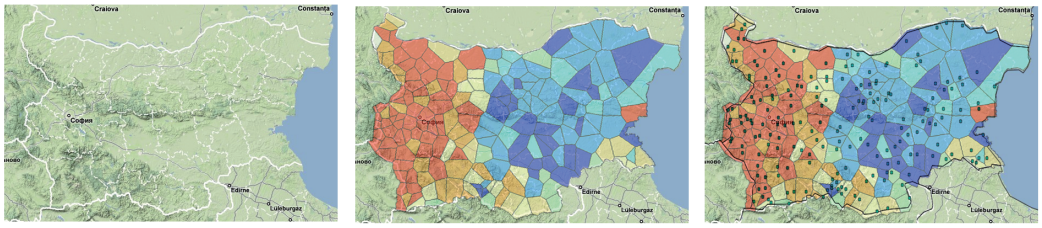


Abbildung 9.2: Beispiel für drei Ebenen in einer GIS-Anwendung, von links nach rechts: Stille Karte, Voronoi-Karte als weiterer Layer, Punkt-Karte als dritte Ebene

Geographische Informationssysteme (kurz GIS)[2] sind im Bereich der Geographie weit verbreitet und stellen ein wichtiges Werkzeug zur Erstellung und Bearbeitung geographischer Karten aller Art dar. Ausgehend von einer stillen Karte können in GIS durch Kombination mehrerer *Layer* dialektale Strukturen sukzessive visualisiert werden (siehe Abbildung 9.2).

9.1.2 Analyse mit VDM

Das Programm VDM ("Visual Dialectometry") wurde von Edgar Haimel an der Universität Salzburg erstellt[3]. Es bietet mannigfaltige Möglichkeiten,

[2]Beispielsweise QuantumGIS und GRASS, beide OpenSource. In dieser Arbeit kommt hauptsächlich QuantumGIS zum Einsatz.

[3]Nähere Informationen zu Entwicklungsgeschichte und Funktionsweise von VDM finden sich auf http://ald.sbg.ac.at/dm/germ/default.htm

dialektometrische Daten zu analysieren und zu visualisieren. In dieser Arbeit wird hauptsächlich von den Visualisierungsmöglichkeiten in VDM Gebrauch gemacht.

Als Grundlage für Visualisierungen benötigt VDM eine Projektdatei mit einer Voronoi-Karte des zu untersuchenden Gebietes. Abbildung 9.3 zeigt die hier verwendete Projektdatei mit der Voronoi-Karte für Bulgarien, eingeteilt in 197 Polygone, entsprechend den 197 Messpunkten des Projekts Buldialects[4].

Ist eine Projektdatei in VDM aktiv, kann anschließend eine Ähnlichkeitsmatrix hinzugeladen und auf diese die vielfältigen Möglichkeiten zur Visualisierung angewandt werden (Abbildung 9.4). Auf Synopsenkarten wird die geladene Ähnlichkeitsmatrix mittels Intervallalgorithmen visualisiert, ausgehend von den hier gesetzten Parametern können auch Strahlen- und Isoglossenkarten erstellt werden. Die Clusteranalyse bietet mehrere hierarchische Clusteralgorithmen, die anschließend als Dendrogramme bzw. auf Voronoi-Karten visualisiert werden können.

[4]Voronoi-Karten können nicht direkt in VDM erstellt werden: Stattdessen muss dies mittels externer Software erfolgen, beispielsweise bieten einige GIS-Programme diese Funktionalität. Anschließend kann die Voronoi-Karte im WMF-Format in eine VDM-Projektdatei importiert werden.

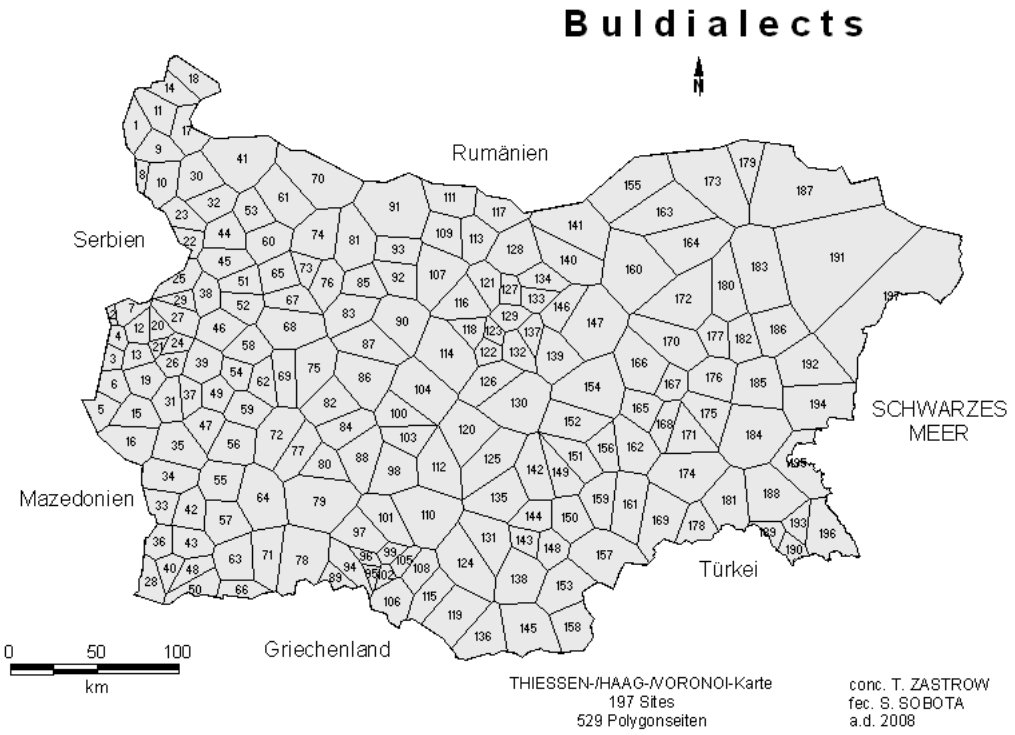


Abbildung 9.3: Leere Voronoi-Karte für das Buldialects-Projekt

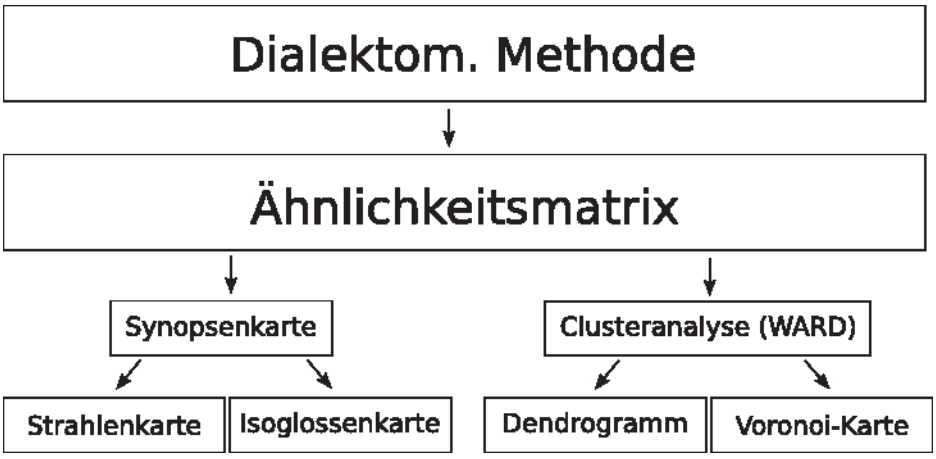


Abbildung 9.4: Analyse mit VDM

Kapitel 10

Bulgarische Dialekte

10.1 Bulgarien: Topographie

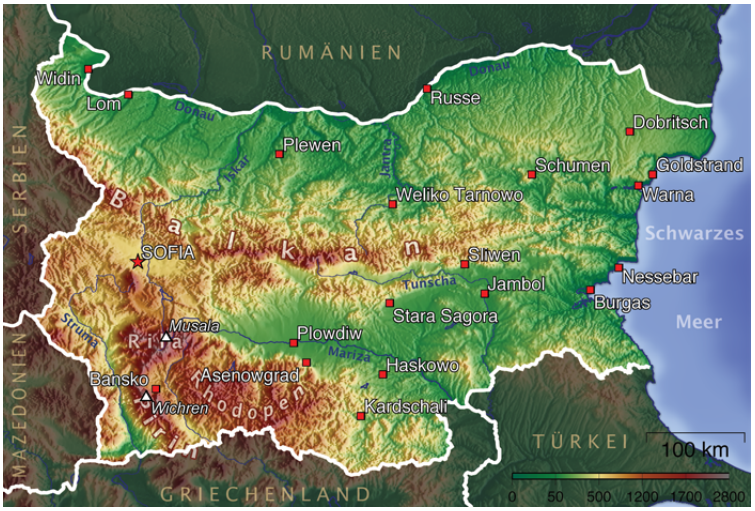


Abbildung 10.1: Das bulgarische Staatsgebiet, topographisch dargestellt.

Die Karte in Abbildung 10.1[1] zeigt das bulgarische Staatsgebiet und die angrenzenden Staaten. Die Hauptstadt Sofia befindet sich im Westen. Im Süden liegt sich das Mittelgebirge der Rhodopen, nördlich darüber das Balkan-

[1]Topographische Karte Bulgariens von
http://www.weltkarte.com/europa/bulgarien/topographie-bulgarien.htm
Lizenziert unter *GNU-Lizenz für freie Dokumentation.*

Gebirge. Zwischen den Gebirgen finden sich Tiefebenen. Im nördlichen Bereich des Balkan-Gebirges fließen die Flüsse Richtung Norden zur Donau hin, im südlichen Bereich und um die Rhodopen herum folgen sie dem Verlauf der Gebirge.

10.2 Die bulgarische Sprache

Bulgarisch ist eine indogermanische, südslawische Sprache mit enger Verwandtschaft zum Mazedonischen und Teilen des Serbischen (Gutschmidt, 2002, S. 229). Außerhalb der Grenzen Bulgariens wird sie von größeren Gruppen der Bevölkerung auch in Teilen der Türkei, Griechenlands, der Ukraine und Rumäniens gesprochen[2]. Sie ist neben Mazedonisch, Serbisch und Kroatisch Teil des *Balkansprachbundes*: Diese Sprachen haben aufgrund jahrhundertelangem Sprachkontakt und sonstigem politischem, kulturellem Austausch etc. sowie der geographischen Nähe zueinander große Ähnlichkeiten zueinander entwickelt. Der Balkansprachbund beinhaltet Sprachen aus verschiedenen Sprachfamilien: indogermanische (Albanisch), romanische (Rumänisch) und slawische Sprachen (Bulgarisch, Mazedonisch, Teile des Serbischen). Dabei sind nur die Sprachen innerhalb einer Sprachfamilie direkt miteinander verwandt, beziehungsweise gehen auf einen gemeinsamen Vorgänger zurück (Grey Thomason, 1999, S. 6).

Es ist zu erwarten, dass diese große Ähnlichkeit zu den Nachbarsprachen auch Einfluss auf die Strukturierung der bulgarischen Dialekte hatte. Da allerdings aus den umliegenden Ländern bzw. deren Sprachen keine mit dem bulgarischen Datensatz vergleichbaren Dialektdaten vorliegen, kann dieser Einfluss momentan nicht mit den hier beschriebenen dialektometrischen Mitteln nachgewiesen werden.

Als Schriftsystem benutzt das Bulgarische eine Variante des kyrillischen Alphabets (Gutschmidt, 2002, S. 223). Bulgarisch ist die älteste Sprache, die nachweislich Kyrillisch in der Schriftsprache benutzt. Historisch wird die bulgarische Sprache in Altbulgarisch (9.-11. Jahrhundert), Mittelbulgarisch

[2]Viele Sprachwissenschaftler konstatieren eine besondere Nähe zwischen dem Bulgarischen und dem Mazedonischen. So klassifizieren bulgarische Sprachwissenschaftler die mazedonische Sprache als bulgarische und umgekehrt.

(12.-14. Jahrhundert) und Neubulgarisch (ab dem 15. Jahrhundert) unterteilt.

In Bulgarien ist Bulgarisch als Erstsprache mit 84,5% der Sprecher die vorherrschende und einzige Amtssprache[3]. 9,6% der Bevölkerung sprechen Türkisch als erste Sprache, 4,7% Romani, eine indischstämmige Sprache der Roma und Sinti sowie 1,8% andere Sprachen (Abbildung 10.2). Eine weitere Gruppe bilden mit ca. 250.000 die muslimischen Pomaken. Diese leben im Süden Bulgariens, sprechen aber im Gegensatz zu dem Gros der muslimischen Bevölkerung Bulgarisch und nicht Türkisch. Die Gesamtbevölkerung Bulgariens beträgt ca. 7,3 Millionen. Zusammen mit im Ausland lebenden Bulgaren kommt das Bulgarische auf ca. 8,8 Millionen Erstsprecher (Gutschmidt, 2002, S. 219). Davon leben ca. 1,2 Millionen in der Hauptstadt Sofia.

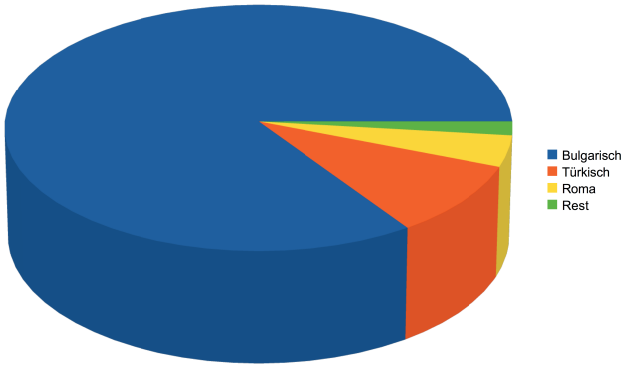


Abbildung 10.2: Sprachen in Bulgarien

Derzeit schrumpft die Bevölkerung Bulgariens jedes Jahr um ca. 0,83%, was im Vergleich zu den umgebenden Nachbarländern ein sehr hoher Wert ist (Rumänien -0,127%, Mazedonien +0,263%, Griechenland +0,163%). Es ist zu erwarten, dass dieser massive Bevölkerungsrückgang in Zukunft auch

[3]Diese und alle anderen demographischen Zahlen stammen, sofern nicht anders angegeben, aus dem CIA World Factbook, https://www.cia.gov/library/publications/the-world-factbook/index.html, eingesehen am 19.2.2008.

Auswirkungen auf die sprachliche Strukturierung Bulgariens haben wird.

10.3 Dialektologie

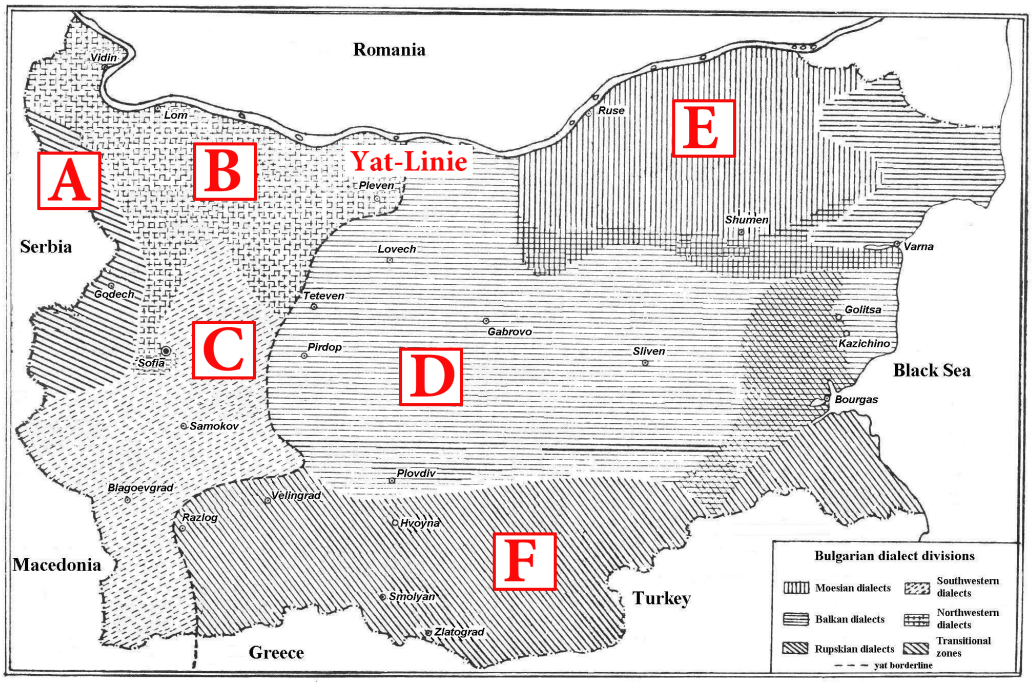


Abbildung 10.3: Einteilung der bulgarischen Dialektgebiete nach Stojko Stojkov, rote Markierungen hinzugefügt

Seit den 1950er Jahren untersuchte Prof. Stojko Stojkov an der Universität Sofia die Dialekte des Bulgarischen (siehe hierzu auch Prokic u. a. 2009). Abbildung 10.3 aus Stojkov (1962) zeigt die dialektale Einteilung Bulgariens durch Stojko Stojkov mit den Methoden der Dialektologie. Die *Yat-Linie* als stärkste Isoglosse teilt Bulgarien in einen West- und einen Ost-Teil[4]. Sie beginnt im Norden Bulgariens in der Nähe von Pleven und verläuft in südlicher Richtung durch das gesamte bulgarische Staatsgebiet. Auf Höhe des

[4]Auf Stojkovs Karte setzt sie sich fort nach Süden in Richtung Griechenland und Mazedonien, was auf eine weitergehende, das bulgarische Staatsgebiet überschreitende Verbreitung hindeutet.

Rhodopen-Gebirges ändert sich ihr Lauf nach Westen, um am Ende der Rhodopen wieder in Richtung Süden zu verlaufen. Das Gebiet der Rhodopen ist somit den östlichen Dialekten zuzurechnen.

Innerhalb der durch die Yat-Linie vorgenommene Grobeinteilung in einen West- und einen Ostteil finden sich insgesamt sechs weitere Dialektareale. Diese sind mittels diskreter Isoglossen voneinander abgetrennt, allerdings existieren im Ostteil an den Rändern der Dialektareale Zonen mit unklarer Zugehörigkeit bzw. gemischten Dialekten. Diese heterogenen Gebiete werden in der Legende nicht als eigenständige Dialektgebiete aufgeführt.

Die sechs klar voneinander abgegrenzten Dialektgebiete im Einzelnen:

- An der Westgrenze Bulgariens zu Serbien finden sich *Übergangs-Dialekte* mit großer Ähnlichkeit zur serbischen Sprache (auf der Abbildung 10.3 Bereich A).

- Der übrige Teil Westbulgariens ist in einen *Nord-* und einen *Südteil* aufgeteilt (Bereich B bzw. C). Die Hauptstadt Sofia liegt genau an der Grenze zwischen diesen Dialektgebieten, sie befindet sich faktisch in einer Ausbuchtung des nordwestlichen Dialektgebietes.

- Im mittleren Teil Ostbulgariens befinden sich die Dialekte des *Balkan-Gebirges* (Bereich D).

- Im Nordosten befinden sich die *Moesischen Dialekte*. Sie unterbrechen die Balkan-Dialekte, die hier im Osten den größten Teil des Gebietes ausmachen und sich östlich von den moesischen Dialekten weiter bis zum Schwarzen Meer hin erstrecken (Bereich E).

- Das Gebiet der Rhodopen mit den *Rupskian-Dialekten* schließt Bulgarien nach Süden hin ab. Hier ist festzuhalten, dass Stojkovs Einteilung nicht exakt dem topographischen Verlauf des Rhodopen-Gebirges entspricht (Bereich F).

In Abbildung 10.4 ist die dialektologische Einteilung der bulgarischen Dialekte nach Stojko Stojkov auf die Polygone der in dieser Arbeit verwendeten Voronoi-Karte übertragen worden. Da die Polygone der Voronoi-Karte

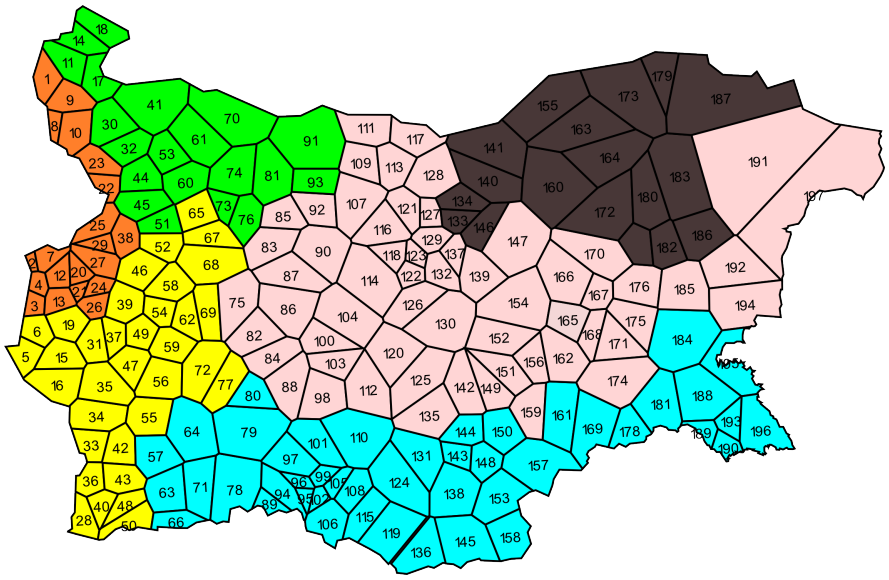


Abbildung 10.4: Übertragung der Stojkovschen Dialekteinteilungen auf die Polygone der Voronoi-Karte

jeweils die Fläche um die analysierten Messpunkte herum markieren, ist eine manuelle 1:1-Umsetzung der Stojkov'schen Karte auf die Voronoi-Karte nicht möglich. Dies betrifft insbesondere den exakten Verlauf der Grenzen zwischen den Dialekten sowie die von Stojkov eingezeichneten Übergangsgebiete im Osten Bulgariens. In diesen Gebieten sind die angrenzenden Dialekte nicht klar voneinander zu trennen, allerdings handelt es sich auch nicht um eigenständige Dialektareale, sondern um Mischungen aus den sich direkt berührenden Dialekten. In der Übertragung auf die Voronoi-Karte werden diese Übergangsbereiche deswegen nicht als eigenständige Dialekte aufgeführt, sondern entweder dem einen oder dem anderen angrenzenden Dialektgebiet zugeordnet. Hier liegt ein Nachteil der diskreten Dialekteinteilung mittels Isoglossen: Kontinuierliche Übergänge zwischen Dialektgebieten lassen sich nur schwer bzw. gar nicht realisieren. Die Tabelle 10.1 weist die Anzahl an Messpunkten des Buldialect Datensatzes im jeweiligen Dialektgebiet nach Stojkovs Einteilung aus.

Abbildung 10.5 zeigt die Isoglossen auf, die zu der oben beschriebenen sechsteiligen Einteilung der bulgarischen Dialektareale führen (manuell nach-

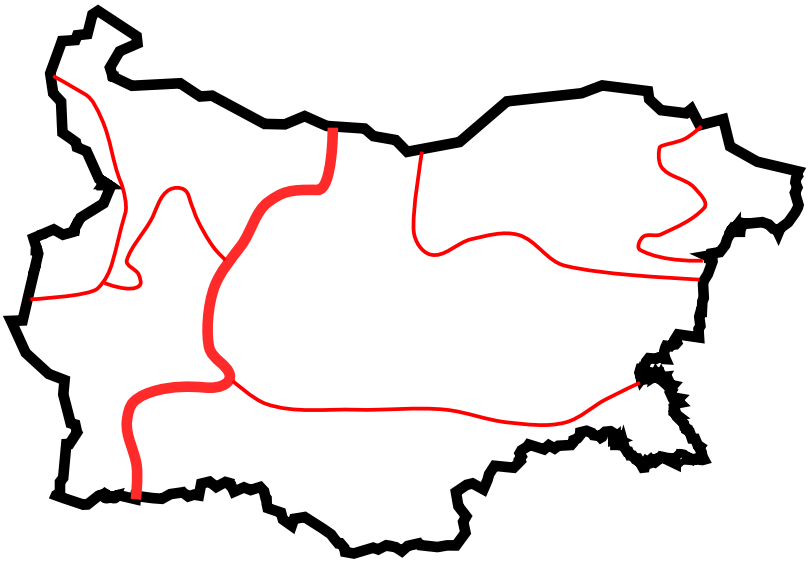


Abbildung 10.5: Übertragung der Isoglossen aus Abbildung 10.3 auf die Stille Karte, auf welche ebenfalls die Polygone der Voronoi-Karte eingezeichnet wurden

Dialekt-Nr.	Beschreibung	Anzahl Messpunkte
1	Übergangsdialekte zum Serbischen	20
2	Nord-West Dialekte	20
3	Süd-West Dialekte	34
4	Rupskian Dialekte	45
5	Balkan Dialekte	60
6	Moesische Dialekte	18

Tabelle 10.1: Anzahl der Messpunkte des Buldialect Datensatzes, aufgeteilt auf Stojkovs sechs Dialektgebiete

gezeichnet). Die Einteilung in Polygone ist in dieser Abbildung nicht berücksichtigt. Die Bereiche mit gemischten Dialekten im Osten sind ebenfalls nicht gesondert aufgeführt, stattdessen zeigen die eingezeichneten Isoglossen einen Mittelweg auf. Die Yat-Linie ist dicker eingezeichnet.

10.4 Das Projekt Buldialects

Das Projekt Buldialects ist ein Kooperationsprojekt zwischen der Universität Groningen, der Bulgarian Academy of Science, der Universität Sofia und der Universität Tübingen. Finanziert wird es von der Volkswagen Stiftung[5]. Der Projektantrag "Measuring linguistic unity and diversity in Europe" (Hinrichs u. a., 2005) definiert als Aufgabe des Projekts, erprobte quantitative Technologien zur Messung von Sprachdiversifikation auf die bulgarische Sprache anzuwenden.

Zu diesem Zweck wurde an der Universität Sofia - in Zusammenarbeit mit der Bulgarian Academy of Science - ein Dialektdatensatz des Bulgarischen erstellt. Enthalten sind sowohl phonetische als auch lexikalische Dialektdaten der bulgarischen Sprache, die getrennt voneinander in unterschiedlichen Teilen des Datensatzes vorliegen. Auf diese formalisierten Daten können sowohl etablierte als auch neue Methoden der Dialektometrie angewendet und die

[5]Weitergehende Informationen wie der Projektantrag, Termine etc. finden sich auf der Homepage des Projekts: http://www.sfs.uni-tuebingen.de/dialectometry/

Ergebnisse anschließend analysiert und visualisiert werden.

Der bulgarische Datensatz wurde von Prof. Vladimir Zhobov und seinem Team an der Universität Sofia zusammengestellt. Dabei wurden hauptsächlich folgende Quellen benutzt (Zhobov, 2006):

1. Studentische Abschlussarbeiten. Diese wurden in der Zeit zwischen den 1950ern und 1985 unter der Leitung von Prof. Stoyko Stoykov sowie zwei seiner ehemaligen Studierenden durchgeführt. Jede dieser Abschlussarbeiten beschreibt einen Messpunkt, meistens den Heimatort des jeweiligen Studierenden. Zur phonetischen Transkription entwickelte Stoyko Stoykov ein eigenes Lautsystem, welches allerdings eins zu eins in das heute international gebräuchliche IPA überführt werden kann.

2. Die Buchreihen *Bulgarian Dialectology. Investigations and Data* und *Studies in Bulgarian Dialectology.* Sie beschreiben einige der Messpunkte ausgiebig. In beiden Buchreihen kommt das von Stoyko Stoykov entwickelte Lautsystem zum Einsatz.

3. In das seit den 1950er Jahren von Stoyko Stoykov gepflegte *Ideographic Dictionary of Bulgarian Dialects* flossen neben den bereits erwähnten studentischen Abschlussarbeiten auch in Feldarbeit erworbene Daten ein. Ebenfalls enthalten ist das Archiv der dialektologischen Abteilung des Bulgarian Language Institute. Zur Zeit findet eine Digitalisierung der ca. 2 Millionen Karteikarten statt.

4. Seit 1981 wurden ca. 250 Stunden Audiomaterial in ca. 100 Messpunkten aufgenommen und liegen in Form von Audiokassetten vor.

5. Der *Bulgarian Dialect Atlas* wird zur Überprüfung des neu erstellten Materials herangezogen.

Der Buldialects Datensatz beschränkt sich auf die bulgarischsprachigen Dialekte Bulgariens. Die Minderheitensprachen Bulgariens, wie beispielsweise das Türkische und die Sprache der Roma und Sinti, sind nicht miterfasst

worden. Ebenfalls außen vor geblieben sind die Dialekte der nicht in Bulgarien lebenden bulgarischen Muttersprachler[6].

Format

Die an der Universität Sofia erhobenen Daten wurden an der Bulgarian Academy of Science elektronisch aufbereitet und in ein valides XML-Format[7] gebracht. Hierzu kommt das ClarK-System zum Einsatz (Osenova u. Simov, 2005).

An der Universität Tübingen wurden die XML-Dateien in die freie XML-Datenbank eXist[8] eingelesen. Die weitere Verarbeitung und Analyse erfolgt mittels selbstgeschriebener Java-Programme, die entweder direkt auf die eXist-Datenbank zugreifen oder die Daten in der benötigten Form exportieren. Unter anderem wird der Datensatz in ein für die Groninger L04-Software verarbeitbares Format gebracht (Nerbonne u. a., 1999).

Der Datensatz ist nach Messpunkten organisiert: Jede Datei enthält alle Wörter und Varianten eines Messpunktes. Dies erlaubt direkt dateiweise Untersuchungen in der SSAW-Richtung. Analysen in der SWAS-Richtung benötigen eine vorangehende Konvertierung der Daten in diese Richtung.

```
<DialectData>
    <site>
        <num>2069</num>
        <name>aldomirovci, slivnica</name>
        <entry id="2069-3">
            <key>агне</key>
            <english>lamb</english>
            <cform ana="Ncnsi">|агне</cform>
            <nform>`agne</nform>
            <variant ana="Ncnsi">`jagne</variant>
            <sampa>
                <nform>"agne</nform>
                <variant ana="Ncnsi">"jAgne</variant>
            </sampa>
        </entry>
```

Abbildung 10.6: Messpunkt Aldomirovci, Wort “Schaf”

[6]Beispielsweise das *Banater Bulgarisch*, eine in Südrumänien von ca. 22.000 Personen gesprochene Variante des Bulgarischen (Dulicenko, 2002, S. 203 ff.).

[7]Zur Validierung wird eine *Document Type Definition* benutzt.

[8]http://exist.sourceforge.net/

Neben den einzelnen Wörtern und deren Varianten enthält der Datensatz einige weitere Informationen (siehe Abbildung 10.6):

- **num:** Die Nummer des Messpunktes, entspricht der Nummer im *Bulgarian Dialect Atlas*
- **name:** Name des Messpunktes

Für jedes Wort der Wortliste:

- **id:** Eine eindeutige (XML-) ID für jeden Eintrag (als Attribut realisiert)
- **key:** Das Wort in kyrillischer Schreibweise ohne Akzent
- **english:** Englische Übersetzung
- **cform:** Das Wort in kyrillischer Schreibweise mit Akzent. Im Attribut "ana" sind morphologische Informationen enthalten
- **nform:** Die transliterierte Form
- **variant:** Die dialektale Variante in diesem Messpunkt
- **sampa / nform:** Normalisierte Form
- **sampa / variant:** Die messpunktspezifische, dialektale Variante des Wortes in XSampa-Codierung

Kapitel 11

Analysen der phonetischen Dialektdaten

In diesem Kapitel sollen die in den vorangegangenen Kapiteln vorgestellten Methoden auf den phonetischen Teil des bulgarischen Datensatzes angewendet werden. Allen folgenden Analysen liegen dieselben Daten zugrunde: Für die 197 Messpunkte des Datensatzes wurden insgesamt 119 Wörter ausgewählt (Tabelle A.1 im Anhang A listet eben diese verwendeten 119 Wörter in kyrillischer Schreibweise, mit deutscher Übersetzung und Wortart auf). Der bulgarische Datensatz enthält eine größere Anzahl verschiedener Wörter, allerdings weisen einige von ihnen Lücken in manchen der dialektalen Varianten auf. Da diese Lücken die Ergebnisse nicht verfälschen sollen, wurden lediglich die 119 Wörter, deren Daten für alle Messpunkte vollständig vorliegen, ausgewählt.

Für einige Wörter existieren für einige Messpunkte mehrere phonetische Varianten. Auch hier wurde, um Verfälschungen zu vermeiden, immer nur eine Variante in der Analyse berücksichtigt.

Die meisten der im folgenden aufgeführten Methoden produzieren als Ergebnis eine symmetrische Ähnlichkeitsmatrix, die mit der VDM-Software[1] analysiert und visualisiert werden kann. VDM bietet eine Vielzahl von Analyse- und Visualisierungsmethoden, wovon die meisten wiederum durch

[1] Eine Konvertierung der VDM-Matrizen in das von der Groninger L04-Software benötigte Datenformat ist ebenfalls möglich.

Parameter weitergehend spezifiziert werden können. Da es aus praktischen Gründen nicht möglich ist, hier alle Varianten und Visualisierungen darzustellen, wurden diese beschränkt auf:

- *Hierarchisches Clustering*: Methode WARD mit 12 Clustern, zusammen mit dem entsprechendem Dendrogramm (siehe Kapitel 8.2, Abbildung 8.3). Die vertikale Linie im Dendrogramm markiert die vorgenommene Einteilung in 12 Cluster. Prinzipiell könnten hier auch andere Einteilungen wiedergegeben werden. Für die Einteilung in genau 12 Cluster sprechen folgende Argumente: Die Synopsenkarte (siehe nächster Punkt) kann ebenfalls in 12 Klassen eingeteilt werden. Desweiteren hat Stojko Stojkov die bulgarischen Dialekte in 6 Hauptareale eingeteilt (siehe Kapitel 10.3). 12 Cluster als Vielfaches hiervon erlauben einen Vergleich, der auch die Gebiete mit vermischten Dialekten entlang den von Stojkov eingezeichneten Isoglossen mit einbezieht.

- *Synopsenkarte* auf der Basis des MinMWMax-Algorithmus mit einer Klassenzahl von ebenfalls 12, dazu das entsprechende Histogramm. Als Referenzpunkt ist Messpunkt 1 (Rakovica, der westlichste Messpunkt) ausgewählt (siehe Kapitel 8.1)

- *Isoglossenkarte*, ebenfalls basierend auf dem MinMWMax-Algorithmus und einer Klassenzahl von 12

Alle drei Analysen werden jeweils auf der Grundlage einer Voronoi-Karte dargestellt. Beim Clustering und der Synopsenkarte erfolgt dies durch flächiges Einfärben der Polygone, auf der Isoglossenkarte markieren die Dicke und die Farbe der Polygongrenzen zueinander den Abstand bzw. die Nähe der jeweils benachbarten Dialekte. Die Analysen und deren Visualisierungen wurden in der VDM-Software (siehe Kapitel 9) durchgeführt.

Die Verwendung einheitlicher Analyse- und Visualisierungemethoden gewährleistet eine Vergleichbarkeit der mit den verschiedenen dialektometrischen Methoden erzielten Ergebnisse. Hinzu kommen methodenspezifische Diagramme, Karten etc. und ein Vergleich der in dieser Arbeit vorgestellten Methoden mit den Ergebnissen der bulgarischen Dialektologie. Ab-

Wortart	Anzahl
Nomen	74
Adjektiv	12
Adverb	5
Verb	11
Pronomen	9
Präposition	4
Numeral	2
Artikel	1
Partikel	1

Tabelle 11.1: Verteilung der verwendeten 119 Wörter auf die Wortarten

schließend sollen in diesem Kapitel einige weitere Analyse- und Visualisierungsmethoden dargestellt werden.

11.1 Grundlegende statistische Kennzahlen

11.1.1 Wortebene

Für die oben genannten 119 Wörter in 197 Messpunkten wird jeweils eine dialektale Variante verwendet. So ergibt sich eine Summe von 23.443 Varianten (siehe auch Tabelle A.1 im Anhang A). Tabelle 11.1 schlüsselt die verwendeten Wörter nach Wortarten auf: Es überwiegen mit weitem Abstand die Substantive, gefolgt von Adjektiven und Verben.

11.1.2 Phon-Ebene

Die in Tabelle 11.1 aufgeführten 23.443 Varianten bestehen insgesamt aus 127.520 XSampa-Codes. Im Durchschnitt ergibt sich eine Anzahl von ca. 5.4 XSampa-Codes pro Variante, wobei nicht weiter nach diakritischen Zeichen oder anderen Phonen unterschieden wird. Die längste Variante besteht aus 14 XSampa-Codes.

Die Tabelle A.2 im Anhang A enthält eine Übersicht aller im Datensatz verwendeten XSampa-Codes, aufgeschlüsselt nach der Frequenz ihres Vorkommens. Der häufigste XSampa-Code ist der Wortakzent mit 22.049 Vorkommen. Es folgen die drei Vokale e, A und o, das i folgt nach zwei Konsonanten. Zwischen i und u liegen weitere 8 Konsonanten. Abbildung 11.1 zeigt die Häufigkeiten aller XSampa Codes *ohne* den Akzent. Klar sichtbar ist eine in die monoton fallende Kurve eingebettete Treppenstruktur, die eine inhärente Gruppierung der XSampa-Codes erkennen lässt.

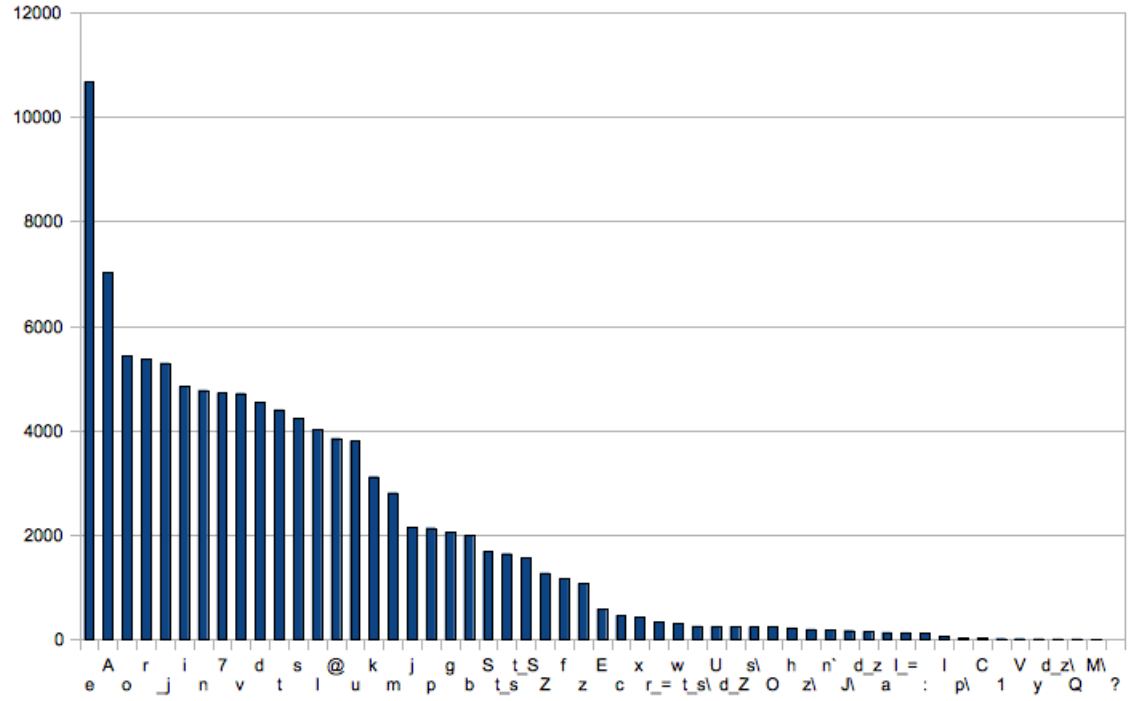


Abbildung 11.1: Frequenzen aller XSampa Codes

11.1.3 Korrelationskoeffizienten

Mittels Korrelationskoeffizienten lässt sich feststellen, inwieweit ein statistischer Zusammenhang zwischen zwei Datenreihen besteht. Dabei stellt die eine Datenreihe erwartete, vorgegebene Werte und die andere Datenreihe empirisch ermittelte Werte dar. In der Dialektometrie können diese Metho-

den eingesetzt werden, um etwaige Korrelationen zwischen der Hochsprache (erwartete Daten) und den Dialektdaten (gemessene Daten) zu ermitteln. Je höher der Korrelationskoeffizient zwischen der Hochsprache und einem Dialekt, desto ähnlicher sind sich die beiden Datensätze. Zur Berechnung des Korrelationskoeffizient gibt es zwei weit verbreitete Standardverfahren.

Der *Bravais-Pearson Korrelationskoeffizient* r berechnet sich wie folgt[2]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{11.1}$$

Im Gegensatz zum Bravais-Pearson Korrelationskoeffizient berechnet der *Rangkorrelationskoeffizient von Spearman* die Korrelation zweier Datenreihen nicht direkt auf Grundlage der gemessenen Werte. Stattdessen werden die Daten in eine diskrete Anzahl Ränge eingeteilt und anschließend die Korrelation r_{sp} zwischen diesen Rängen berechnet:

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n}(R_i - R_i')^2}{(n-1)n(n+1)} \tag{11.2}$$

Für alle 197 Dialekte des bulgarischen Datensatzes und die Hochsprache wurden die Frequenzen der enthaltenen XSampa-Codes berechnet. Anschließend wurden die Korrelationskoeffizienten der einzelnen Dialekte in Relation zur Hochsprache berechnet. Tabelle 11.2 zeigt einige signifikanten Werte der jeweiligen Datenreihen. Das Diagramm in Abbildung 11.2 visualisiert die gesamten Korrelationskoeffizienten, aufgetragen von West nach Ost.

Alle Werte liegen im Bereich > 0.5, was auf eine gegebene Korrelation zwischen der Hochsprache und allen untersuchten Dialekten schließen lässt. Die Kurven der beiden Korrelationskoeffizienten verlaufen ähnlich, aber nicht identisch. Auffallend ist wiederum der relativ ruhige Kurvenverlauf im westlichen Bereich der bulgarischen Dialekte, der im weiteren Verlauf nach Osten durch das Hinzukommen der heterogenen Dialekte in den Rhodopen wesentlich unruhiger wird.

[2]Formeln aller Korrelationskoeffizienten nach Bamberg u. Baur (2008), S. 35 ff. In der Programmierung kam die Implementation der Apache Commons Math Library zum Einsatz (http://commons.apache.org/math/).

	Pearson	Spearman
Minimum	0.7	0.56
Arithemt. Mittel	0.82	0.72
Maxmimum	0.94	0.84

Tabelle 11.2: Korrelationskoeffizienten der bulgarischen Dialekte in Relation zur bulgarischen Hochsprache: Einige ausgewählte Werte

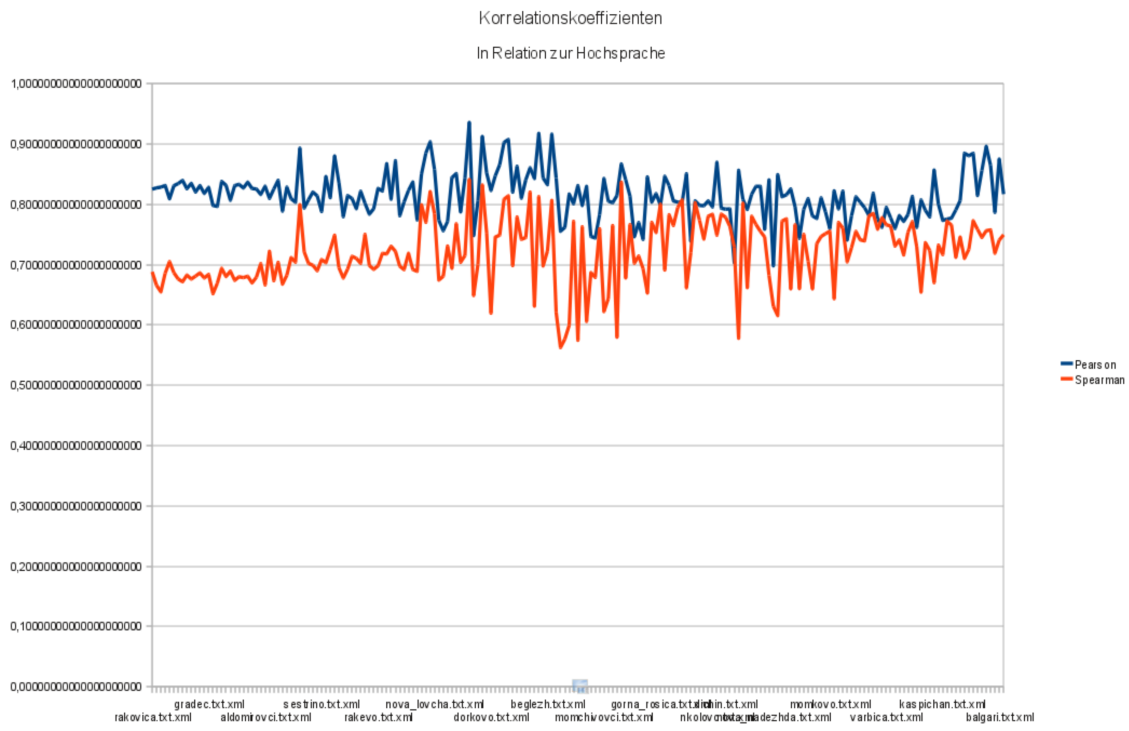


Abbildung 11.2: Korrelationen aller Dialekte zur Hochsprache, aufgetragen auf der Y-Achse von West nach Ost (blau: Pearson, orange: Spearman)

11.2 Informationstheorie

Die bereits erläuterten informationstheoretischen Methoden (siehe Kapitel 7) können nun zur Analyse von phonetischen Dialektdaten auf der Basis von XSampa-Codes herangezogen werden. Mit informationstheoretischen Methoden kann der Informationsgehalt eines Datensatzes bestimmt werden. Anschließend kann dieser Informationsgehalt auf Teile des Datensatzes aufgeteilt und diese Informationswerte dann in Relation zueinander gesetzt werden. Auf Dialektdaten übertragen bedeutet dies, dass die Unterschiede zwischen Dialekten in der Höhe des jeweils enthaltenen Informationsgehalts liegen. Abbildung 11.3 verdeutlicht dies anhand von Farbverläufen: Im linken Bereich sind die Farbverläufe von Messpunkt 1 und 2 identisch, diese Elemente haben somit bezogen auf den gesamten Datensatz eine höhere Frequenz und tragen dementsprechend weniger Information. Im rechten Bereich der Verläufe steigen die Unterschiede (blaue bzw. grüne Elemente) zwischen den Messpunkten und somit auch der Informationsgehalt der einzelnen Elemente.

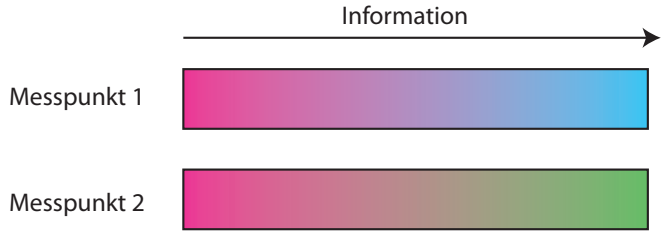


Abbildung 11.3: Dialektale Unterschiede zwischen zwei Messpunkten, farblich dargestellt: Von links (rot) nach rechts nehmen die Unterschiede zwischen den Dialekten zu (blau bzw. grün). Bezogen auf den gesamten Datensatz sind die roten Elemente häufiger und tragen somit *weniger* Information als die grünen bzw. blauen Elemente

Diese unterschiedliche Verteilung der Informationswerte zwischen den einzelnen Messpunkten lässt sich in Bezug auf die Unterschiedlichkeit der jeweiligen Dialekte wie folgt interpretieren:

- Sind zwei Dialekte identisch, ergeben sich auch in Informationsgehalt und Entropiewerten keine Unterschiede.

- Unterscheiden sich zwei Dialekte, sei es auch nur in einem einzigen XSampa-Code, dann werden sich auch die Informations- und Entropie-Werte voneinander unterscheiden. Je größer die Unterschiede in den jeweils enthaltenen XSampa-Codes sind, desto größer werden sich auch die informationstheoretischen Werte voneinander unterscheiden.

Der oben genannte zweite Punkt trifft in einem Fall nicht zu: Wenn die Frequenzen zweier XSampa-Codes innerhalb zweier Dialekte exakt miteinander vertauscht sind und es keinerlei weiteren Unterschiede zwischen den Dialekten gibt. Dieser Fall sollte allerdings extrem unwahrscheinlich sein.

Damit die informationstheoretischen Werte mehrerer Datensätze miteinander verglichen werden können, müssen die Datensätze ungefähr die gleiche Größe aufweisen. Wie oben erwähnt, spiegeln sich erhebliche Größenunterschiede in den Datensätzen nicht unbedingt in den jeweiligen Entropie-Werten wider, wohingegen ein im Vergleich erheblich größerer Datensatz auch immer mehr Information in sich trägt.

Alle informationstheoretischen Methoden basieren auf der Wahrscheinlichkeit, mit der die einzelnen Elemente im Datensatz vertreten sind. Aus deren Berechnung bzw. Aggregierung ergeben sich unterschiedliche Varianten, von denen hier einige angewendet werden sollen:

- Selten auftretende Elemente tragen mehr Information als häufig auftretende. Deswegen beinhalten Messpunkte, die viele seltene Elemente beinhalten, mehr Information als Messpunkte, die weniger seltene Elemente beinhalten.

- Die **Entropie** gibt die Informationsdichte eines Messpunktes an. Hier sollen die einzelnen Messpunkte bzw. die Wahrscheinlichkeiten der enthaltenen Elemente individuell in Bezug auf den jeweiligen Messpunkt berechnet werden. Die Entropien der einzelnen Messpunkte lassen sich im Gegensatz zur Information nicht aufaddieren und können somit nicht in direkte Relation zueinander gesetzt werden. Die Entropie eines Messpunktes ist somit ein Kennwert für die messpunktspezifische

	Mittelwert	Median	Varianz	Stand.Abw.	Variat.-Koeff.
Information	3003,67	3023,82	32511,51	180,31	0,06
Entropie	4,45	4,47	0,01	0,1	0,02
Ent. Pos.	6,49	6,49	0,02	0,14	0,02

Tabelle 11.3: Statistische Kennwerte der informationstheoretischen Methoden

Relation zwischen Elementen die viel, und Elementen, die wenig Information tragen. Dementsprechend wird die Entropie auf Grundlage der Wahrscheinlichkeiten *innerhalb* des jeweiligen Datensatzes und nicht in Bezug zur Gesamtheit der Daten berechnet.

- Um die **Stellung der Elemente** im Wort mitberücksichtigen zu können, wird das einzelne Elemente ergänzt um seine Position im Wort. Ein *a* an erster Stelle im Wort wird somit anders gewertet als ein *a* an zweiter Stelle usw.

11.2.1 Anwendung auf die bulgarischen Dialektdaten

Tabelle 11.3 führt einige statistische Kennzahlen für die informationstheoretischen Methoden auf. Die Werte der Entropie (dritte Zeile) und die der Entropie unter Einbeziehung der Position (vierte Zeile) liegen in ähnlichen Wertebereichen, die Information liegt weit darüber. Varianz, Standardabweichung und Variationskoeffizient liegen bei den beiden Entropie-Analysen sehr eng beieinander. Dementsprechend hat die Position der Elemente innerhalb der Wörter wenig Einfluss auf die prinzipiell lediglich quantitativ definierten Entropiewerte.

11.2.2 Information

Die Abbildung 11.4 zeigt den Informationsgehalt aller im Datensatz vorkommenden XSampa-Elemente[3]. Der Informationsgehalt sinkt von links nach

[3]Die Berechnungen wurden mit 32-Bit-Double-Genauigkeit durchgeführt. Zur besseren Lesbarkeit werden die Werte hier nur mit zwei Nachkommastellen angegeben.

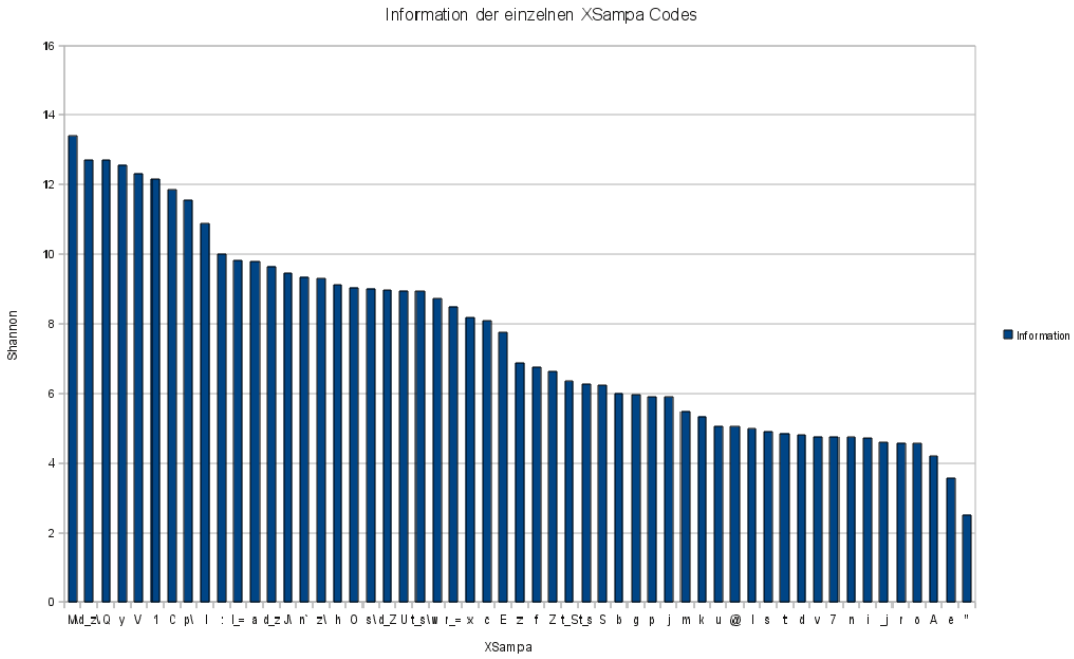


Abbildung 11.4: Informationsgehalt der einzelnen XSampa-Codes

rechts, wobei gleichzeitig die Wahrscheinlichkeit der XSampa-Codes zunimmt. Häufig vorkommende XSampa Codes wie beispielsweise die Vokale finden sich allesamt im rechten Teil des Diagramms. Das häufigste und damit am wenigsten Information tragende Element ist der Wortakzent mit ca. 2.53 Shannon pro Auftreten. Da der Akzent in nahezu jedem XSampa-kodierten Wort ein Mal enthalten ist und seine Position im Wort bei der Berechnung der Information unberücksichtigt bleibt, hat er so gut wie keinen Einfluss auf die Relationen der Informationsgehalte der einzelnen Messpunkte zueinander. Dem Akzent gegenüber steht der XSampa Code M\ (stimmhafter, velarer Approximant, in IPA-Notation: ɰ) mit einem maximalen Informationsgehalt von ca. 13.38 Shannon. Der Mittelwert der Information liegt bei ca. 7.62 Shannon, was in etwa dem XSampa Code E (in IPA ɛ) entspricht (7.75 Sh).

Auffällig ist wiederum die Treppenstruktur des Diagramms, die auf eine Kategorisierung der XSampa-Codes in zusammenhängende Blöcke hinweist.

Ausgehend von den in Abbildung 11.4 dargestellten Informationswerten der einzelnen XSampa-Codes kann nun der Informationsgehalt der einzelnen Messpunkte berechnet werden. Aufsummiert ergibt sich der Informationsgehalt des gesamten bulgarischen Dialekt-Datensatzes: Er beträgt ca. 591.722,25 Shannon. Der Messpunkt mit dem geringsten Informationsgehalt ist Varvara (2774,75 Sh), den größten Anteil an der Information trägt Ustovo mit 3523,32 Shannon.

Das Diagramm 11.5 visualisiert den Informationsgehalt der einzelnen Messpunkte, allerdings nicht der Größe nach, sondern ihrer geographischen Lage nach von West nach Ost. Im westlichen Teil Bulgariens (linke Seite des Diagramms) ist der Informationsgehalt der Messpunkte relativ niedrig und homogen verteilt, nach Osten hin steigt der Informationsgehalt an und der Kurvenverlauf wird unruhiger. Hierfür sind unter anderem die heterogenen Dialekte der Rhodopen mitverantwortlich. Im östlichen Teil Bulgariens (rechte Seite des Diagramms) herrschen weiterhin heterogene Werte vor, allerdings nicht ganz so stark unterschiedlich ausgeprägt wie im mittleren Bereich.

Die Karten in den Abbildungen 11.6, 11.7 sowie 11.8 stellen die Informationswerte aus Diagramm 11.5 mit den oben beschriebenen Parametern für Synopsenkarte, Clustering und Isoglossendarstellung dar. Alle drei Darstellungsweisen lassen die Trennung entlang der Yat-Linie klar erkennen. Der größte Teil des westlichen Teils erscheint homogen, entsprechend dem einheitlichen Verlauf des Diagramms 11.5 im westlichen Teil. Die Rhodopen sind nicht als abgetrenntes Dialektareal erkennbar. Am süd-östlichen Rand, nördlich der Türkei allerdings nicht an das Schwarze Meer anschließend, zeigt sich eine Sprachinsel mit westlich geprägten Dialekten.

11.2.3 Entropie

Im Gegensatz zur Information wird hier die Entropie nicht auf Basis der Wahrscheinlichkeiten der XSampa-Codes im gesamten Datensatz, sondern lediglich auf Grundlage der Werte *innerhalb* des jeweiligen Messpunktes berechnet (messpunktbasierte Methode). Den geringsten Entropiewert weist der

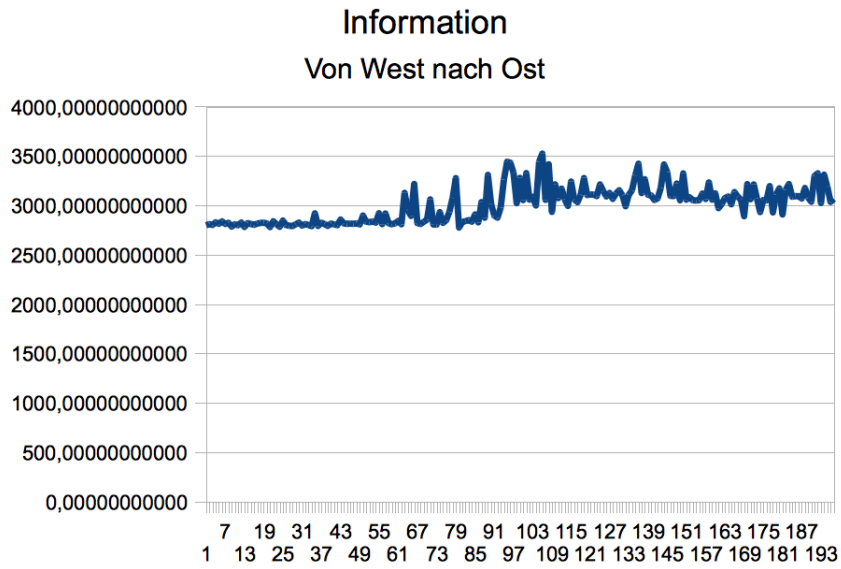


Abbildung 11.5: Informationsgehalt der Messpunkte, dargestellt von West nach Ost

Messpunkt Golema Rakovica mit 4,24 Shannon auf, der Maximalwert liegt wiederum in Ustovo mit 4,69 Shannon. Wird die Position mit eingerechnet, ergeben sich maximal Golica (6,72 Sh) und Voden (6,23 Sh).

Die Diagramme 11.9 für die Entropie bzw. 11.10 für die Entropie inklusive Position weisen beide einen sehr heterogenen Verlauf auf, erkennbar ist ein Sprung in den Entropiewerten ungefähr in der Mitte der Diagramme, was wiederum der Yat-Linie entspricht. Das Diagramm der Entropie unter Einbeziehung der Position (Abbildung 11.10) weist einen etwas ruhigeren Verlauf auf als die Entropie-Werte ohne Einbeziehung der Position.

Die Abbildungen 11.11 bis 11.13 visualisieren die Entropie-Werte der Messpunkte. Es zeigt sich wiederum die Zweiteilung Bulgariens. Im Gegensatz zu den Informationswerten ist die westliche Enklave im Südosten nicht erkennbar.

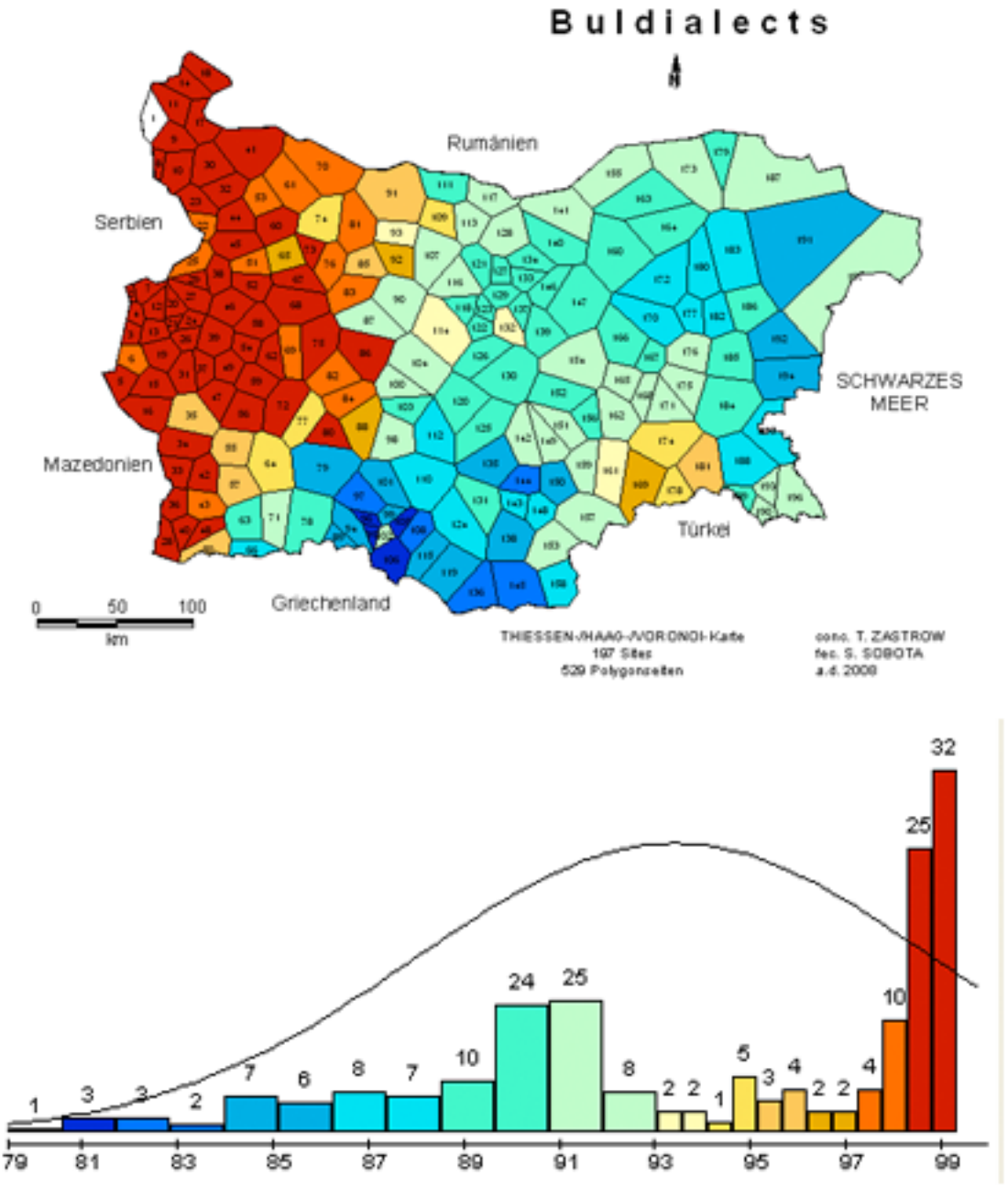


Abbildung 11.6: Synopsenkarte: Information, bezogen auf den gesamten Datensatz

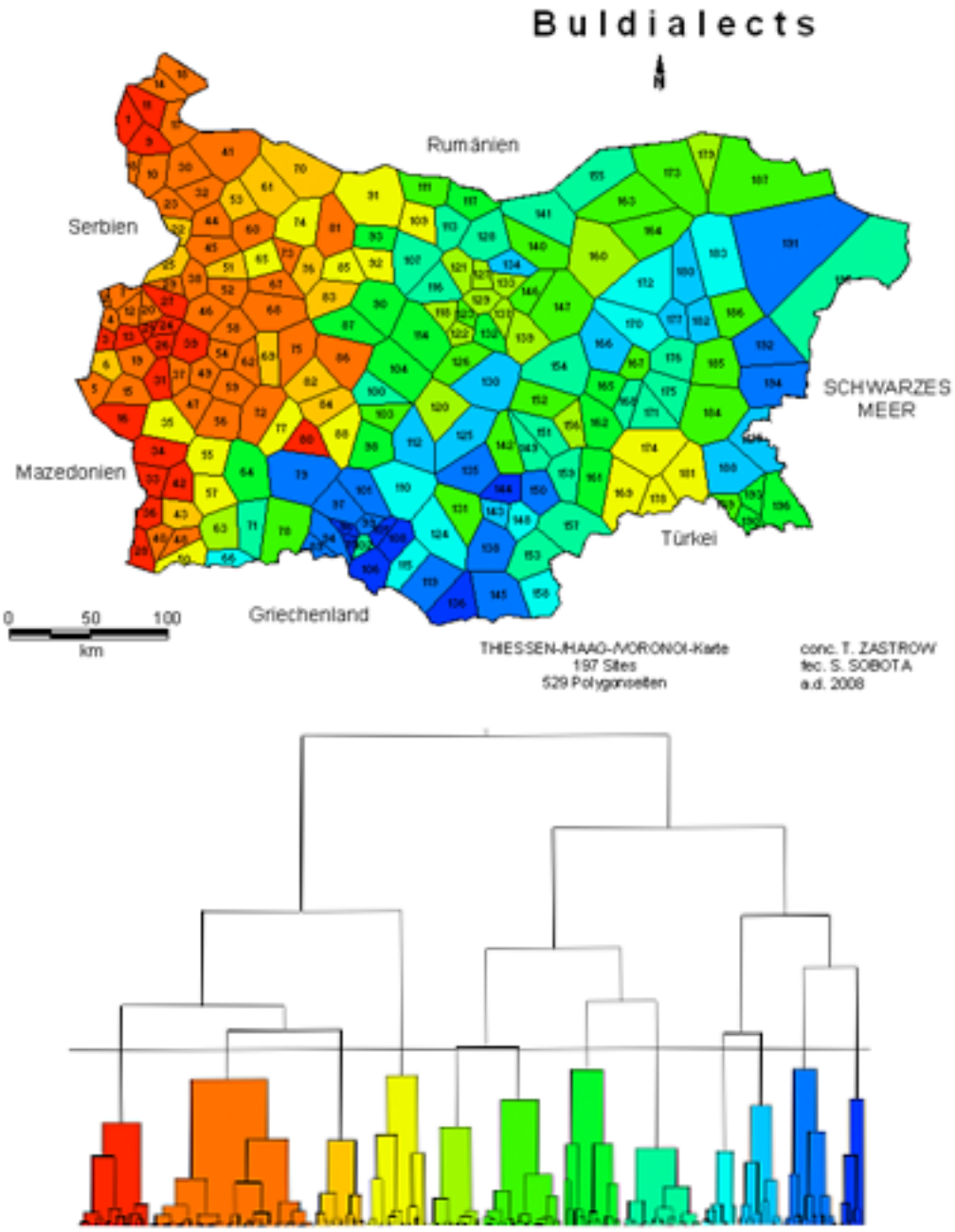


Abbildung 11.7: Clustering: Information, bezogen auf den gesamten Datensatz

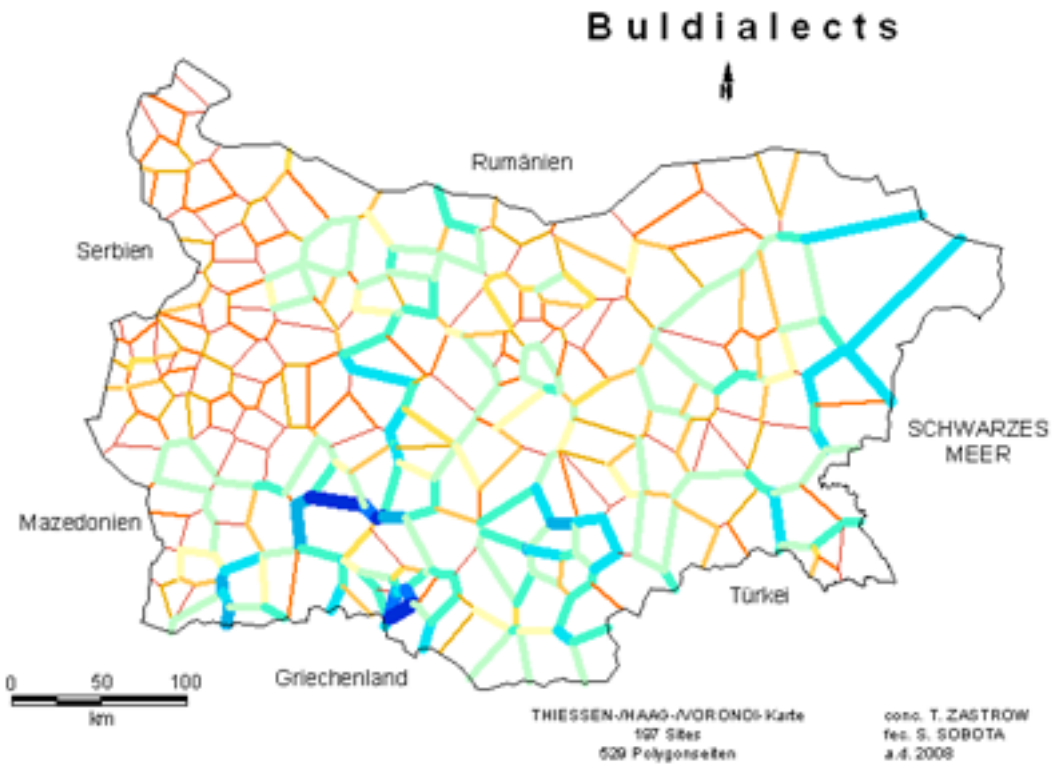


Abbildung 11.8: Isoglossenkarte: Information, bezogen auf den gesamten Datensatz

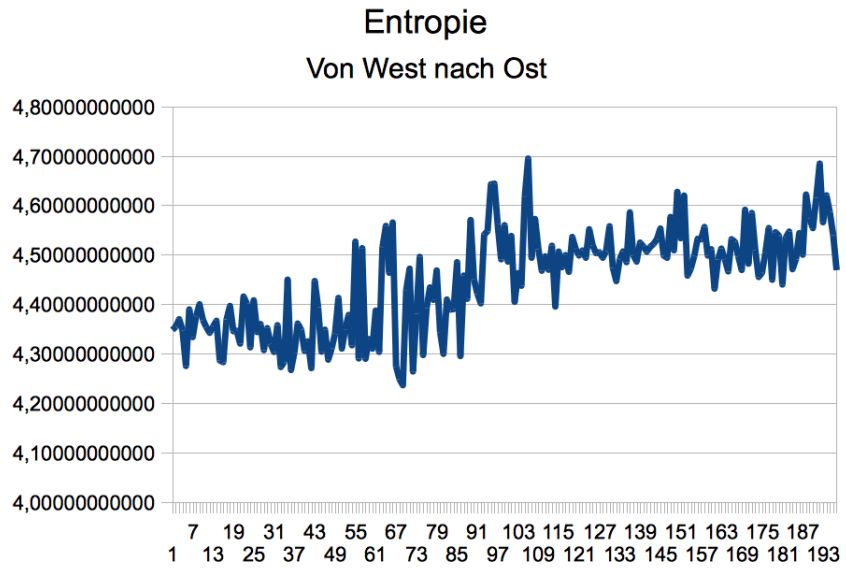


Abbildung 11.9: Entropiewerte der Messpunkte, dargestellt von West nach Ost

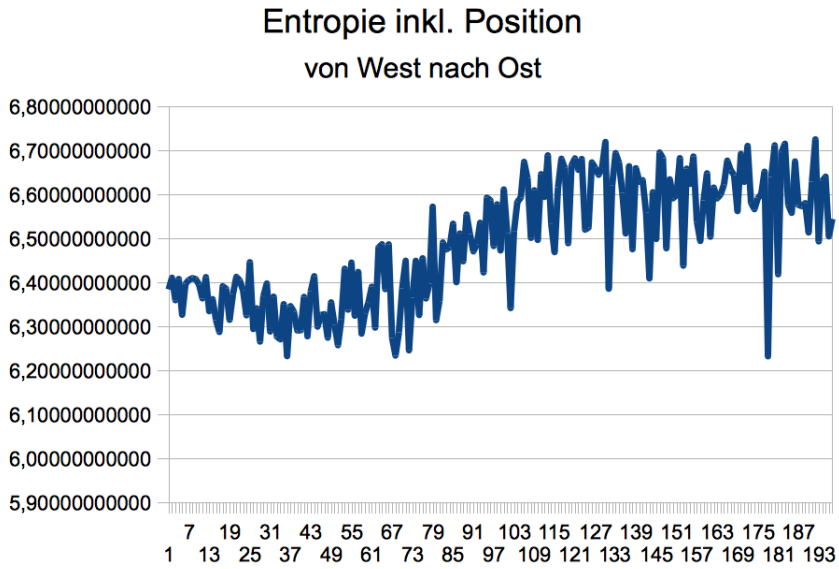


Abbildung 11.10: Die Entropie inklusive der Position, dargestellt von West nach Ost

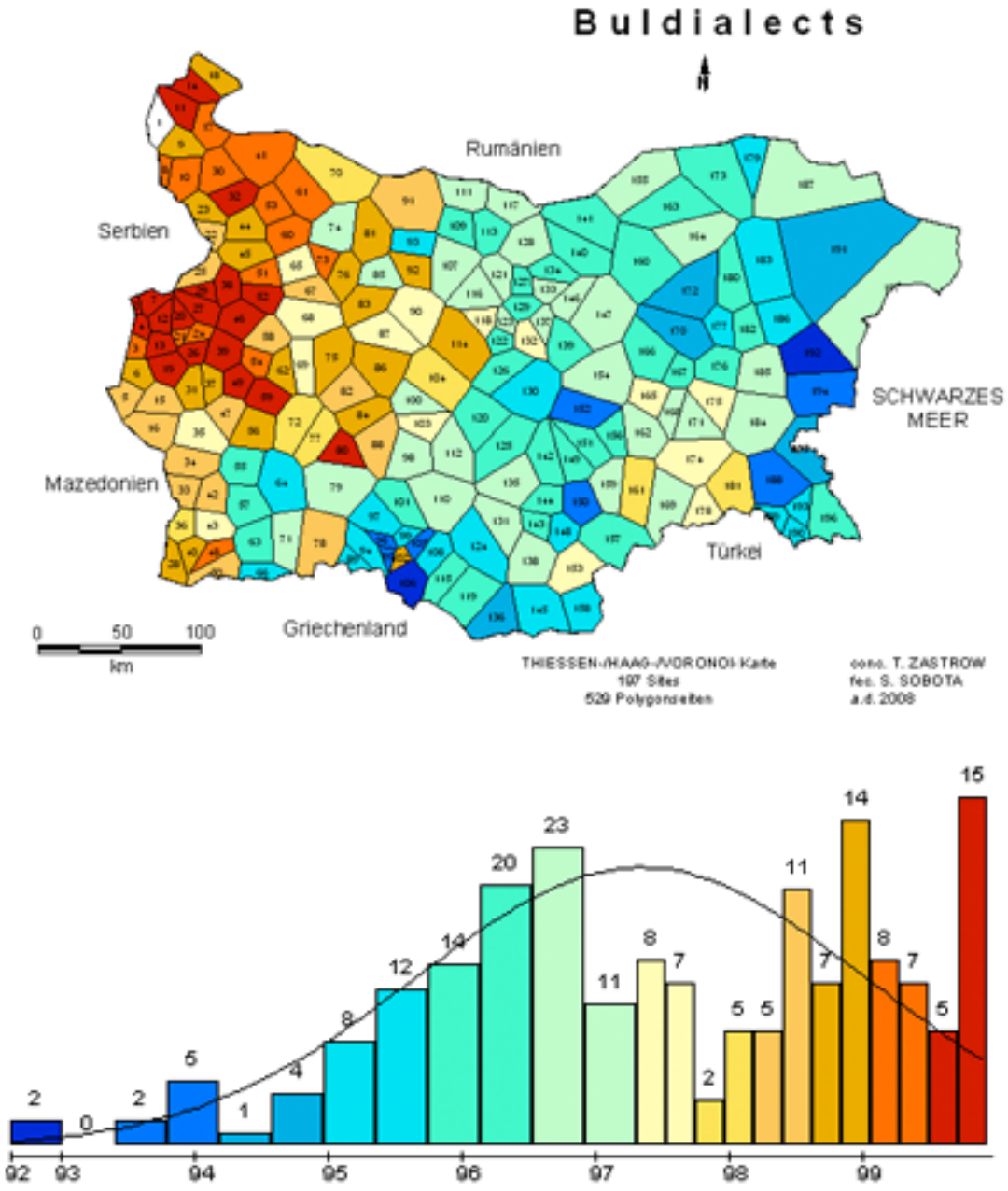


Abbildung 11.11: Synopsenkarte: Entropie

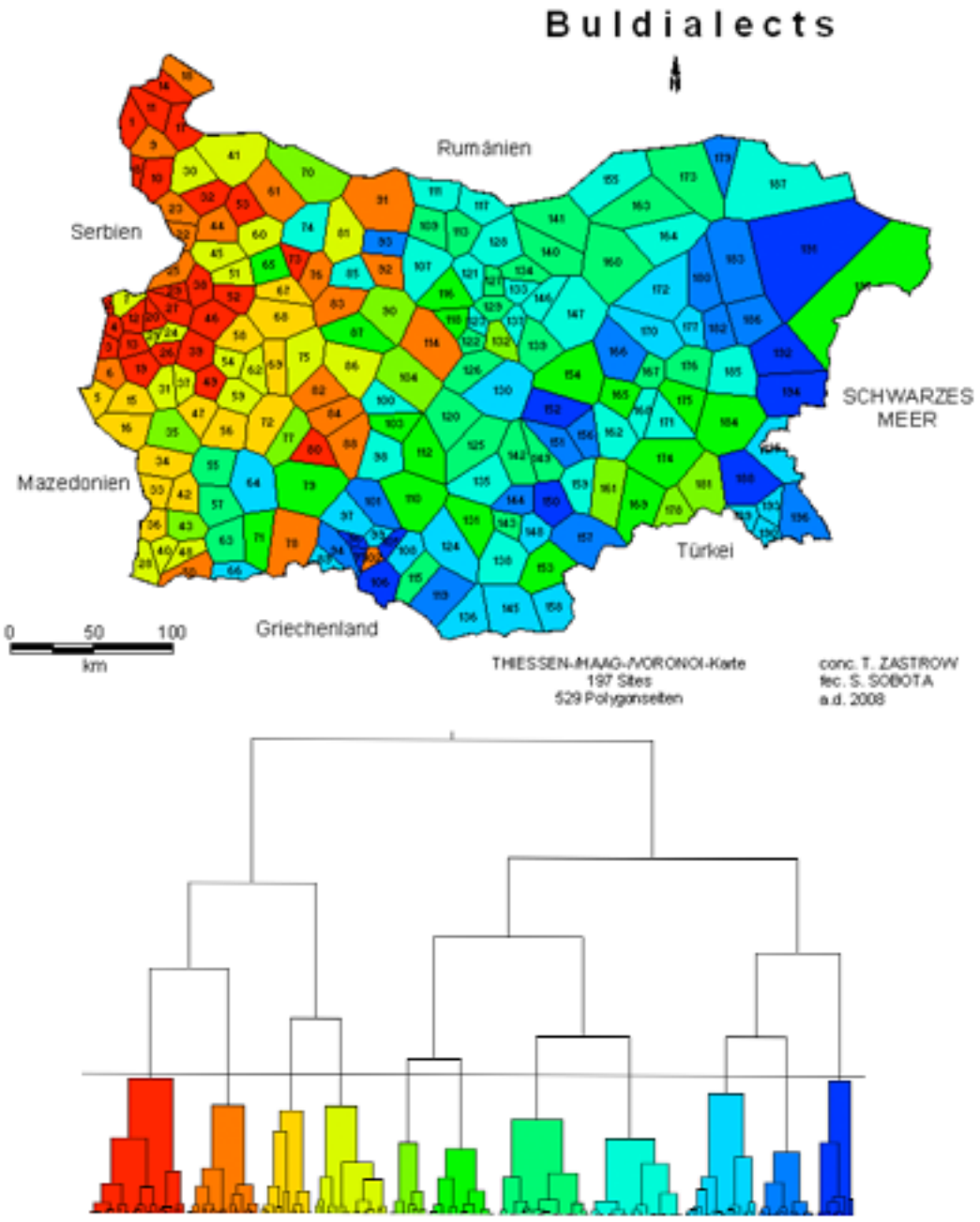


Abbildung 11.12: Clustering: Entropie (aus technischen Gründen sind in dieser Karte 11, und nicht wie auf den anderen Karten, 12 Cluster dargestellt)

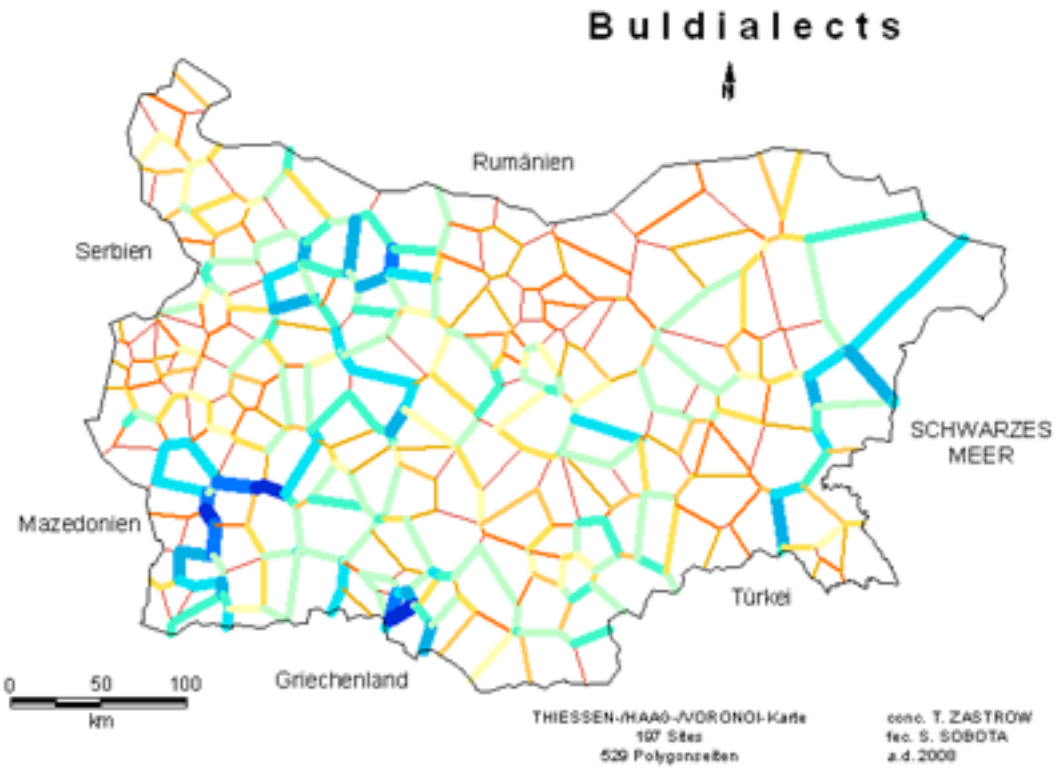


Abbildung 11.13: Isoglossenkarte: Entropie

11.2.4 Noisy Clustering

Wie in Kapitel 8.2.4 beschrieben, ist aus dem Ergebnis eines hierarchischen Clusteringprozeßes nicht mehr ersichtlich, wie die finale Aufteilung der Daten in Cluster zustande gekommen ist. Um dennoch einen Eindruck von der Stabilität der einzelnen Cluster zu gewinnen, kann vor dem Clusteringprozeß per Zufallsgenerator erzeugtes Datentrauschen zu den Daten hinzugefügt werden (Noisy Clustering). Je stabiler die Grenzen zwischen den Clustern sind, desto mehr Datenrauschen kann zu den Daten hinzugefügt werden, bevor sich Änderungen an der endgültigen Clusteraufteilung zeigen.

Abbildung 11.14 zeigt mehrere Durchgänge von auf den Informationsgehalt der Messpunkte angewendetem Noisy Clustering. Die Parameter des Clustering sind dabei nicht verändert worden: Wards Clusteringmethode, Visualisierung von zwei Clustern auf topographischer Karte. Die Anzahl von zwei Clustern wurde gewählt, um den Einfluß des Noisy Clustering auf die Grenze zwischen den beiden Clustern ohne Einwirkung weiterer Clustergrenzen anzeigen zu können. Desweiteren entspricht eine Einteilung in zwei Cluster ungefähr dem Verlauf der Yat-Linie.

Der Parameter γ, der die Obergrenze für das hinzuzufügende Rauschen bestimmt, ist als Prozentwert der Standardabweichung definiert worden. Die erste Karte links oben zeigt die Originaldaten ohne Rauschen (0%). In 5%-Schritten wurde dann der Parameter γ erhöht und Datenrauschen per Zufallsgenerator erzeugt und den Daten hinzugefügt.

Ohne Rauschen zeigt sich die Zweiteilung des bulgarischen Sprachraums entlang der Yat-Isoglosse in einen Ost- und einen Westteil. Ein Bereich im Südosten an der Grenze zur Türkei ist ebenfalls noch dem westlichen Cluster zuzurechnen. Bis zu einem Wert von 15% bleiben die beiden Cluster in ihren Grenzen stabil, was auf eine hohe Stabilität des Clustering hindeutet. Allerdings ist auch bei diesen Werten bereits eine leichte Ausdehung des westlichen Clusters in Richtung Osten erkennbar. Diese erfolgt entlang der Yat-Linie, was darauf hindeutet, dass die Messpunkte, die direkt an die östliche Seite der Yat-Linie angrenzen, doch noch eine größere Ähnlichkeit zu den westlichen Clustern aufweisen als weiter entferntere Messpunkte. Es zeigt auf, dass auch die starken Isoglossen innerhalb einer Sprache durchlässig sind und zu

den benachbarten Dialekten höhere Ähnlichkeiten aufweisen als zu weiter entfernteren. Die direkt an eine Isoglosse angrenzenden Dialekte weisen auf beiden Seiten der Isoglosse immer noch eine gewisse Ähnlichkeit zueinander auf.

Bei Werten ab 20% breitet sich der westliche Cluster in östliche Richtung bis zum Schwarzen Meer aus und eine klare Abgrenzung der beiden Cluster ist nicht mehr erkennbar.

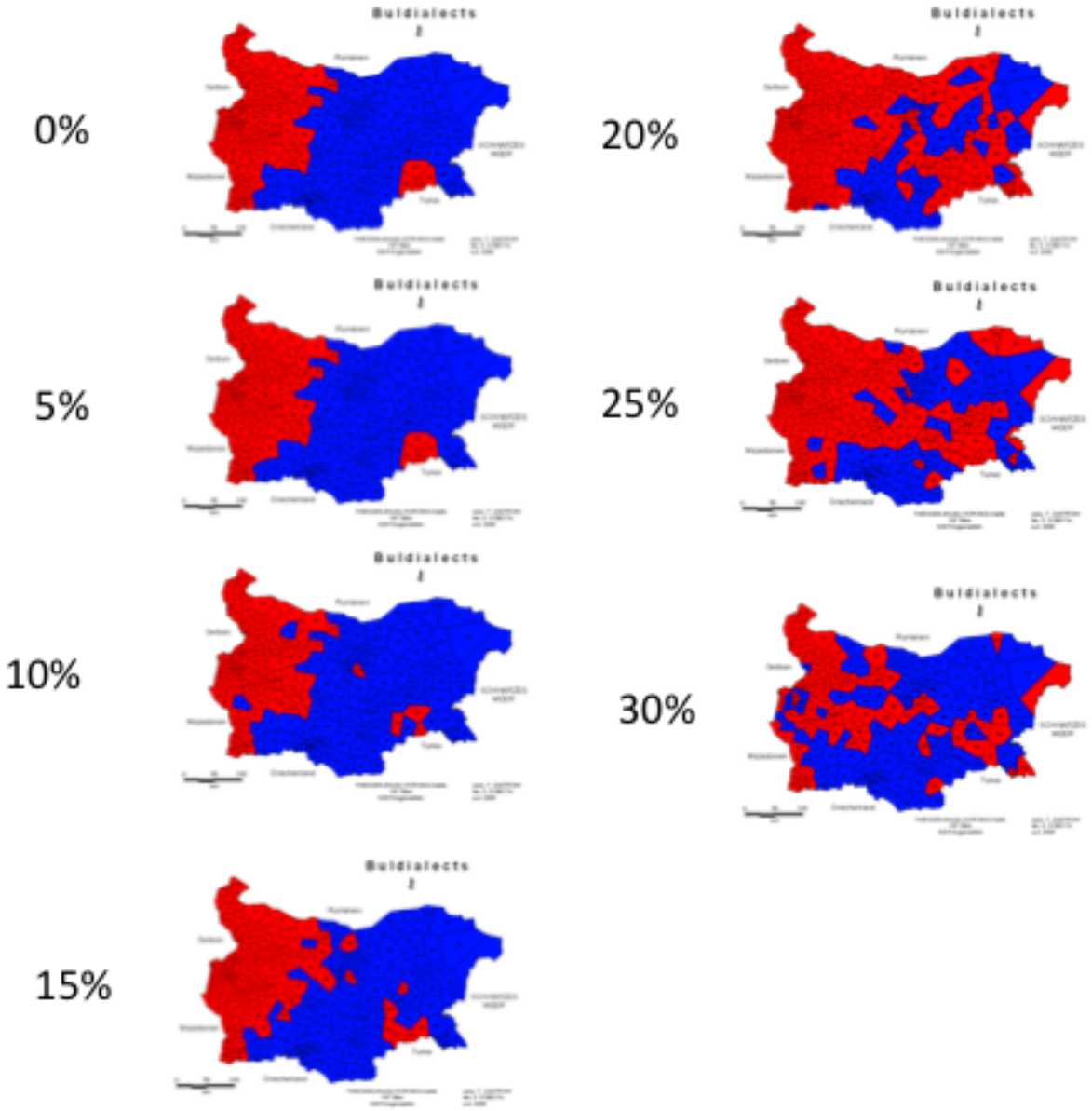


Abbildung 11.14: Noisy Clustering der Information.

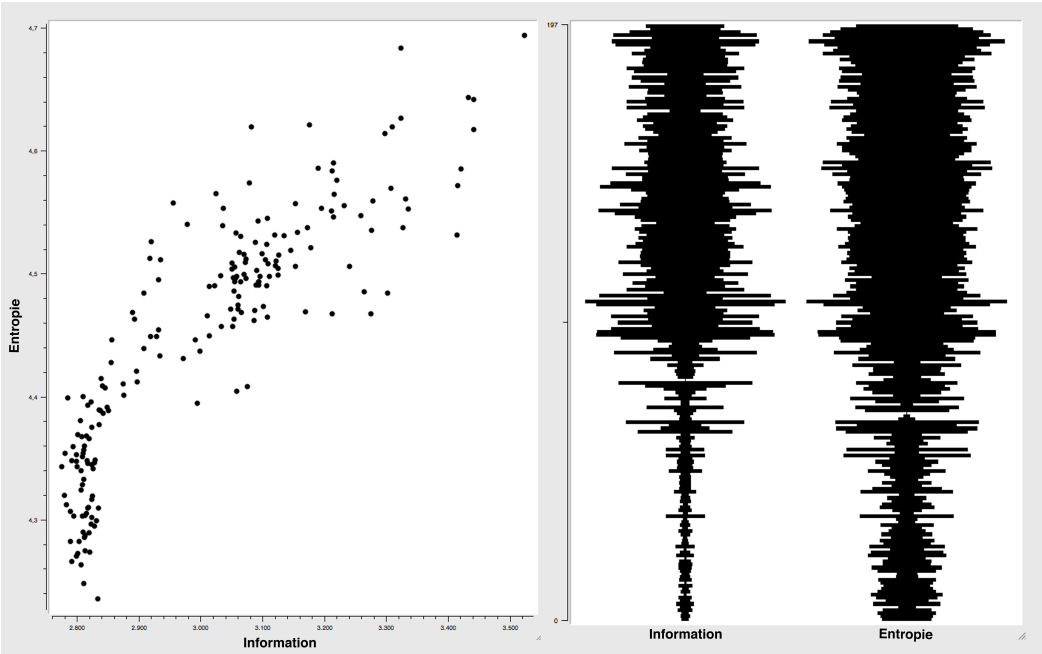


Abbildung 11.15: Vergleich der Informations- und Entropiewerte

11.2.5 Vergleich: Informationswerte und Entropie

Entropie- und Informationswerte werden, wie oben beschrieben, unterschiedlich berechnet und kommen zu unterschiedlichen Ergebnissen. Hier bietet es sich an, die entstandenen Datenreihen bzw. deren Visualisierungen direkt miteinander zu vergleichen.

Abbildung 11.15 visualisiert die Informations- und Entropiewerte (ohne Einbeziehung der Position der einzelnen Elemente) der 197 Messpunkte auf zweierlei Weise. Auf der linken Seite der Abbildung befindet sich ein *Scatterplot*, der auf der X-Achse die Informationswerte und auf der Y-Achse die Entropie der Messpunkte aufträgt. Diese sind als Punkte in das Koordinatensystem eingetragen, in aufsteigender Sortierung von unten nach oben (Y-Achse, Entropie) bzw. von links nach rechts (X-Achse, Information). Es ergibt sich idealisiert eine von links unten nach rechts oben verlaufende Kurve, deren Verlauf abflacht. Es lässt sich eine Korrelation zwischen Entropie und Information erkennen: Je größer der eine, desto größer wird auch der andere Wert sein. Im oberen, höherwertigen Bereich der Kurve streuen die

Messpunkte mehr als im unteren Bereich. Dies deutet darauf hin, dass sich die Informations- und Entropiewerte mit zunehmender Größe auseinanderentwickeln: Je größer Entropie- und Informationswert eines Messpunktes, desto größer ist auch der Unterschied zwischen den beiden Werten. Im unteren und im mittleren Bereich lassen sich jeweils zwei Gruppierungen von Messpunkten ausmachen: In diesen Bereichen sind die Messpunkte dichter vertreten als im restlichen Bereich des Plots.

Rechts neben dem Scatterplot ist ein *Surveyplot* dargestellt. Hier werden die einzelnen Dimensionen für jeden Messpunkt in Form eines horizontalen Balkens dargestellt, die anschließend vertikal zu einer Säule aufeinandergestapelt werden. Der Surveyplot zeigt in der linken Säule die Informationswerte und in der rechten die Entropiewerte der Messpunkte. Die individuellen Werte der Messpunkte (durchnummeriert bis 197 auf der Y-Achse) werden in den jeweiligen Säulen durch die horizontale Länge des entsprechenden Balkens dargestellt. Auch hier zeigt sich in beiden Säulen wieder ein ähnlicher Verlauf, allerdings sind die Werte im unteren Teil der Informations-Säule generell kleiner als die entsprechenden Werte in der Entropie-Säule. In der die Information darstellenden Säule ist die Trennung zwischen dem oberen und dem unteren Teil klarer ausgeprägt.

Es lässt sich festhalten, dass die bulgarischen Messpunkte sowohl durch die Informations- als auch die Entropiewerte in zwei Gruppen getrennt werden. Diese Trennung ist in den Informationswerten klarer zu erkennen als in den Entropiewerten (Surveyplot). Dies deutet auf den stärker von den Daten abstrahierenden Charakter der Entropiewerte hin: Die Unterschiede zwischen den Dialekten sind stärker nivelliert als dies bei den Informationswerten der Fall ist. Die breiter gestreuten Messpunkte im oberen Bereich des Scatterplots stellen die heterogenen Bereiche der Rhodopen dar: Hier finden sich mit die höchsten Informations- und Entropiewerte, die gleichzeitig größere Abstände zueinander aufweisen als dies bei den restlichen Dialektgebieten Bulgariens der Fall ist.

11.3 Vektoranalyse

Im Gegensatz zu den anderen in dieser Arbeit dargestellten Methoden, stellt die Vektoranalyse eine *extrahierende* und keine *aggregierende* Methode dar. Dies bedeutet, dass der Gebrauch einzelner phonetischer Elemente in den verschiedenen Dialekten analysiert wird und der Datensatz nicht als Ganzes betrachtet wird. Gleichzeitig ist die Vektoranalyse *wortbasiert*, was bedeutet, dass die Daten in allen Messpunkten nach einer festen Reihenfolge der Wörter angeordnet werden. Die Wortgrenzen stellen die maximal mögliche Ausdehnung der Vektorkette auf der X-Achse dar. Die einzelnen Vektoren entsprechen somit den *relativen Positionen* des untersuchten Elements innerhalb der Wörter.

Für jeden vorkommenden XSampa-Code in den 197 bulgarischen Messpunkten wurden wie in Kapitel 5 beschrieben Vektorketten erstellt und deren Längen berechnet. Es ergeben sich die Kennwerte:

- Anzahl Vektorketten mit Länge $\neq 0$: 6823
- Länge der kürzesten Kette: 3
- Länge der längsten Kette: 232.91
- Durchschnittliche Länge einer Kette: 110.89

Die Abbildung 11.16 zeigt 6 Vektorketten, jeweils für den XSampa-Code ”e”. Die drei Vektorketten A, B und C repräsentieren die Messpunkte Rakovica, Rani Lug und Dragojchinci. Dabei handelt es sich um die drei westlichsten Messpunkte des Datensatzes. Die weiteren Vektorketten D, E und F entsprechen den Messpunkten Chernomorec, Balgari und Asparuhuvo: Hierbei handelt es such um die drei östlichsten Messpunkte des Datensatzes.

Die Vektorketten stellen individuelle *Fingerabdrücke* des jeweiligen Messpunktes dar. Je ähnlicher sich die Fingerabdrücke zweier Messpunkte sind, desto ähnlicher sind sich die entsprechenden Dialekte in Bezug auf das extrahierte Element (hier XSampa-Code). In der Abbildung 11.16 ergeben sich durch rein visuelle Inspektion große Ähnlichkeiten zwischen den Vektorketten A, B, C und E einerseits und D und F andererseits. Dies entspricht der bereits

häufig gesehenen Einteilung der bulgarischen Dialektareale: der geographisch im Osten verortete Messpunkt E liegt zwar im äußersten Süd-Osten Bulgariens, gehört hier aber zu einer Gruppe *westlich* geprägter Dialekte.

Abbildung 11.17 zeigt die Frequenzen der einzelnen XSampa-Codes (rot) im Vergleich zur Länge der insgesamt 6823 Vektorketten (blau), diese ebenfalls sortiert nach den Frequenzen. Während die Frequenzen eine diskrete Datenreihe darstellen, sind die Längen der Vektorketten stetig: Ihr Werttebereich ist innerhalb jeder Frequenzstufe monoton steigend. Die Längen der Vektorketten weisen somit mehr Varianz auf als eine Betrachtung der reinen Frequenzen.

Für jeden XSampa-Code lassen sich nun die Differenzen in den Längen der Vektorketten der einzelnen Messpunkte errechnen. Diese lassen sich wiederum in einer symmetrischen Datenmatrix darstellen. Die so entstandenen Matrizen sind in den Abbildungen 11.18 und 11.19 visualisiert. Beide Abbildungen enthalten jeweils Voronoi-Karten für die Matrizen der XSampa-Codes A, e, i, o, u und ", letzteres stellt den Wort-Akzent dar. Abbildung 11.18 zeigt Synopsenkarten, Abbildung 11.19 hierarchisches Clustering nach WARD. Auf den Karten der Vokale lassen sich folgende Dialektareale erkennen:

- **A:** Außer der starken Yat-Linie zwischen West- und Ostbulgarien zeigen sich so gut wie keine weiteren Strukturen. Lediglich die moesischen Dialekte im Nordosten sind auf der Synopsen-Karte abgegrenzt. Eine klare Abtrennung dieses Dialektgebietes wird durch die in dieser Arbeit angewendeten aggregierenden Methoden nur selten erreicht.
- **e:** Zusätzlich zu der klar erkennbaren Yat-Linie und dem heterogenen Rhodopen-Areal ist eine vom Südrand Bulgariens nach Nordwesten verlaufende Gruppe von Dialekten sichtbar (dunkelrot auf der Synopsenkarte, hellblau im Clustering).
- **i, o und u:** Hier werden die Rhodopen einmal dem Westen (i) und zweimal dem Osten (o und u) zugerechnet. Die Karten für o weisen

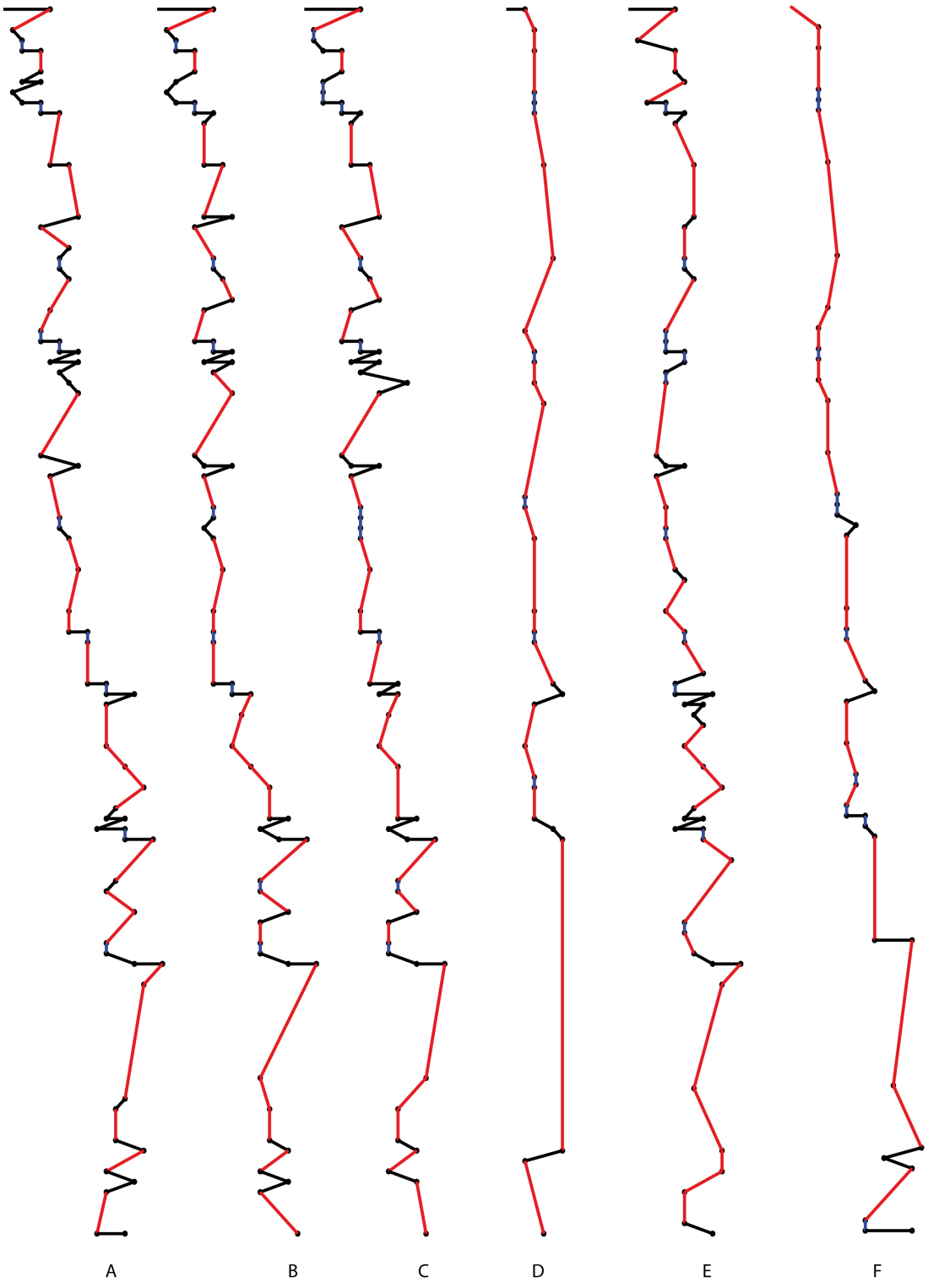


Abbildung 11.16: Sechs Vektorketten des XSampa-Codes ”e” der Messpunkte: A Rakovica, B Rani Lug, C Dragojchinci, D Chernomorec, E Balgari und F Asparuhuvo

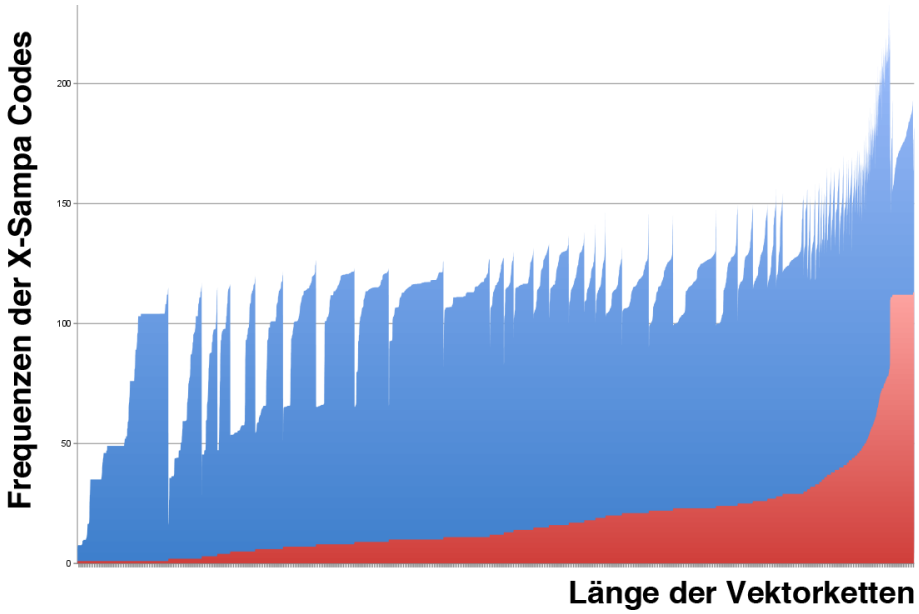


Abbildung 11.17: Frequenzen der einzelnen XSampa-Codes (diskret, rot) und die Längen der Vektorketten (stetig, blau)

im Osten eine große Heterogenität auf: Hier sind lediglich die mittleren Rhodopen und einige der Messpunkte am Schwarzen Meer in eine eigene Gruppe abgetrennt. Die Karten für u zeigen deutlich die Übergangsdialekte zum Serbischen an der Westgrenze auf.

Allgemein zeigen die Vektorketten in den oben genannten Abbildungen keinen Widerspruch zu den Ergebnissen der aggregierenden Methoden auf. In einigen regional begrenzten Bereichen lassen sich allerdings auch feinere Abgrenzungen innerhalb der großen Dialektareale erkennen. Das allgemein heterogene Gebiet der Rhodopen ist nur auf den Karten der e-Vektorkette klar erkennbar als eigenes Dialektareal abgegrenzt, auf den anderen Karten ist es entweder dem Westen oder dem Osten zugeordnet.

Der Wortakzent weicht von den anderen Karten ab: Lediglich in der Mitte Bulgariens erscheint ein zusammenhängendes Gebiet. Im Gegensatz zu allen anderen Karten bildet hier der äußerste Osten eine Gruppe mit dem Westen. Der südliche Rand der Rhodopen ist ebenfalls abgetrennt und bildet ein eigenes Dialektareal. Im Allgemeinen weisen die Karten des Wortakzents keine

klar voneinander abgegrenzten Dialektareale auf.

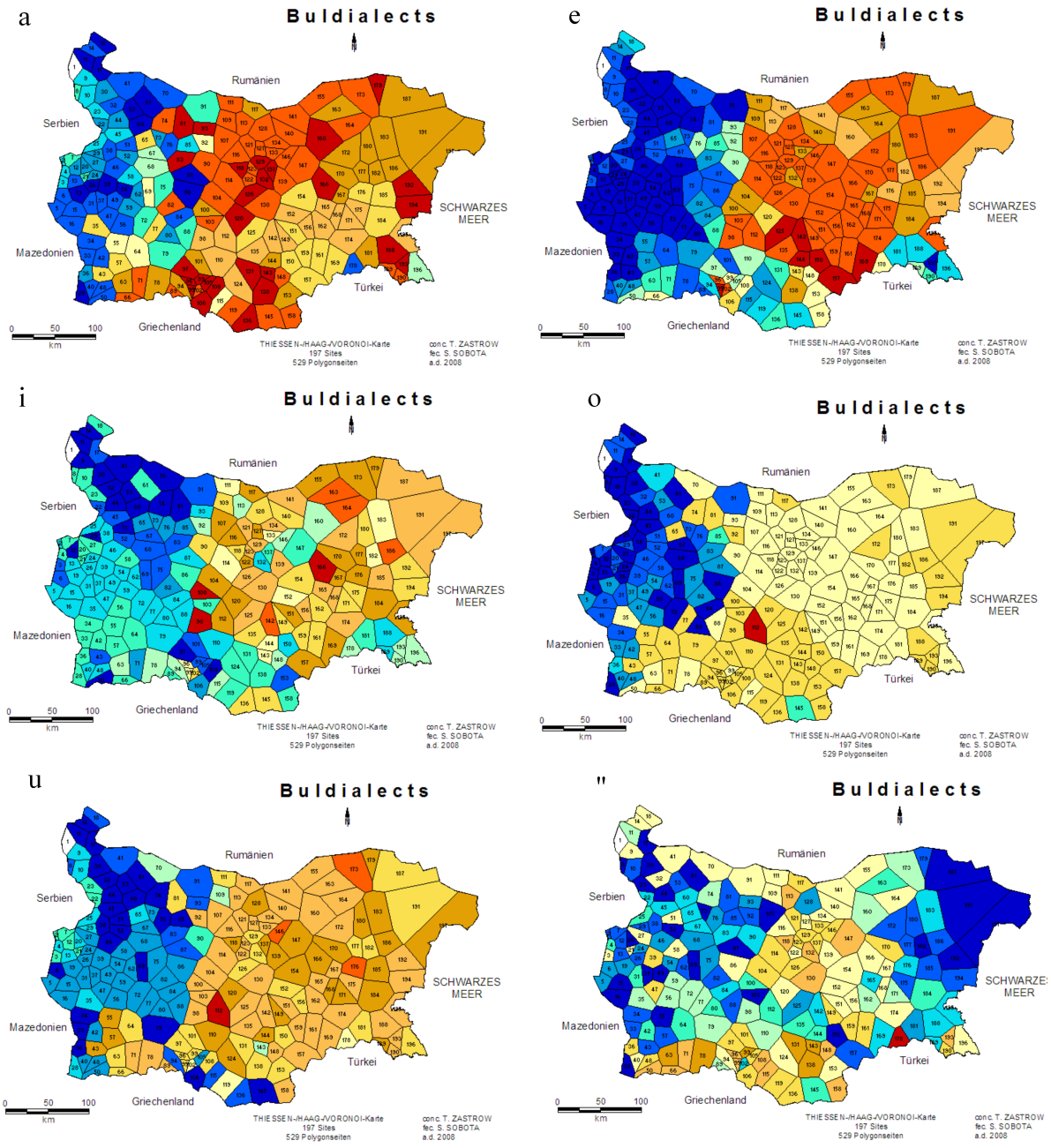


Abbildung 11.18: Vektoranalyse: Synopsenkarte

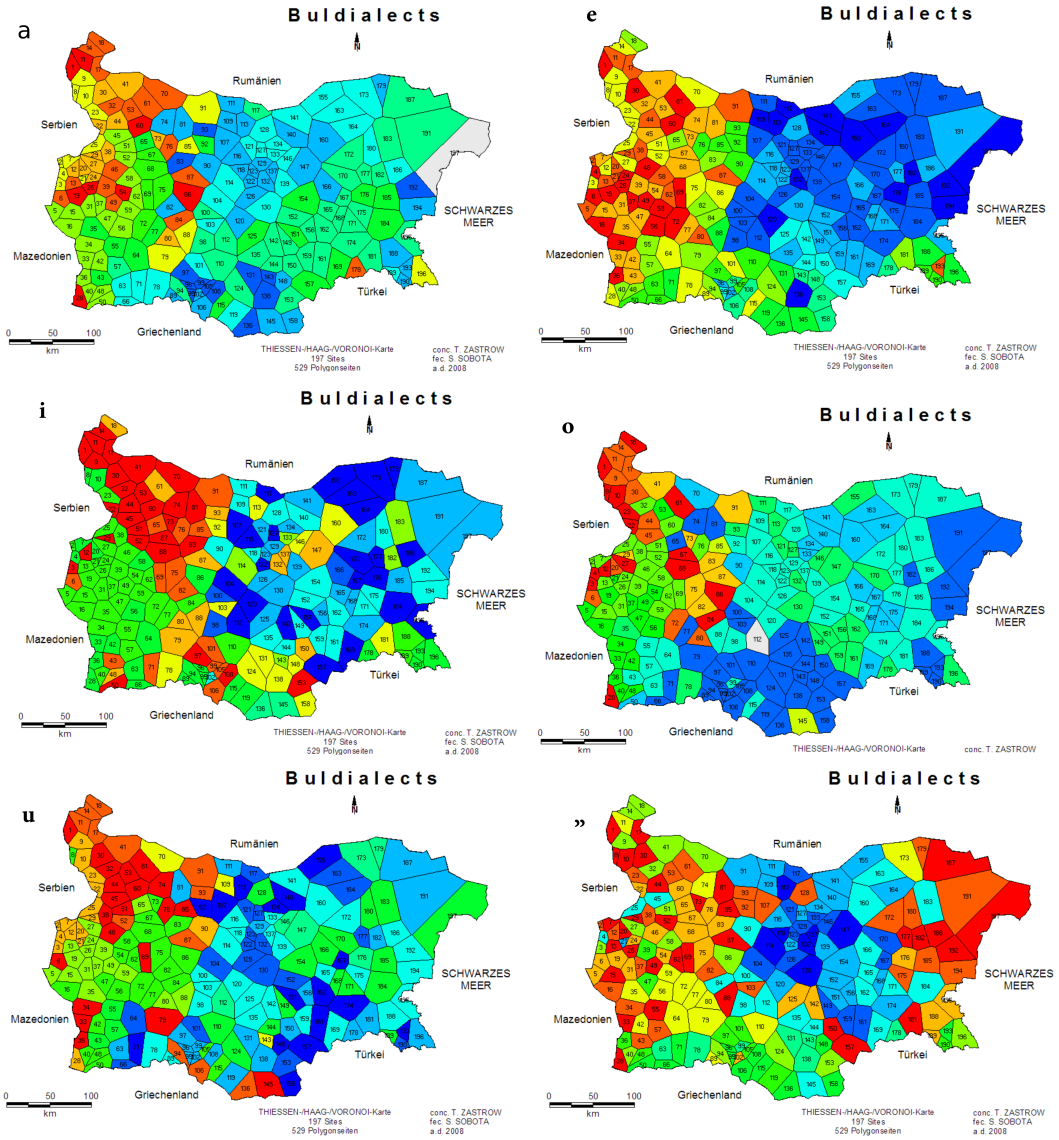


Abbildung 11.19: Vektoranalyse: Hierarchisches Clustering nach WARD, 12 Cluster

11.4 Analyse der Wortakzentverteilung

Ein oder mehrere Teile einer Äußerung können vom Sprecher bzw. der Sprecherin durch den *Akzent*[4] hervorgehoben werden. Durch unterschiedliches Setzen des Akzents kann der Sprechende bewusst die Aufmerksamkeit des Zuhörenden auf einen bestimmten Teil des Gesagten lenken. Der Akzent ist suprasegmental, er gehört nicht zu den einzelnen Elementen einer Sprache, tritt aber immer mit diesen zusammen auf. Gebildet wird der Akzent durch Variation von Tonhöhe, Lautstärke und der Länge des hervorzuhebenden Elements (Altmann u. Ziegenhain, 2010, S. 49 ff.). Erzeugung und Bedeutung des Akzents ist abhängig von der jeweiligen Sprache.

Akzente werden unterschieden in *Wortakzente* und *Satzakzente*:

- Beim Wortakzent wird eine Silbe eines mehrsilbigen Wortes hervorgehoben.
- Der Satzakzent betont ein Wort innerhalb eines Satzes.

Das IPA kennt zwei Varianten des Akzents. Der primäre Akzent wird durch ein hochgestelltes Anführungszeichen (') markiert, der sekundäre Akzent durch ein tiefgestelltes (ˌ). In XSampa entsprechen diese dem doppelten Anführungszeichen (") bzw. dem Prozentzeichen (%). Sie werden jeweils vor die hervorzuhebende Silbe geschrieben. In Wörterbüchern oder vergleichbaren linguistischen Sammlungen wird der Wortakzent häufig durch Unterstreichung der entsprechenden Silbe angezeigt.

In Zhobov u. a. (2004) untersuchen die Autoren das Phänomen der *Additional Accentuation* im Bulgarischen: Hierbei wird unter bestimmten Bedingungen, beispielsweise einer Negation, ein zweiter Satzakzent auf ein Klitikon gelegt[5]. Dies ist ebenfalls abhängig von den verwendeten Konjunktionen, was in einer Hierarchie unter den Konjunktionen resultiert. Die Autoren konstatieren, dass sich die *Additional Accentuation* in einigen Dialekten verändert

[4] Englisch *Stress* - im Rahmen der bulgarischen Dialektologie wird ebenfalls der englische Begriff *Accent* verwendet, siehe bspws. Alexander (2002).

[5] Zum selben Themenkomplex, siehe Alexander (2002).

	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5
Rakovica	82	4	19	5	2
Chepelare	91	3	13	5	0
Asparuhovo	78	3	24	6	1

Tabelle 11.4: Einige Messpunkte mit Angabe, wie häufig der Wortakzent auf den jeweiligen XSampa-Codes liegt

und man hier ein Beispiel für eine gerade stattfindende Wandlung der aktuellen bulgarischen Sprache habe. Es ergibt sich zwischen den Dialekten ein Kontinuum von *aktiver Produktion* der Additional Accentuation, über *passive Akzeptanz* bis hin zu *aktiver Ablehnung*. Im Verlauf der weiteren Entwicklung ist anzunehmen, dass sich die *Additional Accentuation* über die gesamte bulgarische Sprache verbreiten wird.

In der genannten Studie wurden den Probanden komplette Sätze vorgelegt, die diese vorlesen mussten. Abhängig von morphosyntaktischem Kontext und dem gesprochenen Dialekt änderte sich in den gesprochenen Sätzen die Verwendung des Satzakzents. Da der Buldialects Datensatz nur einzelne Wörter und keine ganzen Sätze enthält, kann an ihm der Satzakzent nicht untersucht werden. Eine Analyse des Akzents beschränkt sich hier dementsprechend auf den Wortakzent, der durch das XSampa-Zeichen (“) in den Daten annotiert ist.

Im Bulgarischen ist der Wortakzent generell dynamisch, das heißt, er kann an verschiedene Positionen innerhalb eines mehrsilbigen Wortes gesetzt werden.

11.4.1 Untersuchungsrichtung *Single Site, All Words*

In der SSAW-Untersuchungsrichtung wird die Position des Wortakzents in allen Wörtern eines Messpunktes analysiert. Ein Wortakzent an erster Position im Wort bedeutet somit einen Akzent auf dem zweiten XSampa-Element des Wortes.

Die Tabelle 11.4 zeigt drei Messpunkte mit den Angaben, wie häufig der Wortakzent an der jeweiligen Position innerhalb des Wortes auftritt. Dabei

	Pos. 1	Pos. 2	Pos. 3	Pos. 4	Pos. 5
Maximum	101	12	29	12	4
Minimum	69	0	8	0	0
Arithmetisches Mittel	80,42	3,54	20,32	5,95	1,58

Tabelle 11.5: Aus der Verteilung der Wortakzente errechnete Kennwerte

handelt es sich um die geographisch westlichste (Rakovica), die östlichste (Asparuhovo) und einen Messpunkt, der sich bezüglich des Längengrads in der Mitte Bulgariens befindet (Chepelare). In Tabelle 11.5 finden sich einige statistische Kennzahlen, errechnet aus der Verteilung der Wortakzente in allen 197 Messpunkten. Diese Daten sind in Abbildung 11.20 visualisiert: In allen Messpunkten liegt der Wortakzent mit großem Abstand auf dem ersten XSampa-Code, eine weitere, aber wesentlich kleinere Häufung findet sich an dritter Stelle. Für alle 197 Messpunkte ergibt sich ein ähnlicher Kurvenverlauf.

Ein etwas anderes Bild ergibt sich, wenn nur die Anzahl der Wortakzente an erster Stelle betrachtet wird (Abbildung 11.21). Hier sind die Messpunkte wieder von Westen nach Osten angeordnet. Es ergibt sich eine unruhige Kurve, die aber in ihrem vorderen Teil gleichförmiger verläuft als im hinteren. Dementsprechend werden im östlichen Teil Bulgariens weniger Wortakzente auf dem ersten XSampa-Code angewandt als im westlichen Teil. Die absolute Anzahl der Wortakzente bleibt dabei weitgehend konstant (zwischen 109 und 113 pro Messpunkt).

11.4.2 Untersuchungsrichtung *Single Word, All Sites*

In der Untersuchungsrichtung *Single Word, All Sites* (SWAS) werden die dialektalen Varianten eines Wortes in allen Messpunkten analysiert. Die einzelnen Varianten einer Messreihe sind somit direkt miteinander vergleichbar.

Mit großem Abstand liegt der Wortakzent auch hier wieder auf dem ersten XSampa-Code des Wortes (71,85%), ebenfalls etwas herausgehoben ist der dritte XSampa-Code (18,16%, vollständige Auflistung in Tabelle 11.6).

Von den insgesamt 119 Wörtern des Buldialect-Datensatzes weisen le-

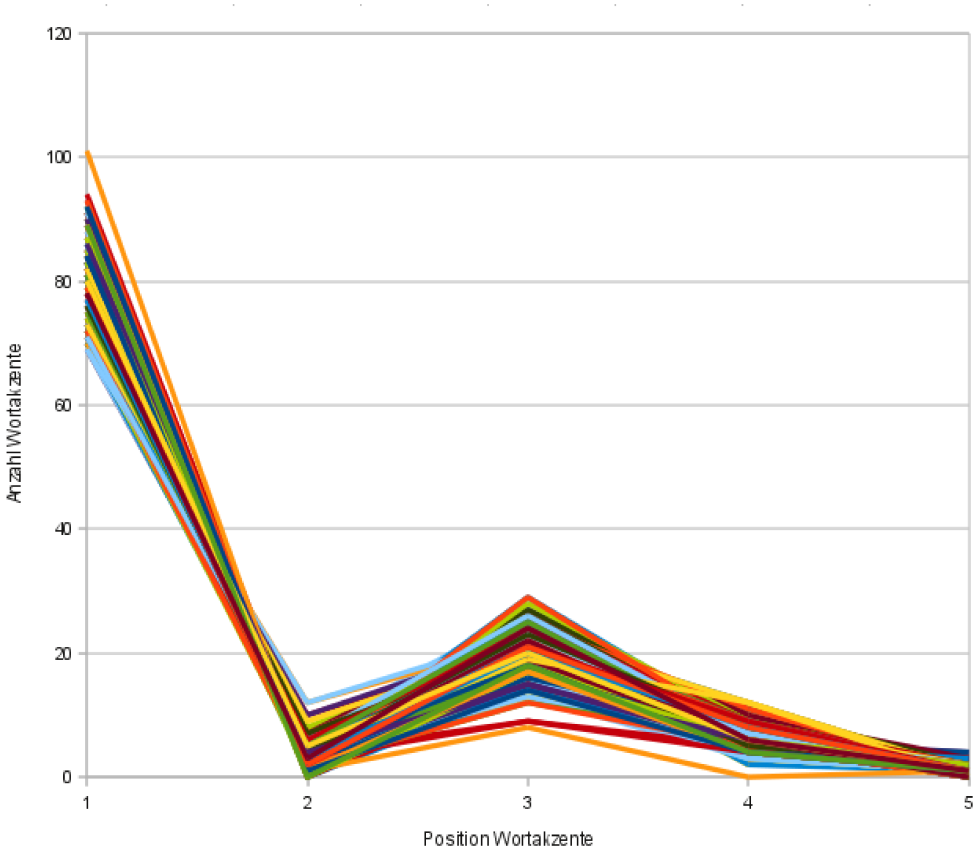


Abbildung 11.20: Visualisierung der Position des Wortakzents in allen 197 Messpunkten

Position	1	2	3	4	5	6
Absolute Anzahl	15843	697	4004	1172	311	22
Prozent	71,85	3,16	18,16	5,32	1,41	0,1

Tabelle 11.6: Verteilung des Wortakzents auf die Positionen innerhalb der Wörter

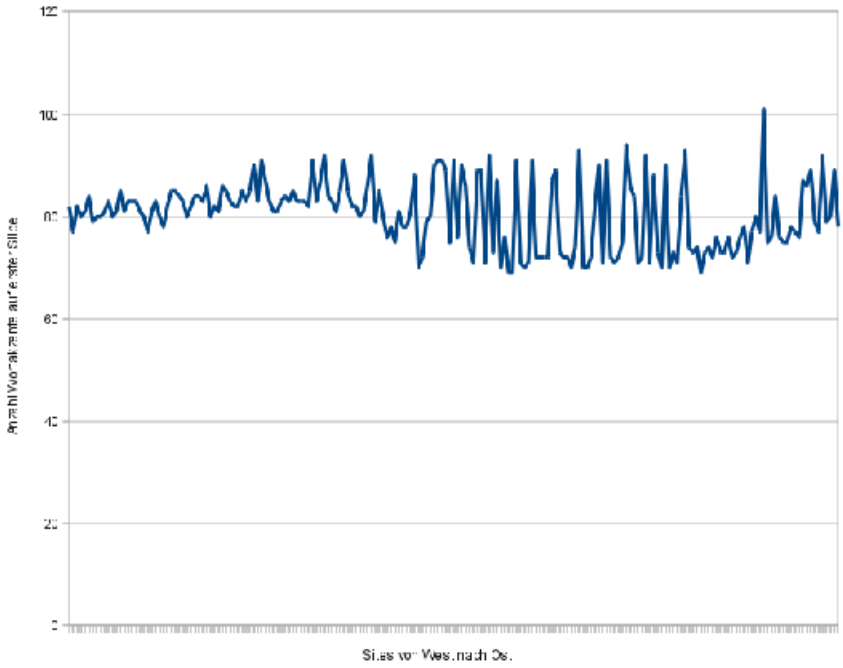


Abbildung 11.21: Anzahl des Wortakzents auf dem ersten XSampa-Code, aufgetragen von Westen nach Osten

Wortakzent an Position	1	2	3	4	5	6
Wort 49	48	0	6	0	120	22
Wort 61	149	1	41	5	0	0
Wort 73	148	10	39	1	0	0
Wort 74	18	8	170	1	0	0
Wort 117	90	53	16	38	0	0
Wort 118	17	8	134	1	37	0

Tabelle 11.7: Sechs Wörter mit dem Wortakzent an jeweils 5 verschiedenen Positionen

diglich 67 Wörter überhaupt Wortakzente an unterschiedlichen Positionen innerhalb der dialektalen Varianten auf. Bei den restlichen 52 Wörtern tritt der Wortakzent immer an derselben Stelle im Wort auf, es ergibt sich so keinerlei Varianz. Im Vergleich aller Wörtern zueinander ist die Dominanz des Wortakzents an erster Stelle bei den 67 Wörtern mit Varianz nicht mehr ganz so stark (52,97% gegenüber den oben genannten 71,85%).

11.4.3 Zusammenfassung

Werden einzelne Wörter ohne weitergehenden Kontext betrachtet, so ist die Position des enthaltenen Wortakzents nicht wirklich ein funktionsfähiges Instrument zur geographischen Einteilung der bulgarischen Dialekte. Der überwiegende Teil der Wörter enthält den Wortakzent auf der ersten Position. Die weitere in der Literatur beschriebene Verteilung der (Satz-) Akzente lässt sich ohne Kontext nicht an einzelnen Wörtern nachweisen. Allerdings ist festzuhalten, dass die östlichen Dialekte eher zur Variation neigen und den Wortakzent häufiger auch in den hinteren Teilen des Wortes anwenden. Im westlichen Teil dagegen findet sich der Wortakzent fast ausschließlich auf dem ersten XSampa-Code.

11.5 Bigramm-Analysen

Auf die atomaren Elemente des bulgarischen Datensatzes bezogen, handelt es sich bei einem *Bigramm* um zwei aufeinanderfolgende XSampa-Codes. Diese werden im Rahmen einer Bigramm-Analyse als eigenständiges, zusammengehörendes Element betrachtet; eine Aufsplittung in die ursprünglichen XSampa-Codes findet nicht mehr statt. Prinzipiell wären auch Analysen an Trigrammen oder generell N-Grammen denkbar. Da die hier verwendeten Wörter des bulgarischen Datensatzes in ihrer Gesamtheit relativ kurz sind, liefert lediglich die Aufteilung in Bigramme eine hinreichend hohe Anzahl unterschiedlicher Elemente.

Aus den gesamten phonetischen Dialektdaten des bulgarischen Datensatzes lassen sich so insgesamt 104.077 einzelne Bigramme erstellen. Diese setzen sich aus 906 verschiedenen Bigrammen zusammen. Das am häufigsten auftretende Bigramm ist die Kombination "v mit 2924 Vorkommen. Unter allen Bigrammen, die häufiger als 1.000 Mal auftreten, sind nur zwei Bigramme ohne Beteiligung des Wortakzents: $\_je$ mit 2167 Vorkommen und $\_jA$ mit 1057 Vorkommen. Teilt man die Bigramme in Häufigkeitsklassen ein, so ergibt sich bei insgesamt 279 Häufigkeitsklassen - von einmaligen bis zum 2924-maligen Auftreten - eine exponentiell abnehmende Kurve: Viele Bigramme kommen nur selten vor, wohingegen einige wenige sehr häufig auftreten (Abbildung 11.22).

11.5.1 Messpunktspezifische Matrizen

Der Buldialects-Datensatz setzt sich aus einem Alphabet von 56 unterschiedlichen XSampa-Codes zusammen. Dies ergibt 56^2 oder 3136 verschiedene, mögliche Bigramme. Hieraus lässt sich für jeden Messpunkt eine zweidimensionale quadratische Matrix mit der Kantenlänge 56 zusammensetzen. In die Zellen werden die Vorkommen der einzelnen Bigramme eingetragen (vertikale Achse: erstes Unigram, horizontale Achse: zweites Unigramm). Für jede dieser Matrizen lässt sich die *Bigrammdichte* berechnen: Sie gibt an, wieviele Zellen der Matrix einen Wert > 0 enthalten in Relation zur Anzahl aller Zellen. Dies lässt sich wie folgt interpretieren: Je größer die Bigrammdichte, desto mehr *unterschiedliche* Bigramme werden in dem entsprechenden

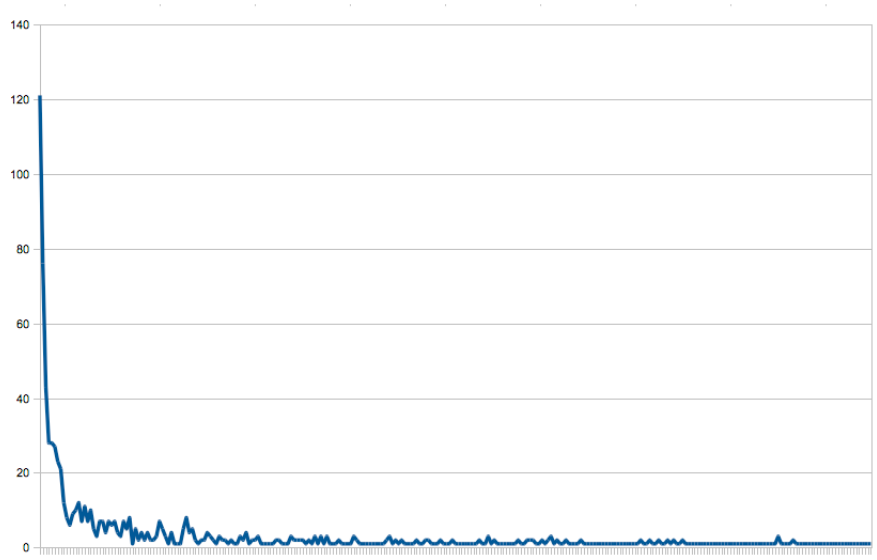


Abbildung 11.22: Verteilung der Bigramm-Häufigkeitsklassen

Messpunkt verwendet. Wobei die Bigrammdichte noch keine Aussage über die absolute Anzahl der verwendeten Bigramme enthält: die Datensätze der einzelnen Messpunkte unterscheiden sich nicht wesentlich in ihrer Größe, was auch zu ähnlichen Anzahlen an Bigrammen pro Messpunkt führt.

Im bulgarischen Datensatz variiert die Bigrammdichte bei insgesamt 3136 verschiedenen möglichen Bigrammen zwischen 5.9% (was 188 verschiedenen realisierten Bigrammen pro Messpunkt entspricht) und 8.4% (entsprechend 266 verschiedenen Bigrammen pro Messpunkt). Der Durchschnitt liegt bei 7,1%. Der Messpunkt Ustovo, der den höchsten informationstheoretischen Informationsgehalt aufweist, hat nach Indzhe Vojvoda die zweithöchste Bigramm-Dichte.

Zur Visualisierung können die Zellen der Matrizen abhängig vom enthaltenen Wert in Grautönen eingefärbt werden. Abbildung 11.23 zeigt die Bigrammmatrizen von 6 Messpunkten. Die oberen drei befinden sich im Westen, die unteren drei im Osten Bulgariens:

- Im Westen liegen die drei Messpunkte Rakovica, Rani Lug und Dragojchinci. Hier weisen die beiden Messpunkte Rani Lug und Dragojchinci

visuell starke Ähnlichkeiten zueinander auf, die Verteilung der Graustufen ist in beiden Matrizen vergleichbar. Aber auch der Messpunkt Rakovica weist ähnliche Bigramm-Verteilungen innerhalb der Matrix auf, hier sind die Graustufen lediglich etwas abgeschwächt.

- Auch die im Osten befindlichen Messpunkte Chernomorec und Asparuhovo weisen dieselbe Verteilungsstruktur der Bigramme auf, die Verteilung der Graustufen ähnelt sich sehr. Die Graustufenverteilung des dritten Messpunktes, Balgari, ist hingegen stärker ausgeprägt als die der beiden anderen östlichen Messpunkte und ähnelt wiederum eher den westlichen Messpunkten. Dies entspricht der Beobachtung, dass es sich bei den Dialekten in diesem Gebiet um eine westlich geprägte Enklave handelt.

11.5.2 Aggregation der messpunktspezifischen Matrizen zu einer Ähnlichkeitsmatrix

Zwei Bigrammmatrizen lassen sich miteinander vergleichen, indem für jede Zelle der beiden Matrizen die Differenz gebildet wird und diese anschließend aufsummiert werden. Das Ergebnis ist eine positive Ganzzahl und stellt ein Distanzmaß für die beiden untersuchten Matrizen dar: Je kleiner der Wert, desto ähnlicher sind sich die beiden Matrizen. Wird das Distanzmaß in der hier beschriebenen Art und Weise für jede Matrix zu jeder anderen Matrix errechnet, so können diese anschließend in eine neue symmetrische Matrix eingetragen und mit den Mitteln der Dialektometrie analysiert werden.

Die Abbildungen 11.24, 11.25 und 11.26 visualisieren diese aggregierte Matrix. Klar erkennbar ist wieder die Yat-Linie, sie trennt die roten bzw. orangefarbenen Bereiche im Westen von den blauen und türkisfarbenen im Osten und den grünen im Süden (Rhodopen). Innerhalb dieser Haupt-Isoglossen haben sich weitere, klar abgegrenzte Gebiete herauskristallisiert: Die Übergangsdialekte an der Westgrenze zu Serbien sind klar erkennbar, die ansonsten heterogenen Rhodopen erscheinen ebenfalls einheitlich in Grüntönen mit einer Ausbeulung nach Norden. Weiterhin lassen sich im Nordosten (dunkelblau) und in der westlichen Mitte zwei zusammenhängende Bereiche

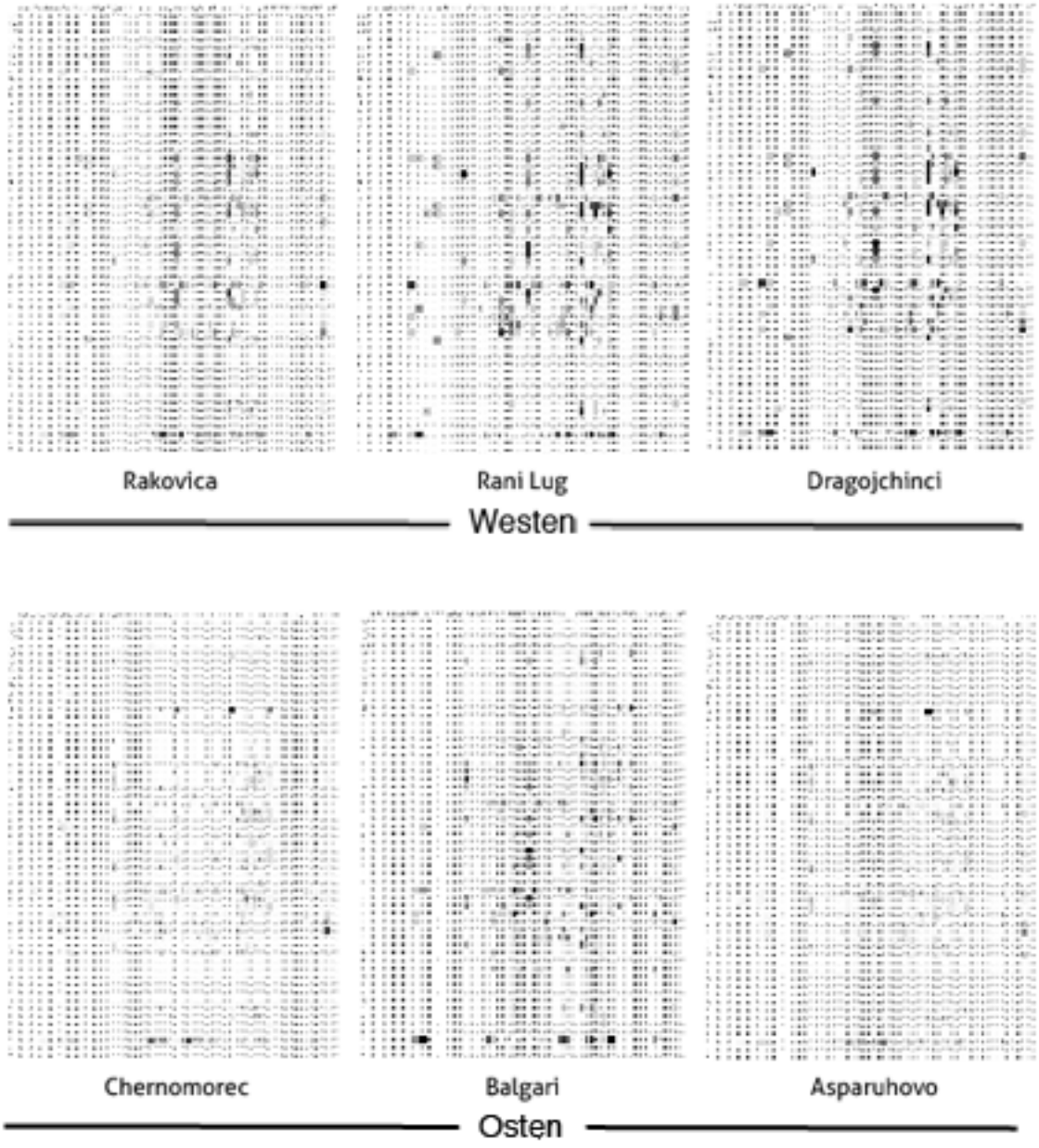


Abbildung 11.23: Visualisierung der Bigrammhäufigkeiten mittels Graustufen: Im oberen Teil drei Messpunkte aus dem Westen, im unteren Messpunkte aus dem Osten

(gelb) entlang der Yat-Linie ausmachen. An der Grenze zur Türkei existiert wieder eine Sprachinsel, die teilweise dem östlichen Großraum und teilweise den Rhodopen zugehörig ist.

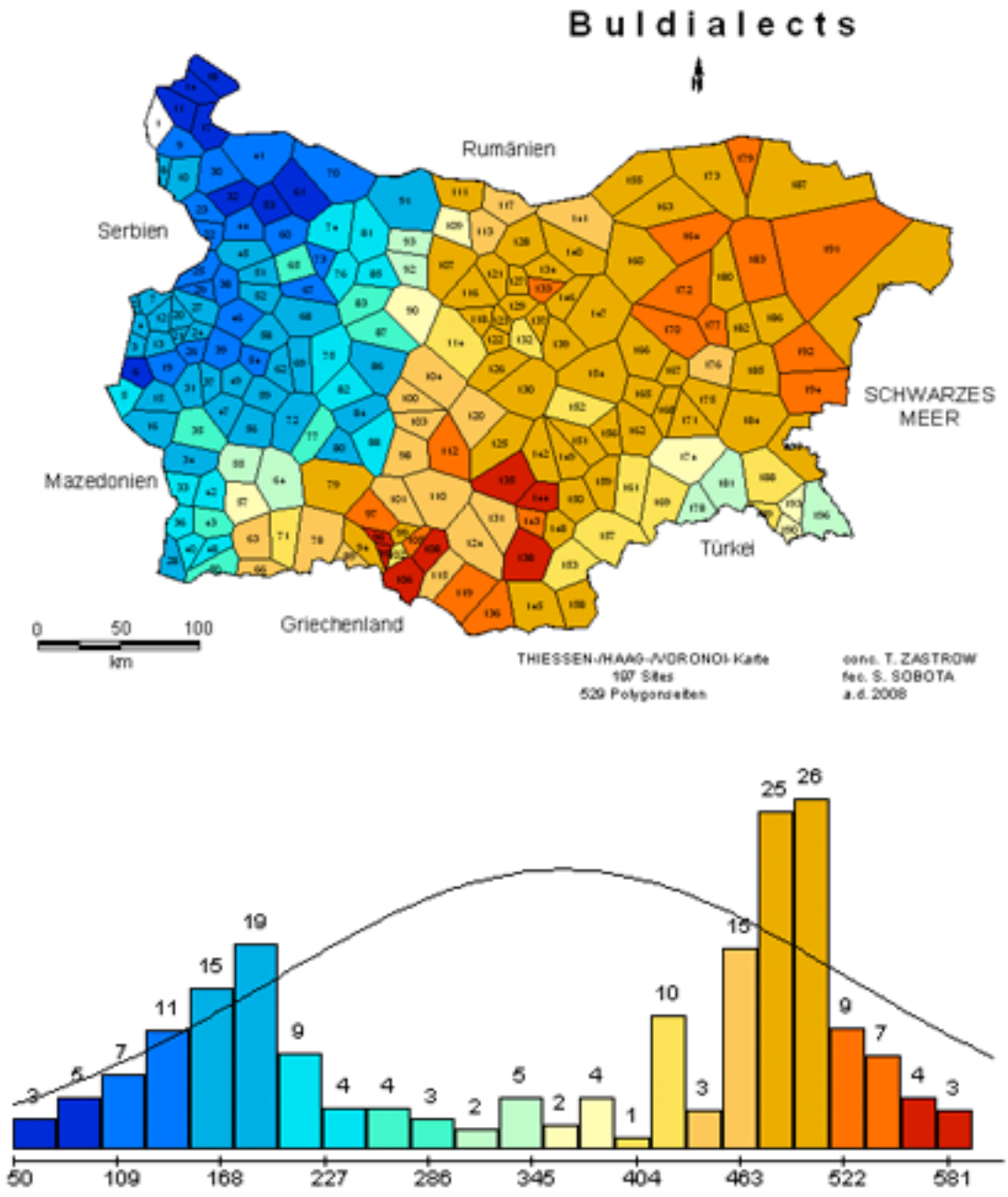


Abbildung 11.24: Synopsenkarte: Aggregierte Bigramme

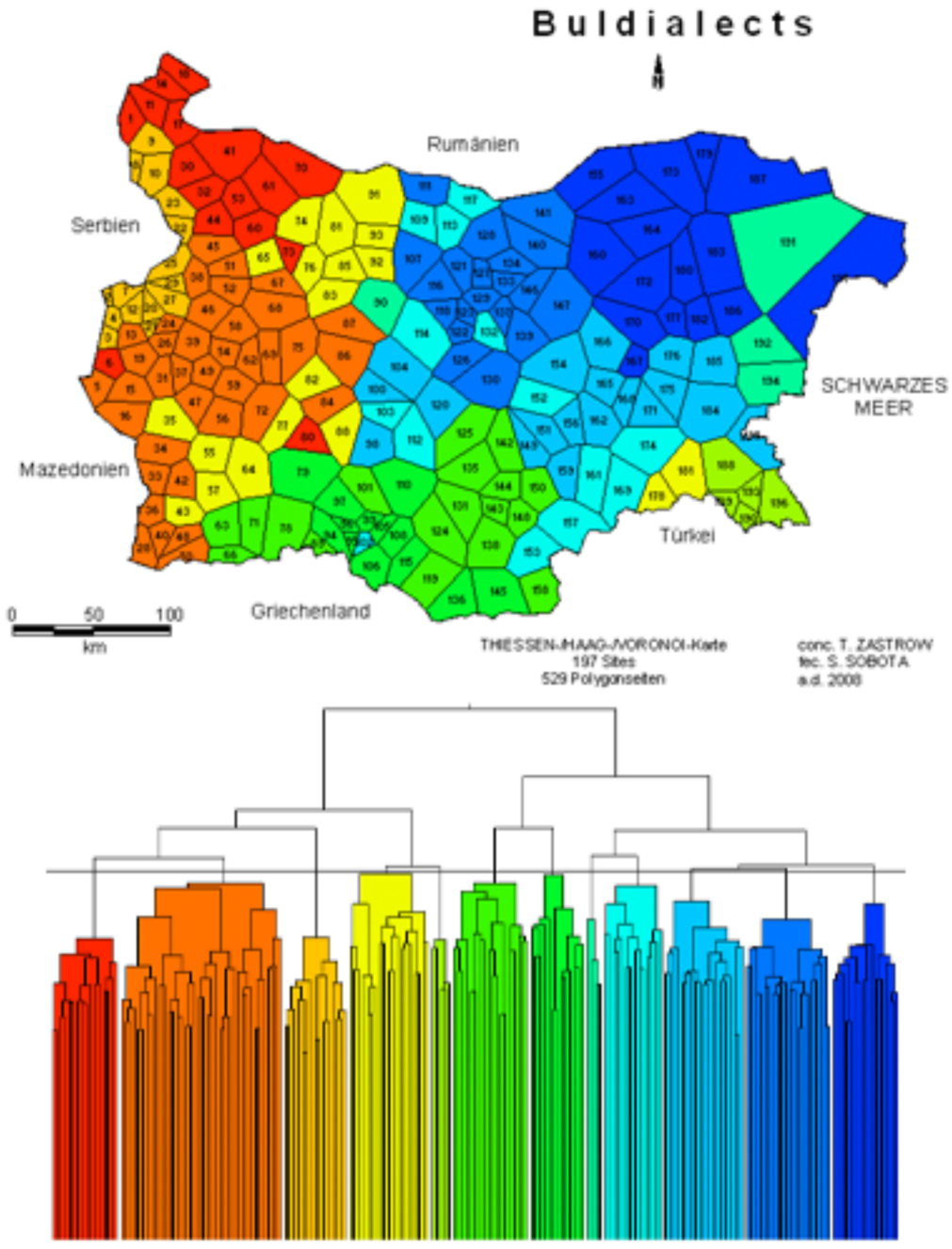


Abbildung 11.25: Clustering: Aggregierte Bigramme

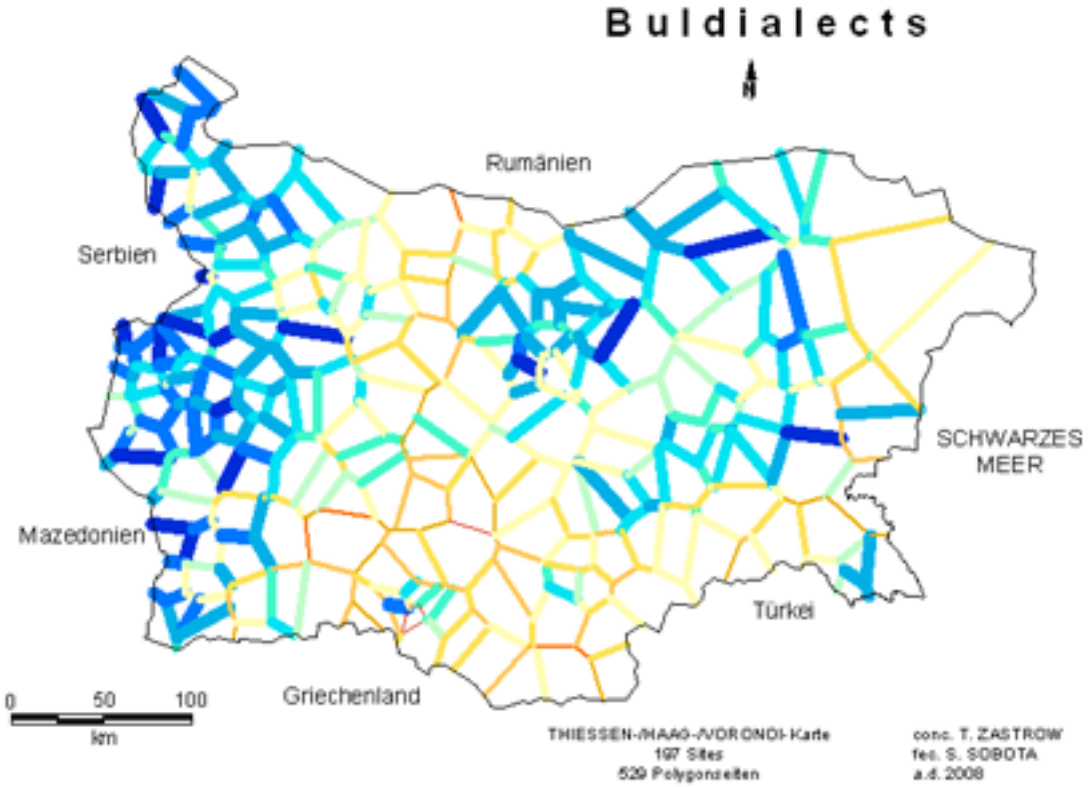


Abbildung 11.26: Isoglossenkarte: Aggregierte Bigramme

11.6 Weitere Visualisierungsmöglichkeiten

11.6.1 Reliefdarstellung

Die Visualisierung von dialektometrischen Kennzahlen auf topographischen Karten erfolgt meistens durch die Verwendung verschiedener Farben. Die zweidimensionalen Polygone, die auf einer Voronoi-Karte den einzelnen Messpunkten entsprechen, werden anhand einer festgelegten Anzahl von Farben oder eines stetigen Farbverlaufs eingefärbt. Dabei visualisieren verschiedene Farben die unterschiedliche Höhe der dialektometrischen Kennzahlen oder die verschiedenen Abstände einer Ähnlichkeitsmatrix.

Eine weitere Möglichkeit, dialektometrische Kennzahlen zu visualisieren, findet sich in der Ausnutzung der dritten Dimension. Abhängig vom Wert der zugeordneten Kennzahl, werden die Polygone der Voronoi-Karte in die dritte Dimension extrudiert und wandeln sich so zu *Prismen* (siehe Abbildung 11.27). Es ergibt sich über das untersuchte Gebiet ein Relief: Höhere Kennzahlen werden durch höhere, niedrigere Kennzahlen durch niedrigere Prismen dargestellt. Sich ähnelnde Messpunkte werden so durch ähnlich hohe Prismen visualisiert. Zusammenhängende Messpunkte ähnlicher Kennzahlen bilden ein *Plateau* ähnlicher Höhe. Die dreidimensionale Reliefdarstellung hat im Gegensatz zur Einfärbung der zweidimensionalen Polygone den Vorteil, dass die Werte der dialektometrischen Kennzahlen direkt visuell durch die Höhe der einzelnen Prismen repräsentiert werden. Bei der Verwendung von Farben ist es nicht ersichtlich, wie die einzelnen Farben interpretiert werden müssen: Welche Farben stehen für hohe, welche für niedrige dialektometrische Kennzahlen? Bei der Kennzahlen-Visualisierung durch Reliefs stellt sich diese Problematik nicht.

Zur besseren Übersichtlichkeit können von den einzelnen Kennzahlen der Messpunkte $K_1 ... K_n$ das Minimum der gesamten Messreihe $min(R)$ subtrahiert werden, sodass die Höhe des Prismas eines Messpunktes $P(S_n)$ lediglich die Unterschiede zwischen den einzelnen Kennzahlen, nicht aber deren absolute Werte darstellt. Damit das dem Minimum der Messreihe entsprechende Prisma sichtbar extrudiert wird, muss anschließend zu allen Messpunkten ein fix vorgegebener Wert q hinzuaddiert werden. Es ergibt sich die Höhe $P(S_n)$

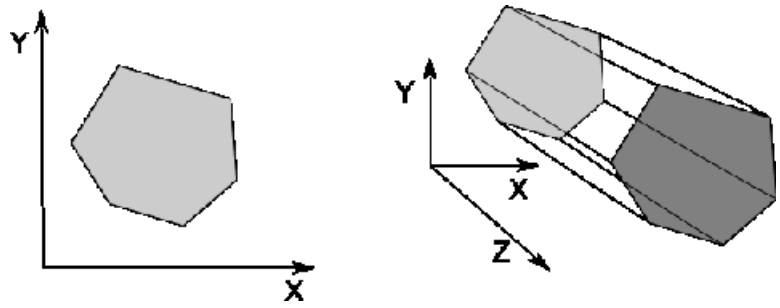


Abbildung 11.27: Links: Zweidimensionales Polygon entlang der Achsen X und Y, rechts: Dasselbe Polygon, zum Prisma extrudiert entlang der Z-Achse

eines Messpunktes:

$$P(S_n) = K_n - min(R) + q \tag{11.3}$$

Raytracing

Raytracing ist eine Technik zur Erstellung fotorealistischer, dreidimensionaler Computergraphiken. Die grundlegenden Algorithmen wurden bereits in den 1960er Jahren entwickelt (Appel, 1968). Beim Raytracing werden Lichtstrahlen auf ihrem Weg durch einen Raum verfolgt und, falls sie auf ein Hindernis treffen, entsprechend eingefärbt bzw. abgelenkt.

Eine Raytracing-Szene besteht aus den darzustellenden Objekten - hier: den Prismen -, einer oder mehrerer Lichtquellen sowie einer Kamera, aus deren Perspektive die Szene betrachtet wird. Die dargestellten Objekte können mit einer individuellen Farbe oder Textur versehen werden. Hier ergibt sich die Möglichkeit, durch die Verwendung entsprechender Farben eine weitere Dimension von Kennzahlen darzustellen oder die betrachteten Messpunkte in Kategorien zusammenzufassen. Die drei Achsen einer Szene (X, Y und Z-Achse) sind nicht mit einer Skalierung versehen: Objekte können einen beliebigen numerischen Raum einnehmen und stehen lediglich in Relation zueinander. So lassen sich die zweidimensionalen Polygone einer Voronoi-Karte eins zu eins auf zwei der drei Achsen abbilden: an der dritten Achse entlang erfolgt die Extrusion des Polygons zum Prisma.

Da in einer Raytracing-Szene sowohl die Kamera als auch deren Blickwin-

kel frei wählbar sind, können die Prismen aus allen möglichen Winkeln und Perspektiven betrachtet werden. Eine Sequenz solcher Szenen mit sich linear verändernder Kameraposition und/oder Blickwinkel kann dementsprechend zu einer Animation verknüpft werden.

Die in dieser Arbeit gezeigten dreidimensionalen Szenen wurden mit *POV-Ray* erstellt. Das Programm POV-Ray, *Persistence of Vision Raytracer*, steht als Open Source zur Verfügung[6]. POV-Ray-Szenen werden in einer C-ähnlichen Syntax geschrieben, so dass sich eine Konvertierung der zweidimensionalen, topographischen Karten in dreidimensionale POV-Ray-Szenen programmatisch durchführen lässt. Als Grundlage der Prismen dienen die Polygone der Voronoi-Karte, die ihren Relationen zueinander getreu in das Povray-Koordinatensystem übernommen wurden.

Die Ergebnisse der hier vorgenommenen dialektometrischen Analysen liegen in den verschiedensten numerischen Bereichen, sowohl diskreten als auch kontinuierlichen, vor. Um diese Werte bei gleichbleibenden Relationen zueinander an die dreidimensionale Visualisierung durch Prismen anzupassen, können zwei Parameter verändert werden:

1. **Multiplikator:** Vor der Berechnung der Höhe der Prismen werden die numerischen Werte der zu visualisierenden Messreihe mit einem gleichbleibendem Wert multipliziert.

2. **Tiefen-Relation:** Nachdem die Formel 11.3 auf die Werte der darzustellenden Messreihe angewendet wurden, kann mittels eines weiteren fixen Parameters die Höhe der zu berechnenden Prismen beeinflusst werden.

Auf den Abbildungen 11.28 bis 11.30 sind die Bigrammdichten der Messpunkte des bulgarischen Datensatzes als unterschiedlich hohe Prismen dargestellt: Hohe Bigrammdichten entsprechen hohen Prismen, niedrige Bigrammdichten niedrigen Prismen. Höhe sowie Blickwinkel der virtuellen Kamera, aus deren Perspektive die Abbildungen erstellt wurden, bleiben unverändert. Die drei Abbildungen zeigen die so entstandene, dreidimensionale Abbildung Bulgariens aus drei verschiedenen Blickrichtungen: Von Süden nach Norden

[6] http://www.povray.org/

(Abbildung 11.28), von Osten nach Westen (Abbildung 11.29) und von Westen nach Osten (Abbildung 11.30). Die Parameter Multiplikator und Tiefen-Relation wurden dabei auf die Werte 150 bzw. 10 gesetzt. Eine weitergehende Analyse findet sich in Kapitel 11.7.3.

Abbildung 11.28: Reliefdarstellung der Bigrammdichten, Blickrichtung von Süden Richtung Norden

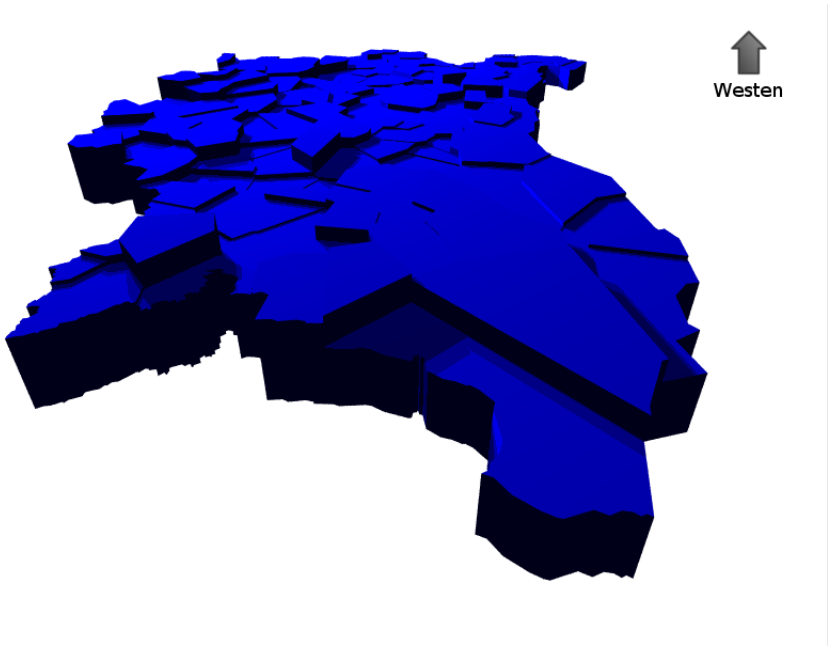


Abbildung 11.29: Reliefdarstellung der Bigrammdichten, Blickrichtung von Osten Richtung Westen

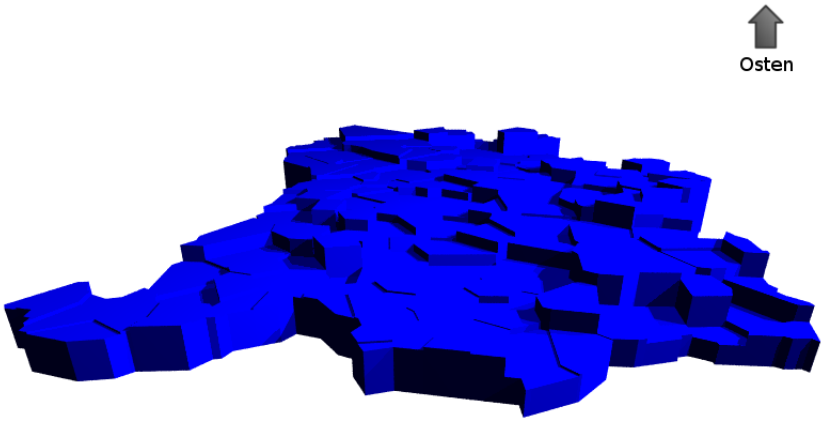


Abbildung 11.30: Reliefdarstellung der Bigrammdichten, Blickrichtung von Westen Richtung Osten

11.6.2 Multidimensional Scaling: Anwendung auf den bulgarischen Datensatz

Mittels *Multidimensional Scaling* (MDS) können die Dialekt-Distanzen zwischen den einzelnen Messpunkten eines Datensatzes in zwei- oder dreidimensionaler Weise visualisiert werden (siehe auch Kapitel 8.3). Dabei werden die geometrischen Distanzen zwischen den einzelnen Messpunkten so berechnet, dass sich Messpunkte mit ähnlichen Dialekten näher zueinander befinden als solche mit unähnlichen Dialekten. Die sich so ergebenden "Wolken" aus ähnlichen Messpunkten haben im Gegensatz zu den Clustern der Clusteranalyse keine scharf abgetrennten Grenzen untereinander.

Die Vielzahl der im MDS zur Verfügung stehenden Näherungsalgorithmen und andere Parameter machen es aus praktischen Gründen notwendig, in dieser Arbeit nur einige wenige der prinzipiell realisierbaren Kombinationen auf die Dialektdaten anzuwenden: Als Distanzmaß wurde wiederum die Euklidische Distanz und als Näherungsalgorithmus der Algorithmus von Kruskal gewählt.

Die Abbildungen 11.31, 11.32 und 11.33 zeigen Multidimensional Scaling, angewandt jeweils auf die Informations-, die Entropiewerte und die Bigrammdichten der bulgarischen Messpunkte. Somit stellen die Abbildungen die Ergebnisse der in dieser Arbeit vorgestellten aggregierenden Methoden dar. Alle drei Abbildungen sind nach 30 Durchläufen des Kruskal-Algorithmus angefertigt worden, gestartet wurde jeweils mit einer Zufallsverteilung der Messpunkte über die anzeigenden Fläche[7]. Die Grautöne, in denen die Messpunkte eingefärbt sind, entsprechen dem jeweiligen konkreten Wert in der dargestellten Messreihe. Zwischen Punkten mit ähnlichen Messwerten wurden von der Software Linien eingezogen.

Alle drei dialektometrischen Methoden ergeben eine ähnliche, hantelförmige Gestalt: Eine Linie, entlang der sich die Messpunkte gruppieren, deutet den idealen Verlauf an. An den Enden der (gedachten) Linie finden sich -

[7]Das hier gezeigte MDS wurde mit Hilfe der Orange-Software angefertigt. Hier gilt die zufallsverteilte Anfangsposition der Messpunkte bereits als erste Iteration, so dass die hier gezeigten Graphiken die Verteilung der Messpunkte nach 31 Iterationen in Orange darstellen, wovon 30 nach dem Kruskal-Algorithmus durchgeführt wurden.

mehr oder weniger ausgeprägt - größere Gruppierungen innerhalb der jeweiligen Messreihe.

Die Informationswerte (Abbildung 11.31) zeigen eine Zweiteilung der Messpunkte, am unteren Rand der Graphik zeigen sich stark heterogene Messpunkte, was zu einer starken Spreizung der Messpunkte über die anzeigende Fläche führt. Hier findet sich eine faktisch abgetrennte Gruppe von Messpunkten. Die Entropiewerte (Abbildung 11.32) zeigen im Vergleich zu den Informationswerten einen fast ausgeglichenen und linearen Verlauf mit größeren Gruppen an beiden Enden der Skala. Hier zeigt sich wieder der höhere Abstraktionsgrad der Entropiewerte: Der Verlauf des MDS ist geradlinieger, die Unterschiede zwischen den Messpunkten sind stärker nivelliert als im MDS der Informationswerte. Beide Gruppen an den Enden der Graphik weisen eine kleinere Spreizung auf als die der Informationswerte.

Das MDS der Bigrammdichten (Abbildung 11.33) weist ebenfalls eine Zweiteilung der Messpunkte mit größeren Gruppen an beiden Enden der Graphik auf. Allgemein weist das MDS der Bigrammdichten Ähnlichkeiten mit dem der Informationswerte auf, allerdings liegen die Messpunkte näher aneinander, die Streuung entlang der idealen Achse ist geringer als bei den Informationswerten.

Die Abbildung 11.34 schließlich vereint die drei Messreihen Information, Entropie und Bigrammdichten in einem einzigen MDS. Es wurde wiederum die Euklidische Distanz (siehe Formel 8.3) zur Berechnung der Abstände der einzelnen Messpunkte und der Kruskal-Algorithmus zur Durchführung des MDS gewählt. Da die drei Messreihen sich in sehr unterschiedlichen Zahlenbereichen befinden, wurden die Daten dahingehend normalisiert, dass sie sich alle im Bereich der Entropie-Werte befinden. Hierzu wurden die Informations-Werte mit 500 dividiert und die Bigrammdichten mit 80 multipliziert.

Im Unterschied zu den oben aufgeführten MDS-Plots sind hier 245 Iterationen durchgeführt worden[8]. Da nun drei unterschiedliche Messreihen in einem einzigen Plot vereint sind, sind die Messpunkte nicht eingefärbt worden. Deutlich zu erkennen sind wieder zwei Wolken von zusammenhängenden

[8]Die Integration von drei Messreihen in ein einziges MDS erzeugt komplexe geometrische Relationen zwischen den einzelen Messpunkten. Diese erfordern mehr Iterationen als die Darstellung einer einzigen Messreihe.

Messpunkten, die wiederum von keiner klaren Grenze mehr voneinander getrennt werden. Zwischen den beiden Wolken und um diese herum liegen weitere Messpunkte, die nicht der einen oder anderen Wolke zugeordnet werden können. Die Auffächerung der Messpunkte ist wesentlich weitergehender als bei Betrachtung der Messreihen in separaten MDS-Plots. Dies verdeutlicht die Unterschiede zwischen den Ergebnissen der einzelnen Methoden: Werden separate MDS-Visualisierungen durchgeführt, so ergeben alle Methoden ein ähnliches Bild (hantel-förmige Anordnung der Messpunkte, Abbildung 11.31, 11.32 und 11.33). Werden die Werte der drei aggregierenden Methoden nun wiederum zu einer Distanzmatrix aggregiert, so ergibt sich ein anderes Bild. Auch hier sind die zwei Hauptgruppen klar erkennbar, es existieren allerdings mehr Ausreißer, die wiederum weiter um die Gruppen herum streuen und sich nicht mehr nur an den Außenrändern befinden. Dies deutet daraufhin, dass die drei einzelnen Messreihen, obwohl sie für sich alleine betrachtet ähnliche (hantelförmige) Strukturen aufweisen, bei gleichzeitiger Betrachtung durchaus Unterschiede zueinander aufweisen. Dennoch ist die immer wiederkehrende Einteilung der bulgarischen Dialekte in zwei große Gruppen mit zusätzlichen Ausreißern erhalten geblieben. Zu einer weitergehenden, geographischen Interpretation, siehe Kapitel 11.7.2.

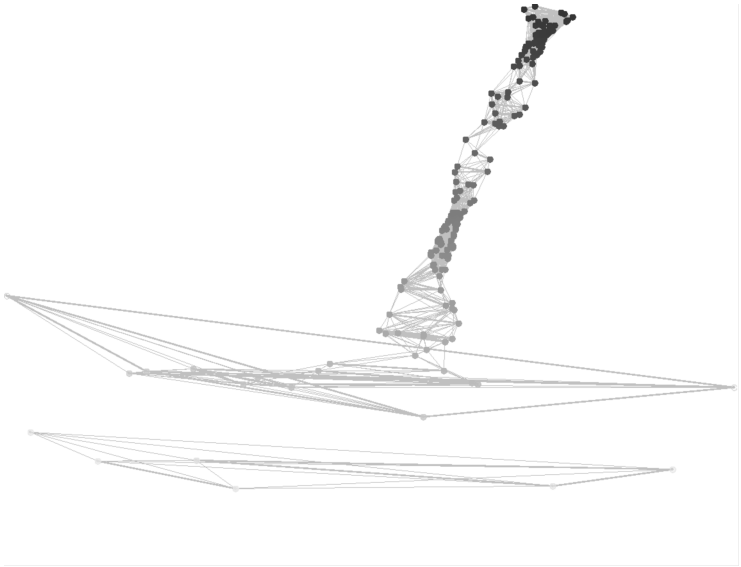


Abbildung 11.31: MDS der *Informationswerte* der Messpunkte, Algorithmus Kruskal, 30 Iterationen

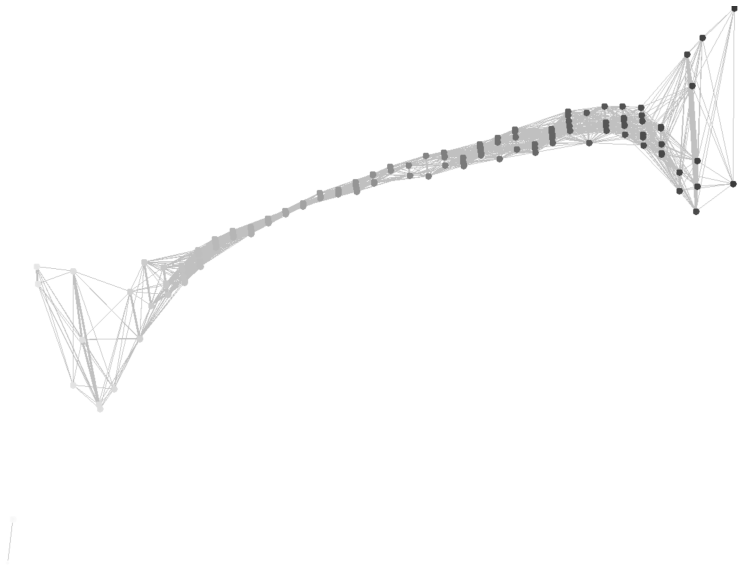


Abbildung 11.32: MDS der *Entropiewerte* der Messpunkte, Algorithmus Kruskal, 30 Iterationen

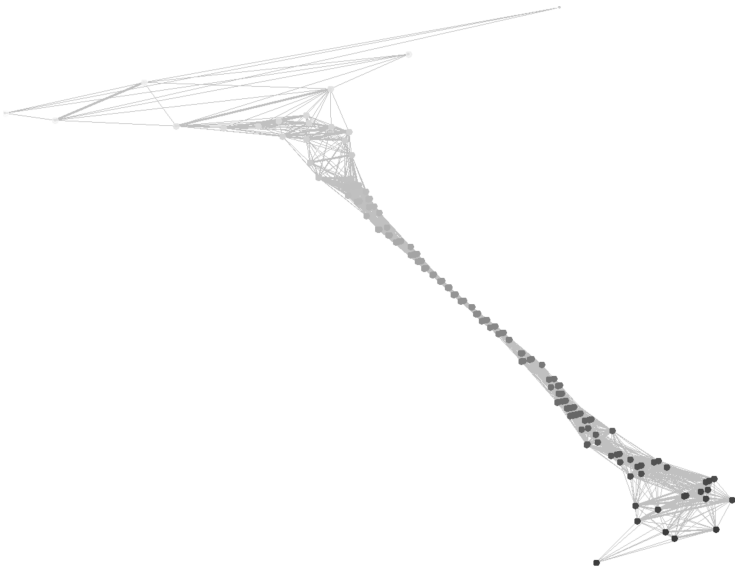


Abbildung 11.33: MDS der *Bigrammdichten* der Messpunkte, Algorithmus Kruskal, 30 Iterationen

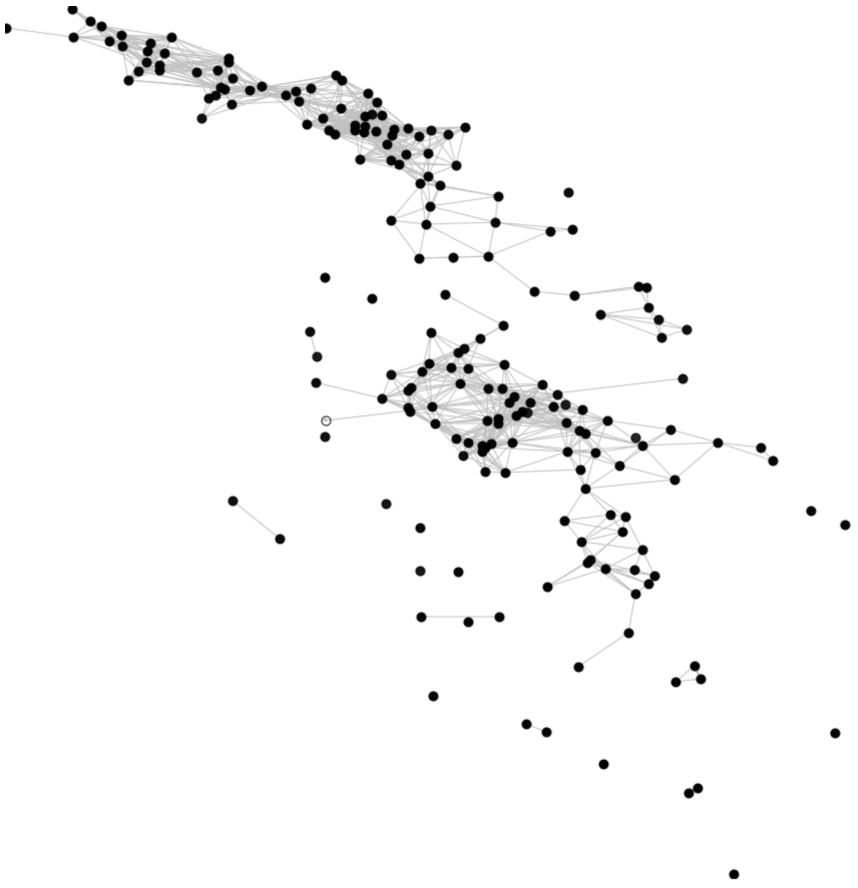


Abbildung 11.34: MDS der aggregierten Informations-, Entropiewerte und Bigrammdichten der Messpunkte, Algorithmus Kruskal, 245 Iterationen

11.7 Vergleich: Dialektologie und Dialektometrie

In diesem Kapitel sollen die Ergebnisse der in dieser Arbeit angewendeten dialektometrischen Methoden mit den Aussagen der bulgarischen Dialektologie verglichen werden. Hierzu dient die Karte von Stojko Stojkov und die dort eingezeichneten Dialektareale als Grundlage (siehe Kapitel 10.3 und hier die Karte in Abbildung 10.3). Die auf der Stojkovschen Karte dargestellten sechs großen Dialektareale stellen den "Maßstab" dar, mit dem die dialektometrischen Ergebnisse verglichen werden sollen. Dies erfolgt in Form von Scatterplots, Multidimensional Scaling und Reliefdarstellungen. Abschließend werden die manuell nachgezeichneten Isoglossen der Stojkovschen Karte über eine Auswahl der in diesem Kapitel erstellten Voronoi-Karten gelegt (Isoglossenvergleich).

11.7.1 Scatterplots

Die Abbildungen 11.36, 11.37 und 11.38 zeigen dieselben Daten wie der Scatterplot in Abbildung 11.15. Hinzugekommen ist mit der Einteilung der Dialekte nach Stojkov eine weitere Datenreihe, die zweifach visualisiert worden ist (in den Graphiken als "Dialektareale" markiert):

- Erstens durch die Farben der einzelnen Messpunkte: Deren Einfärbung entspricht nun den sechs Dialektarealen nach Stojkov. Die grünen Punkte befinden sich im Westen, die roten im Osten Bulgariens.

- Desweiteren wurden die Messpunkte in der dritten Dimension nach Stojkov in sechs Ebenen hintereinander angeordnet (Abbildung 11.35). Diese dritte Dimension ist in Abbildung 11.37 sichtbar: Der dreidimensionale Kubus, der auf den ersten zwei Dimensionen die Information und die Entropie anzeigt, wurde um 90 Grad gedreht, so dass nun ein seitlicher Blick die Entropie-Werte und die sechs hintereinander geordneten Ebenen nach Stojkov anzeigt. Abbildung 11.37 zeigt diese sechs Ebenen aus einer seitlichen Perspektive, sichtbar sind die Entropie und

die Dialektareale. Abbildung 11.38 zeigt die Dialektareale und die Information. In frontaler Ansicht ergibt sich die Abbildung 11.36.

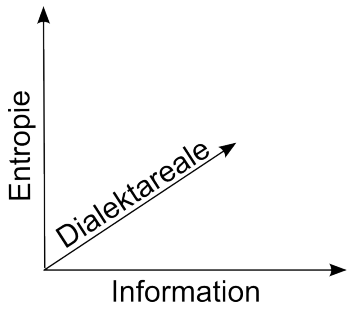


Abbildung 11.35: Schematische Darstellung der Anordnung der drei Messreihen in den Abbildungen 11.36 bis 11.38

In den Abbildungen 11.37 und 11.38 zeigt sich, dass die Informationswerte im Westen Bulgariens (grün) besser mit der Einteilung nach Stojkov übereinstimmen als die Entropiewerte. Hier verteilen sich die Messpunkte über eine kleinere Fläche als das bei den Entropiewerten der Fall ist. Je geringer die Streuung der Messpunkte innerhalb der sechs Ebenen, desto besser stimmen die Dialekteinteilungen der beiden einander gegenübergestellten Messreihen miteinander überein. Im Osten hingegen sind beide informationstheoretischen Methoden in etwa gleich auf, die Messpunkte streuen hier in ähnlichem Maße. Dies könnte wiederum an der geringeren Datendichte des Buldialect-Datensatzes im Osten Bulgariens liegen.

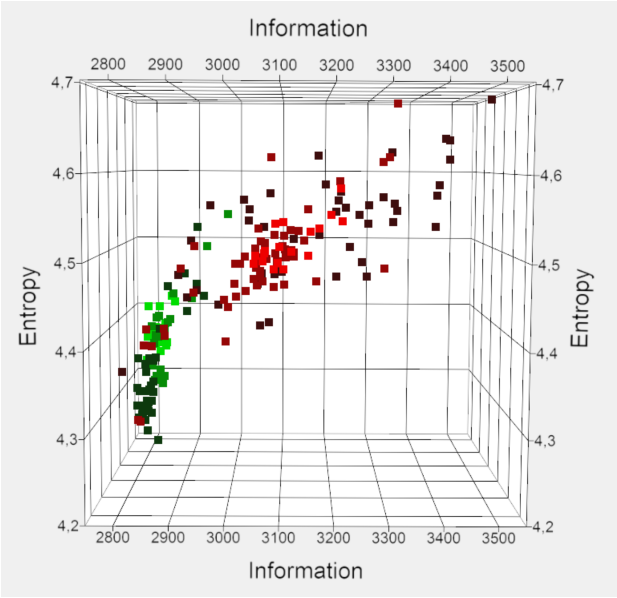


Abbildung 11.36: Dreidimensionaler Scatterplot von Information, Entropie und Dialektologie: Perspektive von oben

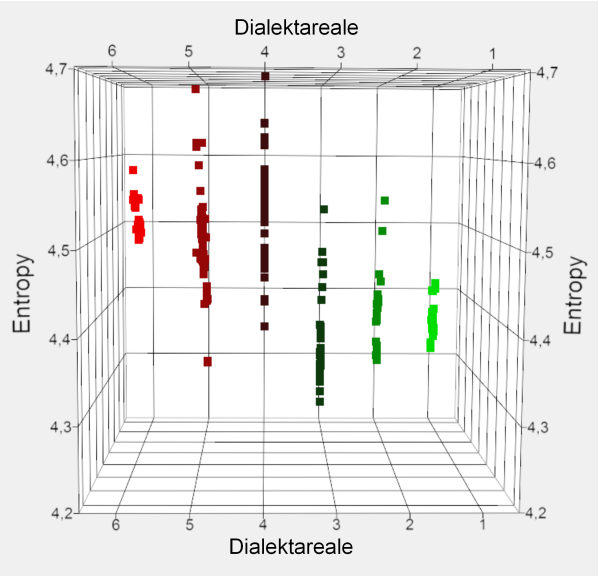


Abbildung 11.37: Dreidimensionaler Scatterplot von Information, Entropie und Dialektologie: Perspektive von der Seite, Entropie und Dialektareale

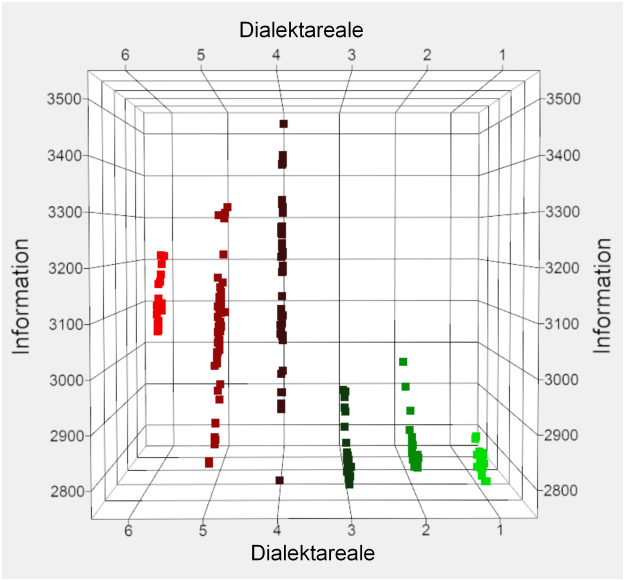


Abbildung 11.38: Dreidimensionaler Scatterplot von Information, Entropie und Dialektologie: Perspektive von der Seite, Information und Dialektareale

Dialekt-Nr.	Beschreibung	Farbe	Bereich
1	Übergangsdialekte zum Serbischen	Blau	W
2	Nord-West Dialekte	Rot	W
3	Süd-West Dialekte	Grün	W
4	Rupskian Dialekte	Orange	R
5	Balkan Dialekte	Gelb	O
6	Moesische Dialekte	Violett	O

Tabelle 11.8: Sechs Dialektgebiete nach Stojkov, eingefärbt und den Regionen Bulgariens zugeordnet (W = Westen, O = Osten, R = Rhodopen

11.7.2 Multidimensional Scaling

Die Abbildung 11.39 zeigt dasselbe Multidimensional Scaling wie Abbildung 11.34, allerdings wurden nun zusätzlich zu den drei Datenreihen (Informations- und Entropiewerte sowie Bigrammdichten) eine Einfärbung nach den sechs Stojkovschen Dialektareale vorgenommen. Die Farben der Messpunkte wurden vergeben wie in Tabelle 11.8 angegeben.

Es zeigt sich, dass das MDS der drei aggregierten dialektometrischen Methoden gut übereinstimmt mit der dialektologischen Isoglosseneinteilung der bulgarischen Dialekte nach Stojkov. Die westlichen Dialektgruppen 1, 2 und 3 (mit den Farben blau, rot und grün) bilden die obere Gruppe des MDS-Plots in Abbildung 11.39. Die Südwest-Dialekte (grün) bilden hier innerhalb der oberen Wolke wiederum den oberen Bereich, die Nordwest-Dialekte sowie die Übergangsdialekte zum Serbischen stehen darunter (blau und rot). Die östlichen Dialektareale bilden die untere Gruppe im MDS (violett und gelb). Die orange eingefärbten Dialekte der Rhodopen bilden die Ausreißer um die Gruppe der östlichen Dialekte herum. Sie stehen somit dem östlichen Bereich näher als dem westlichen, was auch ihrer geographischen Lage entspricht.

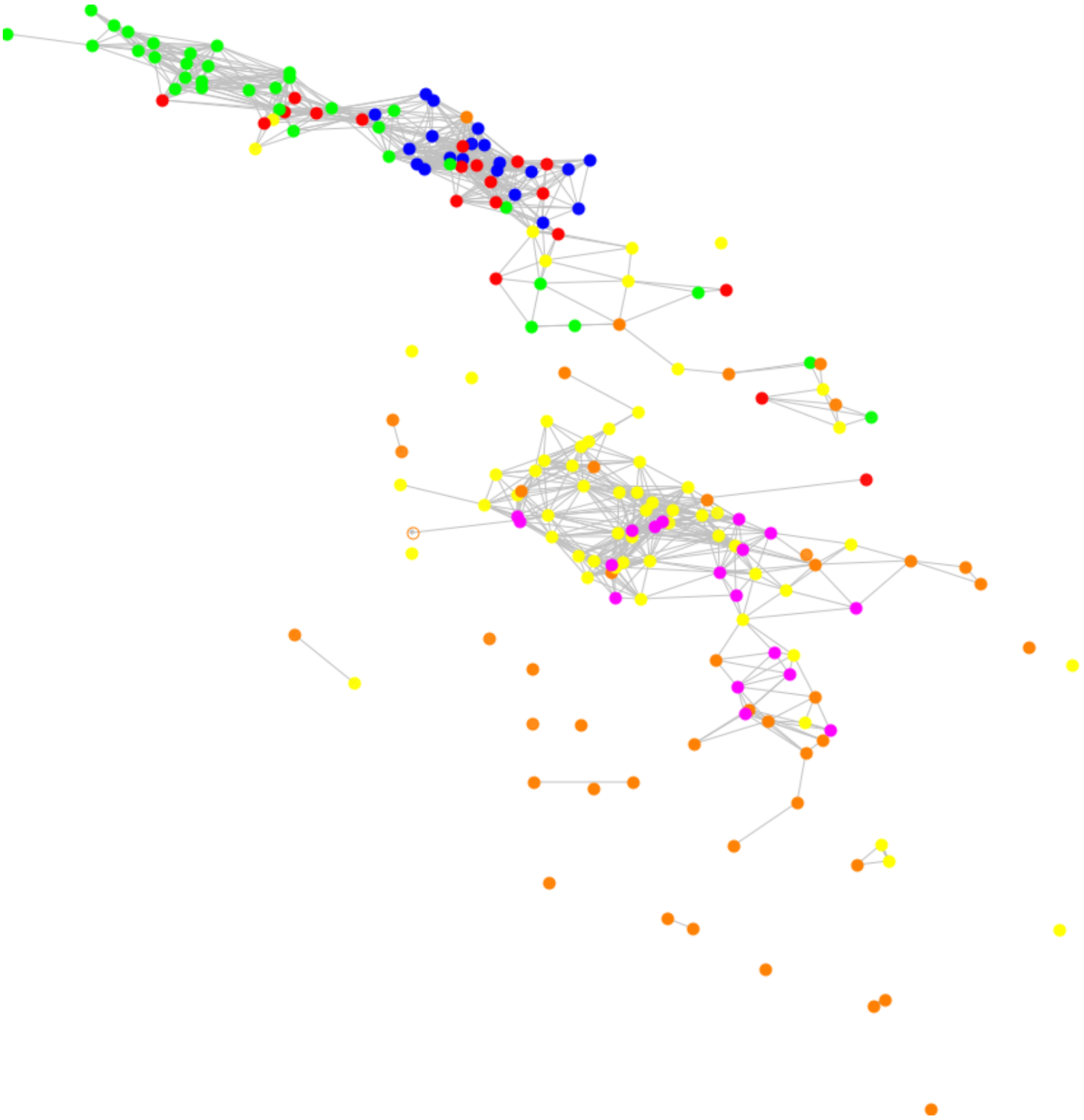


Abbildung 11.39: MDS der drei aggregierenden Methoden und dialektologische Einteilung nach Stojkov: Die obere Gruppe (blaue, rote und grüne Messpunkte nach Stojkov) stellt die westlichen Dialektareale dar, die untere Gruppe die östlichen (violette und gelbe Messpunkte). Die heterogenen Dialekte der Rhodopen (orange) streuen um die östlichen Dialekte herum

11.7.3 Reliefdarstellungen

Die Reliefdarstellung dialektometrischer Analyseergebnisse in Kombination mit der Einfärbung der entstandenen Prismen erlaubt die Kombination zweier voneinander unabhängiger Messreihen. Die eine Messreihe wird visualisiert durch die exakte Anordnung der Prismen in korrekten geographischen Relationen und Koordinaten zueinander. Die Visualisierung der zweiten Messreihe erfolgt durch die Einfärbung der Prismen.

Die Abbildung 11.40 zeigt verschiedene Reliefdarstellungen der Bigrammdichten der bulgarischen Messpunkte, berechnet wie beschrieben in Kapitel 11.5. Die Höhe der Reliefs entspricht den Werten der Bigrammdichten bzw. den jeweiligen aus den Bigrammdichten berechneten Clustern. Messpunkte, die dem gleichen Cluster angehören und geographisch nebeneinander liegen, bilden somit ein zusammenhängendes Plateau gleicher Höhe. Zusätzlich sind die Prismen nach den sechs Stojkovschen Dialektarealen eingefärbt (vergleiche Tabelle 11.8, aus Gründen der besseren Sichtbarkeit wurden hier andere Farben als im MDS verwendet). Diese Einfärbung ist in allen Reliefdarstellungen fix und wird nicht verändert. Je besser die Übereinstimmung zwischen den Bigrammdichten und den Stojkovschen Dialektarealen ist, desto mehr Messpunkte auf gleicher Höhe werden in der dem Dialektareal entsprechenden Farbe eingefärbt sein. Idealerweise würden sich Plateaus gleicher Höhe ergeben, die durchgängig in einer Farbe eingefärbt sind. Abweichungen zwischen der Einfärbung und der Reliefhöhe entsprechen Abweichungen zwischen Stojkovschen Dialektareal und der Bigrammdichte. Auf diese Weise ist eine direkte Gegenüberstellung der beiden Datenreihen - Bigrammdichten und Stojkovsche Dialektareale - in einer einzigen Visualisierung möglich.

Da zur Berechnung der Reliefs die Bigramm*dichten* vewendet wurden, sind die hier gezeigten Werte bzw. Ergebnisse des Clustering nicht identisch mit den in Kapitel 11.5.2 gezeigten Clusteringergebnissen, da dort direkt Distanzwerte aus den verwendeten Matrizen errechnet und geclustert wurden.

Die Abbildung 11.40 setzt sich aus zwei Spalten mit drei Zeilen mit insgesamt 6 einzelnen Graphiken zusammen (A1 bis C2). Die linke Spalte enthält eine Sicht auf das bulgarische Staatsgebiet von Süden in Richtung Norden. Die Graphiken in der zweiten Spalte zeigen dasselbe Gebiet aus der Perspek-

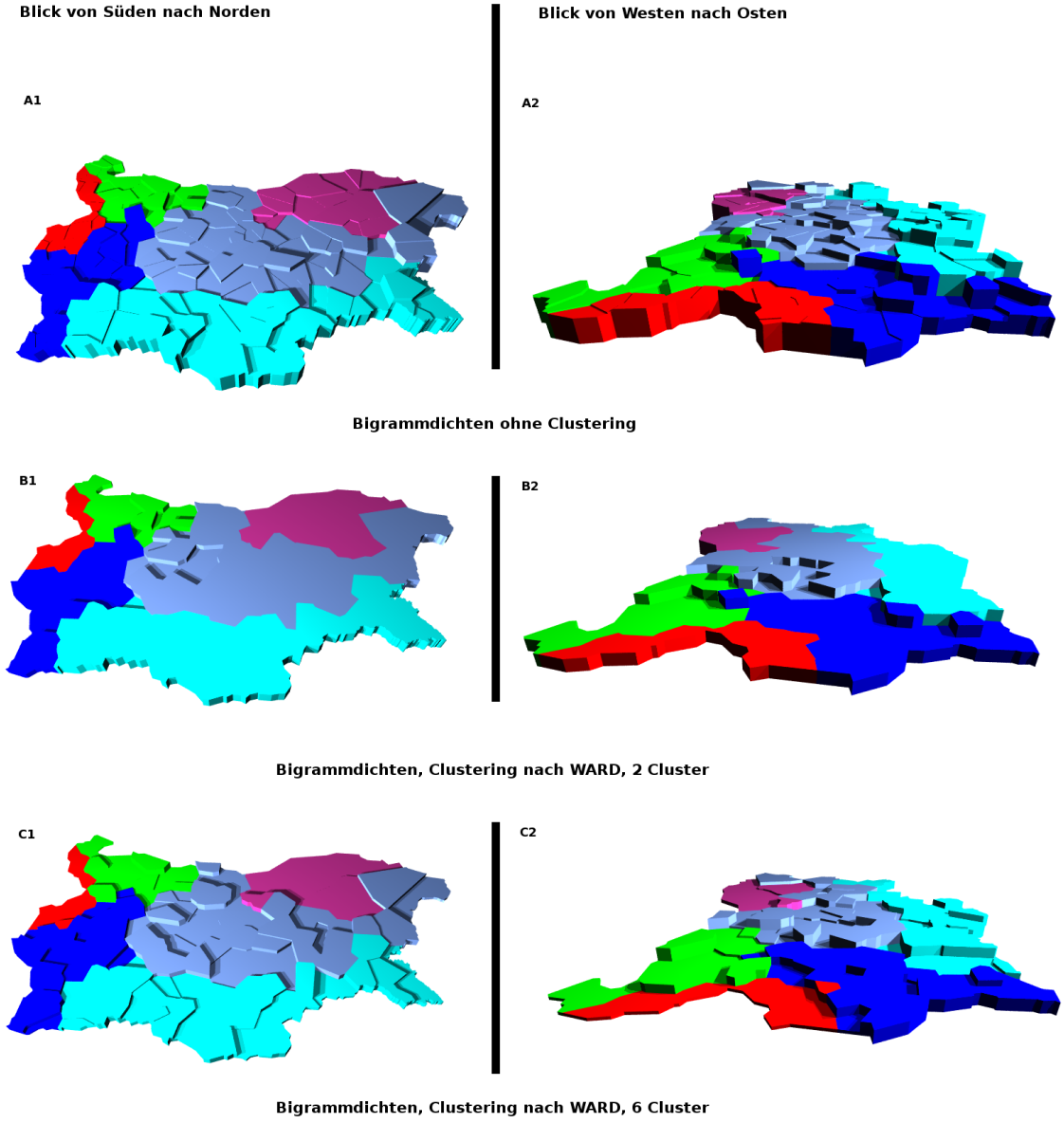


Abbildung 11.40: Reliefdarstellungen der Bigrammdichten. Die Einfärbung der Prismen entspricht den sechs Stojkovschen Dialektarealen

	Multiplikator	Tiefen-Relation
Bigrammdichten (A1, A2)	150	10
2 Cluster (B1, B2)	1	10
6 Cluster (C1, C2)	1	2,5

Tabelle 11.9: Parametereinstellungen der Reliefdarstellungen in Abbildung 11.40

tive von Westen nach Osten.

Die erste Zeile zeigt die Bigrammdichten der Messpunkte ohne Clustering, jeder Messpunkt hat faktisch eine andere Höhe (A1 und A2). Die Zeilen zwei bis drei visualisieren die Bigrammdichten nach angewandtem Clustering (B1 bis C2). Es wurde jeweils die Methode WARD benutzt, mit einer Clusteranzahl von 2 und 6 in den Zeilen zwei und drei[9]. Die Graphiken in der ersten Zeile (Bigrammdichten ohne Clustering) repräsentieren somit einen numerisch kontinuierlichen Bereich zwischen 0 und 1, wohingegen die weiteren Zeilen diskrete Zahlenbereiche von 1 bis zur Anzahl der jeweils berechneten Cluster darstellen. Tabelle 11.9 listet die verwendeten Parameter (Multiplikator und Tiefen-Relation) zur Berechnung der Reliefs auf.

Die Reliefs weisen mehrere charakteristische Strukturen auf:

- Die Bigrammwerte steigen von Westen nach Osten hin an. Dies ist besonders deutlich sichtbar in den Graphiken der rechten Spalte (Blickrichtung von West nach Ost): Die Reliefs steigen hier von niedrigen zu größeren Höhen an (siehe hierzu auch Abbildung 11.41, die die Reliefs der Bigrammdichten von Westen aus einem flacheren Blickwinkel darstellt).

- Das Relief A2 weist neben dem von West nach Ost ansteigendem Niveau der Prismen einen weiteren, von Süden nach Norden verlaufenden dreistufigen Anstieg der Reliefs im westlichen Teil Bulgariens auf (Abbildung 11.41 zeigt eine flache Aufsicht auf die Prismen von Westen

[9]Das Clustering wurde mit der Software SAS JMP berechnet.

Richtung Osten). Die südlichen Prismen haben hier die geringsten Höhen, die Übergangsdialekte stellen einen Mittelwert dar und die nördlichen Dialekte weisen die höchsten Prismen auf. Die Prismen im Ostteil Bulgariens sind im Durchschnitt höher als die westlichen (in der Abbildung 11.41 im Hintergrund).

- Das Gebiet der Rhodopen ragt hervor, hier finden sich in einigen Messpunkten die höchsten Bigrammdichten und dementsprechend auch die höchsten Prismen (Graphik A2 und Abbildung 11.42). Im westlichen Bereich der Rhodopen weisen die Prismen eine weitestgehend einheitliche Höhe auf. Im östlichen Bereich der Rhodopen werden die Höhen der Prismen heterogener, hier weisen die Prismen die größten Unterschiede zueinander im gesamten Untersuchungsgebiet auf.

- In den Graphiken B1 und B2 werden zwei unterschiedliche Reliefhöhen dargestellt, entsprechend der Einteilung der Messpunkte in zwei Cluster. Es bilden sich zwei geographisch zusammenhängende Plateaus, deren Grenze zueinander größtenteils dem Verlauf der Yat-Linie nach Stojkov entspricht. Abweichungen ergeben sich an den Rändern der Dialektareale: Hier sind einige Messpunkte abweichend von der Stojkovschen Dialekteinteilung dem jeweils anderen Plateau zugeordnet worden. Desweiteren existiert eine nach Osten ragende Einbuchtung des westlichen, tiefer gelegenen Reliefs in der Mitte Bulgariens (Nord-Süd-Richtung).

- Auch die Reliefs der Clusteranalyse mit 6 Clustern zeigt einen ansteigenden Verlauf der Reliefs von Westen nach Osten. Hier fallen vor allem die Übergangsdialekte zum Serbischen auf, die bis auf zwei Messpunkte Übereinstimmung zwischen den Stojkovschen Dialektarealen und den Bigrammdichten aufweisen.

- Die Moesischen Dialekte im Nordosten (violette Einfärbung) erscheinen im Relief als relativ homogener Bereich, der in der Reliefhöhe allerdings ähnliche Höhe wie die umgebenden Balkan-Dialekte aufweist.

Die hier gezeigte Kombination aus Bigrammdichten und den Dialektarealen der Stojkovschen Karte weist eine weitgehende Übereinstimmung

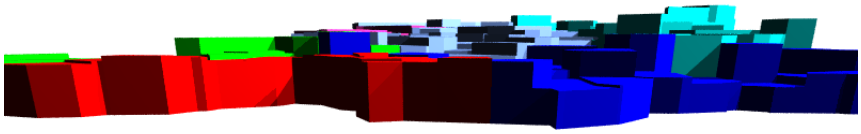


Abbildung 11.41: Das Relief der Bigrammdichten (ohne Clustering) mit Blickrichtung von Westen nach Osten, flacher Winkel

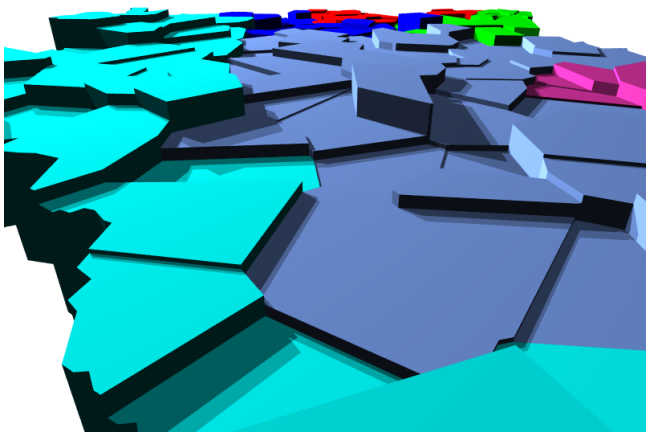


Abbildung 11.42: Blick von Osten, Position der Kamera auf dem höchsten Prisma

der beiden Datenreihen auf. Abweichungen ergeben sich hauptsächlich an den Rändern der Dialektareale, hier weichen einige Messpunkte in der Höhe ihrer Prismen von der Einfärbung der Stojkovschen Dialektareale ab. Dies zeigt sich in den Reliefs des Clustering mit 6 Clustern (C1 und C2): Die Übergangsdialekte zum Serbischen, hier rot eingefärbt nach der Stojkovschen Areal-Einteilung, bilden ebenfalls von der Höhe der Prismen her ein zusammenhängendes Plateau. Lediglich ein Messpunkt im Süden weicht ab und ist von seiner Höhe her den südwestlichen Dialekten (blau) zugeordnet worden, im Gegenzug ist ein Messpunkt des nordwestlichen Areals den Übergangsdialekten zugeordnet worden (grün, äußerster nördlicher Messpunkt). Eine ähnliche Situation zeigt sich im Bereich der Moesischen Dialekte. Diese stellen durchgängig ein Plateau dar, es weicht nur ein Messpunkt im Westen ab (Graphik C1).

Durch die dialektometrischen Methoden wird die Dialekt-Einteilung Bulgariens nach Stojkov bestätigt. Allerdings weisen die abweichenden Dialekte an den Grenzen der Dialektareale auf durchlässige bzw. nicht absolut fixe Grenze zwischen den Arealen hin. Dies kann auf zweierlei Weise interpretiert werden: An diesen Stellen findet ein Wandel in den bulgarischen Dialekten statt. Es werden Veränderungen aus den benachbarten Dialekten adaptiert, was diachron betrachtet schließlich zu einer anderen Dialekteinteilung führen wird. Darüber hinaus ist die Datenbasis, anhand derer die Analysen in dieser Arbeit durchgeführt wurden, eine andere als die, auf deren Grundlage Stojkov die Einteilung Bulgariens vornahm. In den voneinander abweichenden Grenzziehungen spiegeln sich die (diachronen) Unterschiede zwischen den beiden Datensätzen wider. Andererseits könnte es sich bei den Übergängen zwischen den Dialekten um Dialektkontinua und keine isoglossenbasierten Grenzen handeln. In diesem Falle wären die bulgarischen Dialektareale fließend im Übergang zueinander. Der Übergang zwischen den Messpunkten wäre nicht abrupt und die Zuweisung der Messpunkte an den Grenzen kann sich auch von Methode zu Methode ändern.

Methode	Abweichungen
Information	8
Entropie	6
Vektoranalyse (e)	7
Aggregierte Bigramme	8

Tabelle 11.10: Anzahl der Messpunkte, die nach der angegebenen dialektometrischen Methode zusätzlich dem westlichen Bereich zugeordnet worden sind, laut Isoglossenverlauf aber im Osten liegen

11.7.4 Isoglossenvergleich

In Abbildung 11.43 wurden die aus der Stojkovschen Karte extrahierten Isoglossen (siehe hierzu auch Abbildung 10.5) auf vier Karten der dialektometrischen Analysen angewendet. Dargestellt sind der Informationsgehalt der Messpunkte, die Entropie, die Vektoranalyse anhand des XSampa-Codes "e" sowie die aggregierten Bigramme. Alle vier Karten zeigen das Clustering nach der Methode Ward mit jeweils 12 Clustern. Die Isoglossen der Stojkovschen Karte wurden über die Polygone der Voronoi-Karte eingezeichnet.

Zwischen den einzelnen dialektometrischen Ergebnissen und der Isoglosseneinteilung ergeben sich Gemeinsamkeiten, aber auch Unterschiede. So ist die Yat-Linie als Trennung zwischen dem Osten und dem Westen Bulgariens in allen dialektometrischen Analysen deutlich erkennbar, im Gegensatz zu der stojkovschen Isoglosseneinteilung aber nach Osten verschoben (siehe Tabelle 11.10).

Die Übergangsdialekte zum Serbischen sind in der Vektoranalyse und den aggregierten Bigrammen gut sichtbar, in der Entropie-Analyse erscheinen sie weiter nach Osten hin ausgedehnt.

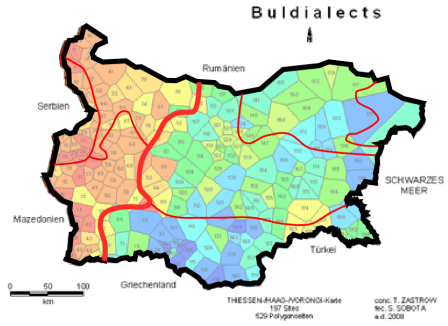
Die weitere Einteilung des Westens in einen Nord- und einen Südteil ist lediglich in der Bigramm-Analyse klar abgezeichnet.

Die Moesischen Dialekte im Nordosten sind auf keiner der dialektometrischen Karten erkennbar, allerdings ist die Datendichte des Buldialects Datensatzes im Nordosten auch am geringsten. Lediglich in der Bigramm-Analyse bildet der dunkelblaue Cluster ein zusammenhängendes Gebiet, wel-

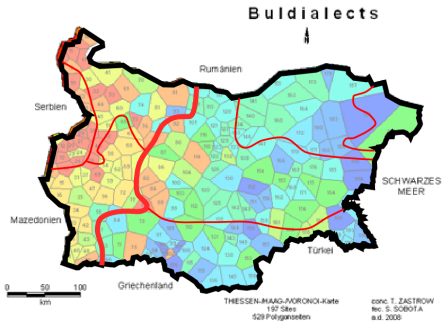
ches in etwa der Stojkovschen Abgrenzung der moesischen Dialekte entspricht.

Das heterogene Gebiet der Rhodopen ist in den Visualisierungen der informationstheoretischen Methoden (Entropie und Information) nicht sichtbar, ansonsten folgt es weitestgehend dem stojkovschen Isoglossenverlauf. Im mittleren Bereich ist das Dialektareal häufig unterbrochen, hier verläuft auch die Isoglosse scharf an der südlichen Grenze Bulgariens. Unterschiedlich stellt sich das westliche Ende der Rhodopen dar: In der Visualisierung der aggregierten Bigramme ragt der westliche Cluster über die Isoglosse hinaus in den Osten hinein. Die anderen dialektometrischen Einteilungen folgen weitestgehend der stojkovschen Isoglosse.

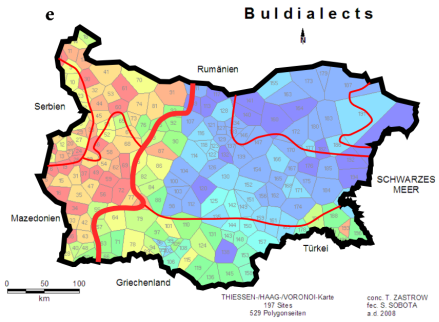
Die hier geschilderten Abweichungen der dialektometrischen Analysen von der Stojkovschen Isoglosseneinteilung können unterschiedliche Gründe haben: Zum einen sind die Karten von unterschiedlichem Typus, andererseits hätte eine andere Parametrisierung des Clustering unter Umständen besser passende Ergebnisse produziert. Am nähesten an die Isoglossen-Einteilung kommt die Karte der aggregierten Bigramme. Hier sind die einzelnen Dialekte innerhalb der Areale am homogensten und es ist die einzige Analyse, die die Einteilung der westlichen Dialekte in Übergangs-, Nord- und Süd-Dialektgebiete klar aufzeigt. Generell zeichnet sich ein im nördlichen Bereich leicht nach Osten verschobener Verlauf der Yat-Linie ab. Die westlichen Dialektgebiete um die Hauptstadt Sofia herum konnten sich somit auf Kosten der östlichen Dialektgebiete erweitern.



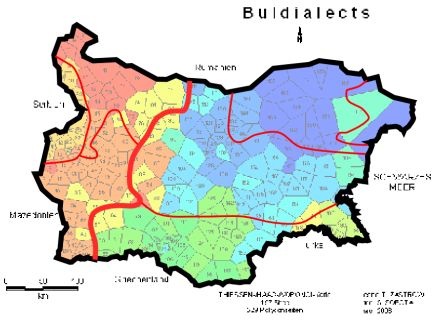
Information



Entropie



Vektoranalyse
(XSampa-Code "e")



Bigramme, aggregiert

Abbildung 11.43: Die Isoglossen der Stojkovschen Karte, eingezeichnet über die Cluster-Ergebnisse

11.7.5 Zusammenfassung

Die in dieser Arbeit mittels dialektometrischer Methoden gewonnenen Erkenntnisse über die bulgarische Dialekteinteilung bestätigen größtenteils die Ergebnisse der Dialektologie (siehe hierzu auch Kapitel 10.3). Die Gemeinsamkeiten überwiegen, an einigen Stellen lassen sich allerdings auch Abweichungen bzw. detailliertere Definitionen der einzelnen Dialektgebiete vornehmen. Die gefundenen Abweichungen lassen sich durch eine Vielzahl möglicher Gründe erklären:

- Verwendung unterschiedlicher Datensätze, auf deren Grundlage die jeweiligen Analysen durchgeführt wurden.
- Die von Stojko Stojkov erstellte Karte wurde von Hand gezeichnet, wohingegen die in dieser Arbeit verwendete Karte eine automatisch erstellte Voronoi-Karte ist.
- Die verwendeten Analysemethoden wie Clusteringalgoritmen etc. sowie die darauf aufbauenden Visualisierungen können durch eine Vielzahl unterschiedlicher Parameter konfiguriert werden. Je nach gewählter Einstellung dieser Parameter können sich unterschiedliche Ergebnisse in der Einteilung der Dialektareale ergeben.
- Wie bereits oben erwähnt, sind die meisten der in dieser Arbeit vorgestellten Methoden und Visualisierungen parametrisierbar. Dies bedeutet, dass eine andere Einstellung der Parameter unter Umständen auch eine größere Nähe zu Stojkovs Karte hätte ergeben können. Aus praktischen Gründen können in dieser Arbeit nur eine begrenzte Zahl unterschiedlicher Parametrisierungen berücksichtigt werden.

Von kleineren Abweichungen abgesehen, ergeben sich größtenteils Übereinstimmungen zwischen der Dialektologie und der Dialektometrie. In allen dialektometrischen Untersuchungen spielt die Yat-Linie bzw. die durch sie vorgenommene Einteilung des bulgarischen Sprachraums in einen Ost- und einen West-Teil die wichtigste Rolle. Sie ist jederzeit klar erkennbar, sowohl Intervall-Algorithmen als auch Clustering-Verfahren machen sie bei einer Anzahl von zwei Klassen bzw. Clustern sichtbar. Dabei ist festzuhalten, dass die

Yat-Linie in den dialektometrischen Analysen etwas weiter Richtung Osten verschoben ist.

Die Übergangsdialekte zum Serbischen sind ebenfalls in den meisten Analysen klar erkennbar. Sind sie nicht in einen eigenen Bereich abgetrennt, so werden sie den westlichen Dialekten zugeschlagen. Der übrige Bereich Westbulgariens ist häufig weiter unterteilt. Hier finden sich neben der traditionellen Nord-/Süd-Einteilung auch weitere Muster.

Die Aufteilung der östlichen Dialekte in Moesische und Balkan-Dialekte tritt nicht in allen Analysen klar zutage. Ein Grund hierfür kann in der relativ geringen Datendichte des Buldialect Datensatzes im Nordosten liegen: Die Messpunkte liegen hier weiter auseinander als im restlichen Teil des untersuchten Gebietes. Dies führt zu größeren Polygonen auf der Voronoi-Karte. Gut zu sehen ist die Unterteilung in Moesische und Balkan-Dialekte auf den Karten der Bigramm-Analyse (siehe Abbildungen 11.24, 11.25 und 11.26). Auf der Stojkovschen Karte werden die Balkan-Dialekte kurz vor dem Schwarzen Meer von einem nach oben reichenden Ausbrecher der Rupskian-Dialekte unterbrochen. In den dialektometrischen Untersuchungen erscheint die am Schwarzen Meer gelegene, abgetrennte Region häufig als ein einzelner, aus dem Rahmen fallender Messpunkt.

Ebenfalls meistens klar erkennbar sind die im Süden gelegenen Rupskian-Dialekte im Gebiet der Rhodopen. Sie ziehen sich von der Yat-Linie im Westen bis hin zum Schwarzen Meer im Osten, wobei hier an der Südost-Grenze zur Türkei häufig eine Dialekt-Insel mit zum Westen gehörigen Dialekten erkannt wird. Die westliche Spitze des Rhodopengebirges wird in seltenen Fällen dem westlichen Teil Bulgariens zugeordnet. Generell liefern die verschiedenen dialektometrischen Methoden in dieser Region leicht voneinander abweichende Ergebnisse:

- Der südliche Teil Bulgariens wird als eigenständige Dialektregion aufgefasst (Bigramm-Analyse).
- Bei geringer Cluster- bzw. Klassenanzahl werden die Rhodopen entweder dem Westen oder dem Osten zugehörig dargestellt.
- Rhodopen und Rupskian-Dialekte erscheinen als äußerst heterogen geprägtes Gebiet. Ihre Abgrenzung zu den westlichen und östlichen Dia-

lektgegenden ist stets klar ausgeprägt und deutlich sichtbar. Allerdings sind hier eigenständige Dialekte durchmischt mit Dialekten aus Westen oder Osten.

11.8 Vergleich: Informationstheoretische Methoden und Bigrammdichten

Informationstheoretische Methoden (siehe Kapitel 7) und die Bigrammdichten (Kapitel 6) sind aggregierende Methoden: Die gesamten Daten eines Messpunktes werden in die Analyse einbezogen und hieraus messpunktspezifische Kennzahlen errechnet. Die so entstandenen Reihen von Kennwerten können in Relation zueinander gestellt und miteinander verglichen werden.

In Abbildung 11.44[10] sind Entropie bzw. Information (X-Achsen) gegen die Bigrammdichten (Y-Achse) aufgetragen. Die Farben der einzelnen Messpunkte entsprechen wieder der sechsfachen Einteilung der bulgarischen Dialekte nach Stojkov. Die blauen Linien kennzeichnen den idealen, von der Software errechneten Verlauf der Datendistribution.

Entropie und Bigrammdichten (linke Graphik) bilden eine fast linear ansteigende Kurve, während der Verlauf der Kurve der Information in Relation zu den Bigrammdichten (rechte Graphik) unregelmäßiger verläuft und die einzelnen Messpunkte weiter um die Ideallinie herum streuen. Somit haben die Entropie-Werte der Messpunkte eine größere Ähnlichkeit zu den Bigrammdichten der Messpunkte als die Informationswerte.

Dies lässt sich begründen mit der unterschiedlichen Art und Weise, in der die jeweiligen Meßreihen erstellt wurden: Entropiewerte und Bigrammdichten wurden jeweils mit den Daten des entsprechenden zugrundeliegenden Messpunktes errechnet, wohingegen die Informationswerte auf Basis der Wahrscheinlichkeiten der XSampa-Codes in *allen* Messpunkten errechnet wurden. Die durch die Verwendung unterschiedlicher Farben gekennzeichnete Einteilung der Messpunkte nach Stojkov lässt in beiden Graphiken zwei Gruppen erkennen: Die grün bzw. rot markierten Messpunkte finden sich jeweils in zwei zusammenhängenden Bereichen im oberen bzw. unteren Bereich der Diagramme. Somit zeigt sich wieder die prinzipielle Zweiteilung der bulgarischen Dialekte

[10] Scatterplot, erzeugt mit JMP 8.

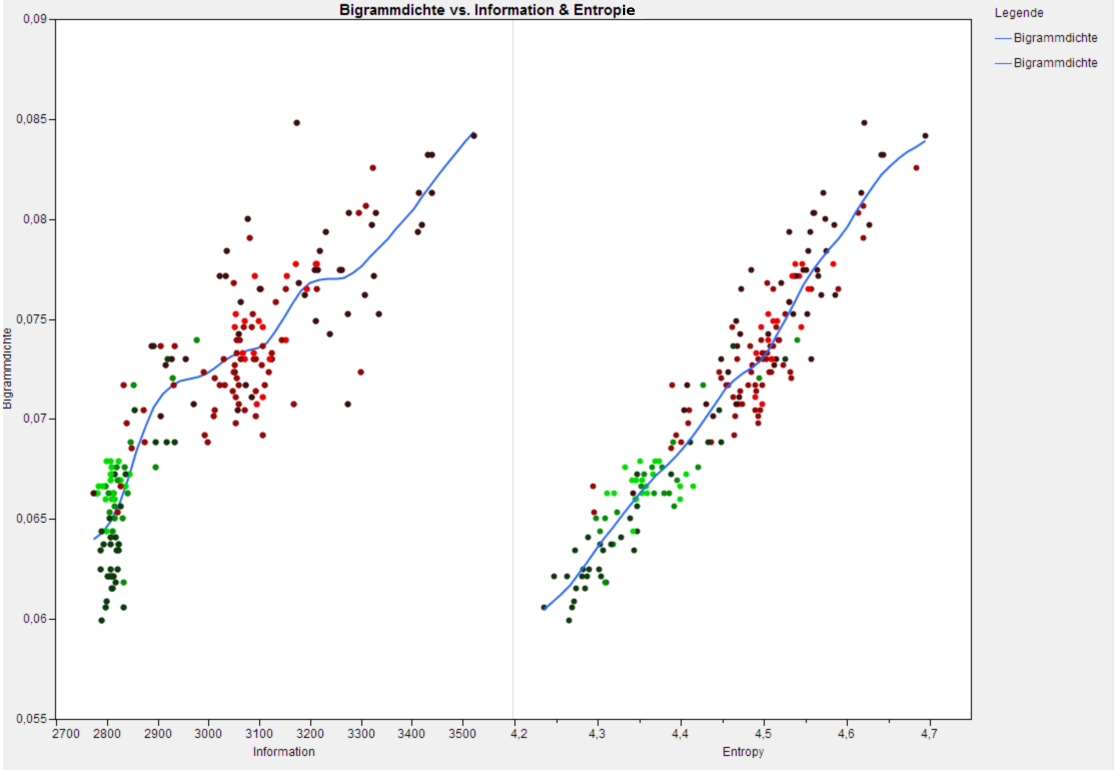


Abbildung 11.44: Scatterplot: Information und Entropie, aufgetragen gegen die Bigrammdichten

11.9 Zusammenfassung

Die hier gezeigten dialektometrischen Methoden sind alle geeignet, um die größten dem bulgarischen Datensatz innewohnenden Dialektareale aufzudecken. Abhängig von der verwendeten Methode und den jeweiligen feinjustierenden Parametern der Analysemethoden und der anschließenden Visualisierungen, werden mehr oder weniger Dialektareale identifiziert. Die den Westen vom Osten trennende Yat-Linie stellt dabei die wichtigste Trennungslinie dar - sie ist mehr oder weniger deutlich immer sichtbar. Hinzu kommen Übergangsdialekte zum Serbischen an der Westgrenze sowie eine westlich geprägte Enklave am südöstlichen Rand hin zur Türkei. Das südliche Mittelgebirge, die Rhodopen, setzen sich heterogen zusammen. Methodenabhängig werden sie mal dem Westen, mal dem Osten zugehörig oder als eigenständiges Dialektareal dargestellt. Zu diesen häufig sichtbaren, stark voneinander abgegrenzten Dialektarealen kommen weitere, schwächer abgegrenzte hinzu. Diese werden nur durch wenige oder in einigen Fällen sogar nur durch eine einzige dialektometrische Methode sichtbar gemacht.

Kapitel 12

Lexikalische Daten

Dialekte unterscheiden sich auf allen Ebenen der Sprache, auch auf lexikalischer. Im Gegensatz zu phonetisch variierenden Dialekdaten, in denen ein Lexem in verschiedenen Dialekten lediglich unterschiedlich betont ausgesprochen wird, unterscheiden sich lexikalische Dialektdaten in der Verwendung verschiedener Lexeme für ein und dasselbe Konzept. Damit weisen Dialekte, die sich lexikalisch voneinander unterscheiden, eine größere Distanz zueinander auf als solche, die sich lediglich auf phonetischer Ebene voneinander unterscheiden.

Bei lexikalisch differenzierenden Dialektdaten stellen die Wörter an sich die atomaren Einheiten dar. Im Gegensatz zu phonetischen Daten lassen sie sich nicht weiter aufsplitten. So lassen sich Ähnlichkeiten auf Wortebene nur binär - identisch oder nicht identisch - bestimmen, eine feinere Abstufung der Ähnlichkeiten zueinander ist nicht möglich. Somit ist auch die Anzahl der zu untersuchenden Elemente eingeschränkt.

Um in der Dialektometrie Verwendung finden zu können, müssen auch lexikalische Dialektdaten einem hohen Maß an Formalität entsprechen. Die untersuchten Konzepte müssen in allen Dialekten vorhanden sein: Wörter, die regionalspezifische Kontexte beschreiben, können nicht verwendet werden. So haben beispielsweise an einer Meeresküste beheimatete Dialekte eher eine wasserspezifische Terminologie als Dialekte in einer Bergregion.

Der morphologisch oder syntaktisch variierende Gebrauch eines Lexems kann zwar ebenfalls dialektal begründet sein, eine entsprechende Analyse darf

allerdings nicht mit der Analyse unterschiedlicher Lexeme vermischt werden.

Von den in dieser Arbeit vogestellten Methoden können zwei zur Analyse lexkalischer Dialektdaten herangezogen werden:

- *Phänomenologische Darstellung* der einzelnen Wörter auf topographischen Karten. Dies entspricht weitgehend dem klassischen Sprachatlas der Dialektologie.
- Bestimmung des *Relativen Identitätswerts* oder verwandter Ähnlichkeitsmaße auf Wortebene.

12.1 Sprachatlas

Im bulgarischen Datensatz liegen 114 lexikalisch variierende Konzepte vor, die sich jeweils auf einer topographischen Arbeitskarte darstellen lassen. Die Anzahl der verschiedenen Lexeme pro Konzept reicht von einem (keinerlei dialektale Varianz), bis hin zu 35 verschiedenen Lexemen.

Im Gegensatz zu den phonetischen Daten geben die Arbeitskarten der lexikalischen Daten kein einheitliches Bild ab. In einigen Karten finden sich Übereinstimmungen zu den phonetischen Daten, andere Karten weisen neue Strukturen auf. Allerdings bilden sich auch hier in den meisten Karten zusammenhängende Gebiete heraus.

Die einzelnen Karten lassen sich in mehrere Kategorien einordnen (siehe auch Abbildung 12.1):

- Inselförmige Strukturen: Ein Messpunkt oder eine geringe Anzahl von Messpunkten verwenden ein anderes Lexem als der Großteil der anderen Dialekte (kakvo - ”was”).
- Ost-West-Teilung: Die Trennung entlang der Yat-Linie ist auf vielen Karten klar erkennbar (kotka - ”Katze”).
- Die Rhodopen, das Mittelgebirge im Süden Bulgariens, verhält sich auch in den lexikalischen Daten heterogen. In einigen Karten ist es dem westlichen (bulka - ”Braut”), in anderen dem östlichen Teil Bulgariens

([yrvuli - ”Sandale”) zugeordnet. In wiederum anderen Karten bilden die Rhodopen ein eigenes, unabhängiges Gebiet (zapyr-vam - ”braten”).

12.2 Relativer Identitätswert

Der Methode *Relativer Identitätswert (RIW)* von Hans Goebl (siehe Kapitel 4.3.1) ist die einzige in dieser Arbeit beschriebene *aggregierende* Methode, die direkt auf lexikalische Dialektdaten angewandt werden kann. Der RIW zwischen zwei Messpunkten stellt den Quotienten aus der Menge der jeweils identischen Elementen zu der gesamten Menge der Elemente dar. In Bezug auf lexikalische Dialektdaten sind die einzelnen Wörter, lexikalische Realisierungen eines Konzepts, die zu untersuchenden Elemente. Der RIW zweier Messpunkte entspricht so dem Quotienten aus der Anzahl der gleichen Wörter zu der Gesamtzahl der erhobenen Wörter.

Abbildung 12.2 zeigt ein Clustering der RIW-Werte nach der Methode Ward mit 12 Clustern. Eine Struktur der Dialekte lässt sich nicht mehr erkennen: Die den einzelnen Clustern zugeordneten Messpunkte sind geographisch über ganz Bulgarien verteilt. Hier zeigt sich ein starker Gegensatz zu den phonetischen Daten: Letztere zeigen zumeist eine klare Einteilung in Ost- und Westbulgarien, die bei den lexikalischen Daten noch nicht einmal ansatzweise vorhanden ist. Die Aufsplitterung der lexikalischen Dialektdaten ist zu ausgeprägt und heterogen, als dass sich bei gemeinsamer Betrachtung aller Wörter Strukturen erkennen lassen. Zumindest im Fall der bulgarischen Daten lassen sich diese nicht mit aggregierenden Methoden analysieren.

12.3 Zusammenfassung

Abschließend lässt sich feststellen, dass zumindest im Fall des bulgarischen Datensatzes die lexikalischen Daten andere Strukturen als die phonetischen Daten aufweisen. Ein wirklicher Widerspruch zwischen den verschiedenen Arten von Dialektdaten lässt sich allerdings nicht ausmachen: Verlaufen phonetische und lexikalische Grenzen zwischen Dialekten nicht parallel, so widersprechen sie sich zumindest in den allermeisten Fällen nicht. Zumeist bil-

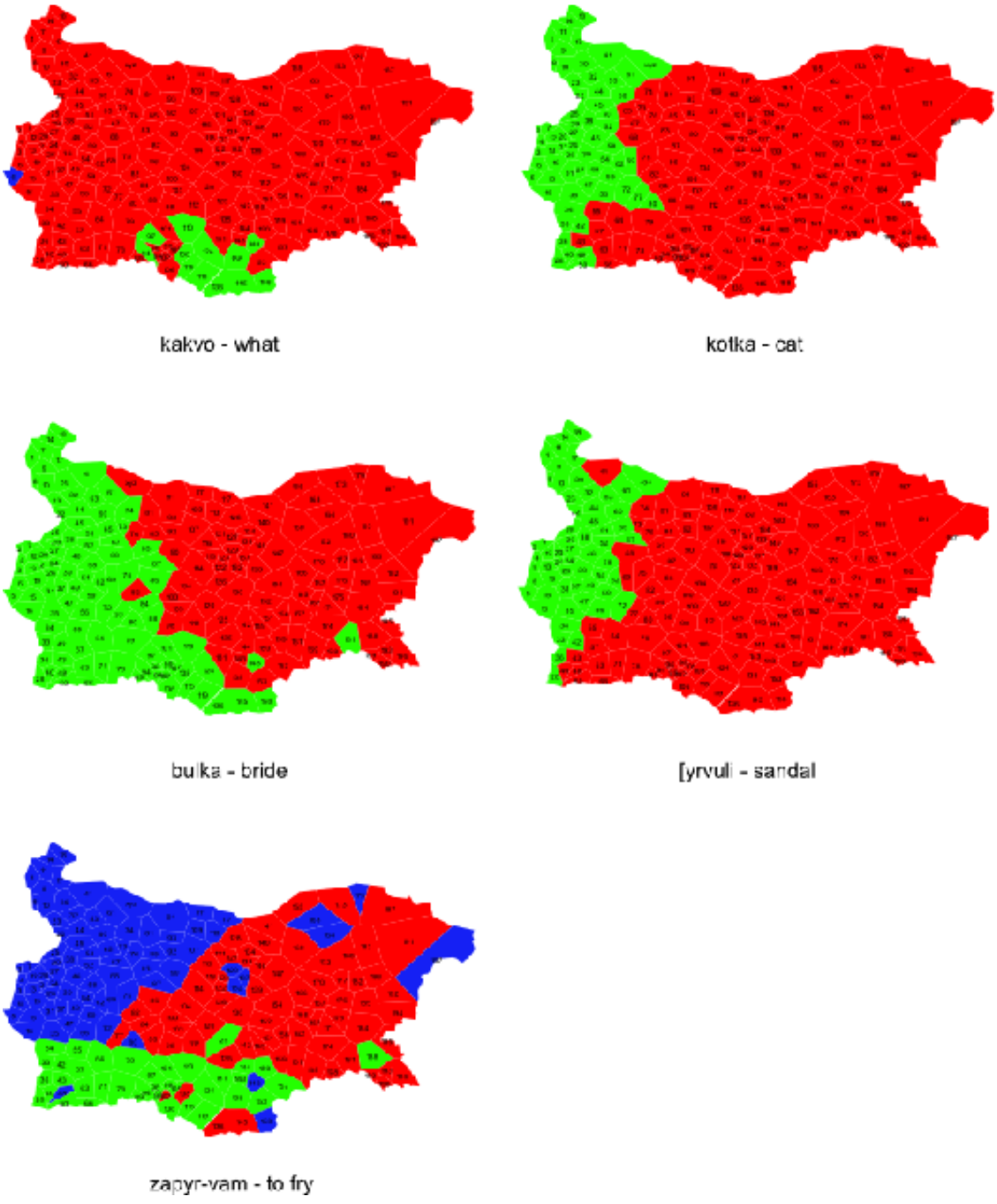


Abbildung 12.1: Einige Wörter des lexkalischen Datensatzes, phänomenologisch auf Voronoi-Karten dargestellt

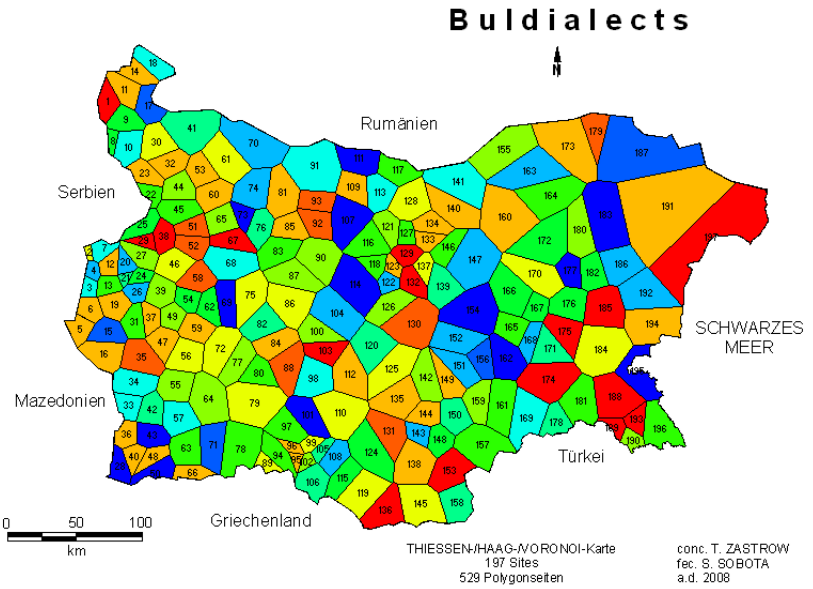


Abbildung 12.2: Der Relative Identitätswert, auf den lexikalischen Teil des bulgarischen Datensatzes angewandt, ergibt keine sichtbaren Strukturen (hier: Ward Clustering, 12 Cluster)

den die lexikalischen Daten zusammenhängende Gebiete aus, die entweder den Gebieten der phonetischen Dialekteinteilung entsprechen oder zumindest einen Teil davon abdecken. Die lexikalischen Daten weisen meist eine wesentlich gröbere Einteilung auf als die phonetischen Daten und decken häufig nur einen Teil des entsprechenden Gebiets ab.

Dies gilt allerdings nur bis zu einer gewissen Anzahl verschiedener Lexeme pro Konzept. Je höher hier die Anzahl der aufgetretenen Lexeme pro Konzept ist, desto verworrener wird die Einteilung der Dialektgebiete. Dies ist wahrscheinlich auch einer der Gründe, warum die aggregierende RIW-Methode kein zufriedenstellendes Ergebnis erzielt. Es stellt sich die Frage, ob die Verwendung der jeweiligen Lexeme tatsächlich nur auf die geographische Verteilung zurückzführen ist, oder ob auch noch weitere Kriterien wie besipielsweise das soziale Milieu Einfluss auf die Nutzung der Lexeme haben.

Wie eingangs erwähnt, sind lexikalische Differenzen zwischen Dialekten stärker zu bewerten als phonetische. Dies ist zurückzuführen auf die Tatsache, dass durch lexikalische Unterschiede das gegenseitige Verstehen stärker beeinträchtigt wird als durch lediglich phonetische Differenzen. Im Zuge dessen sind lexikalische Dialektvariationen innerhalb einer Sprache nicht so stark ausgeprägt *aggregierend* wie dies bei phonetischen Daten der Fall ist. Würden sich immer mehr lexikalische Differenzen zwischen Dialektgebieten aufsummieren, so würde sich ab einem bestimmten Zeitpunkt unweigerlich die Frage ergeben, ob die Unterscheidung noch auf dialektaler Ebene zu treffen ist oder ob es sich doch um separate Sprachen handelt.

Kapitel 13

Vergleich: Edit Distance und Relativer Identitätswert anhand des *Sprachatlas des Dolomitenladinischen und angrenzender Dialekte* (ALD-1)

In diesem Kapitel soll ein Datensatz des *Sprachatlas des Dolomitenladinischen und angrenzender Dialekte* (ALD-1) mit zwei Methoden der Dialektometrie analysiert und die Ergebnisse anschließend miteinander verglichen werden.

Der ALD-I ist eine Sammlung dialektaler Daten zu 217 österreichischen, schweizerischen und norditalienischen Messpunkte, zusammengestellt von Hans Goebl an der Universität Salzburg (siehe hierzu auch Kapitel 3.2.1). Abbildung 13.1 zeigt das abgedeckte Gebiet in Norditalien[1]. Der ALD-I ist ein wichtiger Sprachatlas der Romanistik. An seinen Daten wurden bereits viele dialektometrische Studien durchgeführt, vorzugsweise unter Anwendung

[1]Graphik eingesehen am 15.12.2010 auf der Webseite des ALD-I: http://www.sbg.ac.at/rom/people/proj/ald/ald_home.htm

des *Relativen Identitätswerts*, siehe hierzu beispielsweise Bauer (2003). In Form des Projekts *Sprechender Sprachatlas* stehen die Originalaufnahmen als Audiodateien verbunden mit entsprechender geographischer Software im Internet und auf CD-ROM zur Verfügung[2].

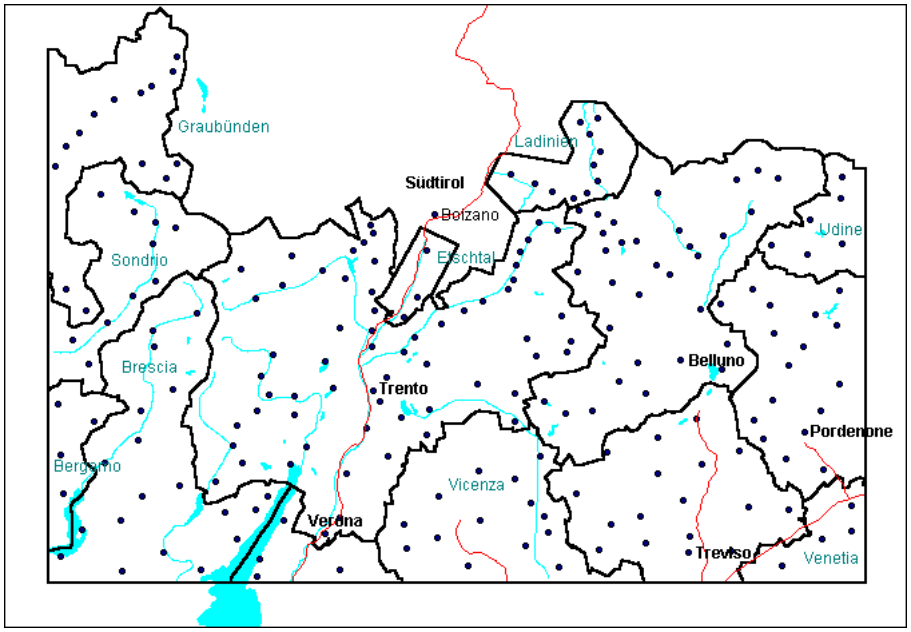


Abbildung 13.1: Das von ALD-I umfaßte Gebiet mit den punktuell eingezeichneten Messpunkten (Graphik von der Homepage des ALD-I)

Der ALD-I enthält sowohl lexikalische als auch phonetische Daten. Da eine *Edit Distance*-Analyse mit Hilfe des Levenshtein Algorithmus nur an phonetisch variierenden Daten möglich ist, wurden in Zusammenarbeit mit Hans Goebl insgesamt 147 rein phonetisch variierende Wörter ausgewählt. Über die 217 Messpunkte des ALD-I ergeben sich 31899 Varianten, somit ist der ALD-I-Datensatz größer als der Buldialects-Datensatz.

Der ALD-I enthält zwei Arten phonetischer Transkription: Das Datenfeld *Hilfstrans* beinhaltet eine einfache phonetische Transkription ohne weitergehenden Diakritika, im Datenfeld *Deftrans* hingegen findet sich eine fein granulierte, Diakritika verwendende Transkription (siehe Tabelle 13.1). Die

[2]http://www.sbg.ac.at/rom/people/proj/ald/sprech/einleitung.htm

Messpunkt	Feature	Hilfstrans	Deftrans
27	l'aceto	aze	l'az\3d\1
27	acido	acit	á\6c\1it

Tabelle 13.1: Ausschnitt aus den Dialekt Daten des ALD-I

folgende Untersuchung bezieht lediglich die einfache Hilfstrans-Transkription ein. Um auch die exaktere Deftrans-Transkription mittels der L04-Software analysieren zu können, wäre eine Umkodierung der Deftrans-Daten in IPA bzw. X-Sampa notwendig.

13.1 Vergleich: Relativer Identitätswert und Levenshtein Distance

Mit Hilfe der L04-Software von Peter Kleiweg wurde eine Levenshtein-basierte Analyse der ALD-I-Daten durchgeführt. Hierfür wurden die oben genannten 147 Wörter über die 217 Messpunkte in das L04-eigene Datenformat überführt (siehe Programm 2). Anschließend nimmt das Programm *leven* aus dem L04-Paket das Verzeichnis mit diesen Dateien als Eingabe-Parameter und berechnet mit Hilfe des Levenshtein-Algorithmus eine Distanzmatrix aus den gegebenen Daten[3]. Im Falle der ALD-I-Daten wurde keine weitere Parametrisierung des *leven*-Programms vorgenommen. Die entstandene Distanz-Matrix im L04-Format wurde anschließend in das VDM-Format umgewandelt.

Von Hans Goebl wurde eine mittels RIW errechnete Ähnlichkeitsmatrix der ALD-I-Daten zur Verfügung gestellt. Aufgrund der identischen Struktur der beiden mittels RIW bzw. Levenshtein Distance errechneten Matri-

[3]Der exakte Programmaufruf lautete:

leven -n 217 -l tbl.tbl -o fon.dif dir/.csv*

wobei die Datei tbl.tbl die Namen der Messpunkte enthält (hier: fortlaufende Nummern) und die erzeugte Distanz-Matrix in der Datei fon.dif zu finden ist

```
:1
-azai
:2
-azai
:3
-azai
```

Programm 2: Ein Ausschnitt aus den ins L04-Datenformat konvertierten ALD-I-Daten. Die Zeilen mit Doppelpunkt und Zahl beziehen sich auf den jeweiligen Messpunkt, darunter die phonetische Variante des entsprechenden Wortes, eingeleitet durch −. Für jedes Wort wird eine Datei benötigt, die die phonetischen Varianten in allen untersuchten Messpunkten enthält

zen lassen sich diese nun direkt in der VDM-Software miteinander vergleichen. Hierbei ist allerdings zu beachten, dass durch die unterschiedliche Natur der Algorithmen eine Matrix mit *Ähnlichkeits-* (RIW) bzw. *Distanz*-Werten (Levenshtein Distanz) entstanden ist. Die eine Matrix stellt also faktisch die Negation der anderen dar. Um eine direkte Vergleichbarkeit erreichen zu können, muss eine der beiden Matrizen dementsprechend in das Format der jeweils anderen überführt werden. Aus 30% *Ähnlichkeit* werden so 70% *Distanz* bzw. umgekehrt. Für diese Untersuchung wurde die Distanz-Matrix der Levenshtein Distanz in eine Ähnlichkeits-Matrix überführt.

Die Abbildungen 13.2 bzw. 13.3 zeigen Visualisierungen als Synopsenkarte der RIW- bzw. der Edit Distanz Matrix. Verwendet wurde der Intervallalgorithmus MinMwMax, jeweils mit 6 Klassen und dem Messpunkt Nummer 68 als Referenzpunkt. Auf beiden Karten lassen sich ähnliche Dialektareale ausmachen:

- Im Norden zeigen sich drei blau eingefärbte Bereiche.
- In der Mitte schiebt sich, ausgehend vom Referenz-Messpunkt, ein rot eingefärbter Keil von Norden Richtung Süden.
- Die grünen Bereiche im Osten und Westen weisen jeweils ähnliche Bereiche aus.

- Unterschiede zwischen den beiden miteinander verglichenen dialektometrischen Methoden zeigen sich in den gelben bzw. pinkfarbenen Arealen.

Die Abbildungen 13.4 und 13.5 zeigen ein Clustering nach der Ward-Methode mit jeweils 4 Clustern. Auch hier sind die beiden nördlichen Cluster nahezu identisch (rot und türkis), Abweichungen zeigen sich im südlichen Bereich. Die mittels RIW-Algorithmus erzeugte Matrix weist hier einen wesentlich größeren zusammenhängenden Bereich von Südosten nach Nordwesten (blau) auf als die Edit-Distanz Matrix.

Sowohl Relativer Identitätswert als auch Edit-Distanz Analyse zeigen im nördlichen Bereich des ALD-I weitgehend übereinstimmende dialektale Areale auf. Im Süden zeigen sich Unterschieden in der Einteilung der beiden vorwiegenden Dialektareale: Dies weist auf eine geringere Stabilität der südlichen Dialektgrenzen hin. Weitere Untersuchungen mit anderer Parametrisierung, vor allem aber die Einbeziehung der vollständigen, mit diakritischen Zeichen versehenen *Deftrans*-Daten des ALD-I könnten hier weitergehende Aussagen ermöglichen.

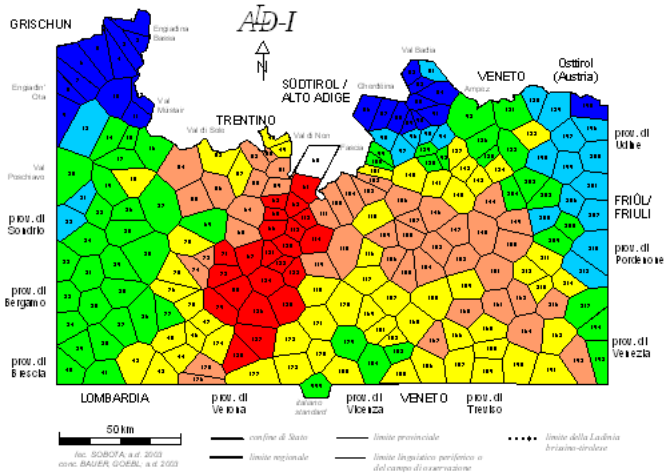


Abbildung 13.2: RIW, Klassifikation mit MinMwMax, 6 Klassen, Messpunkt 68 als Referenzpunkt

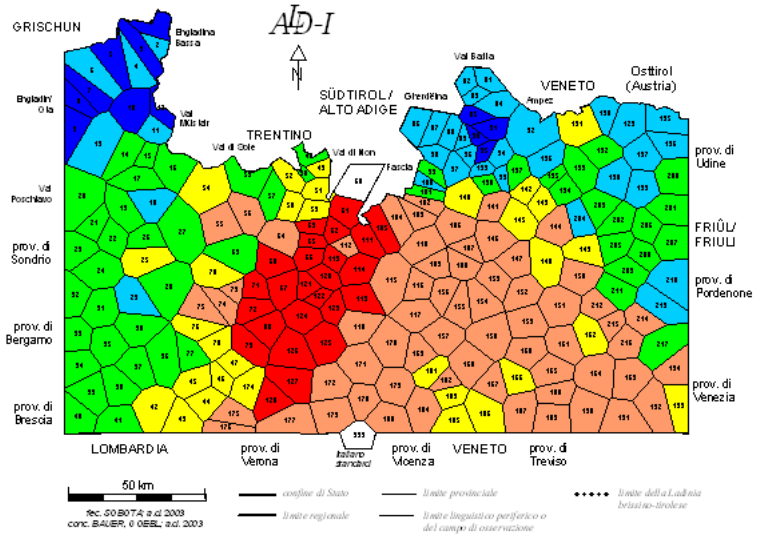


Abbildung 13.3: Levenshtein Distanz, Klassifikation mit MinMwMax, 6 Klassen, Messpunkt 68 als Referenzpunkt

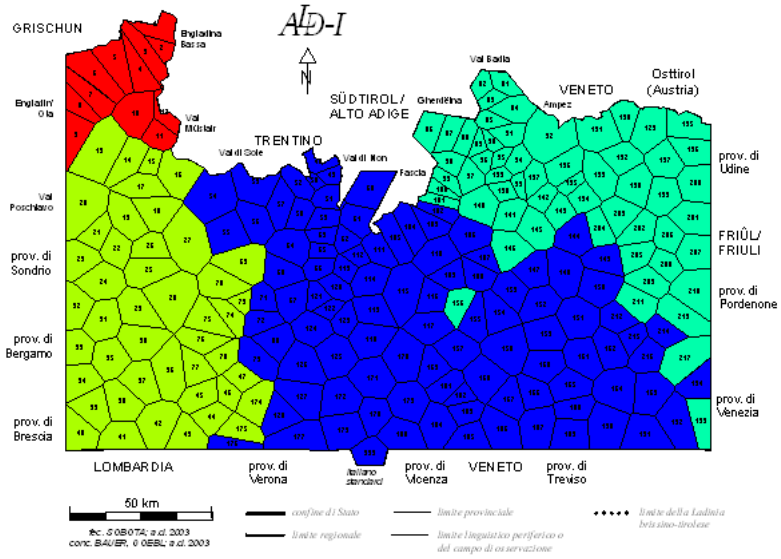


Abbildung 13.4: RIW, Clusteranalyse mittels WARD-Methode, 4 Cluster

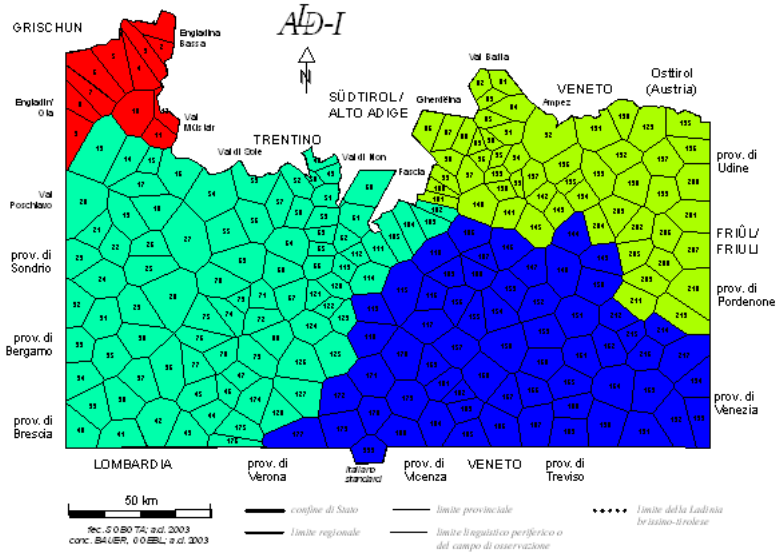


Abbildung 13.5: Levenshtein Distanz, Clusteranalyse mittels WARD-Methode, 4 Cluster

Kapitel 14

Schlußbetrachtungen

Methodik

In dieser Arbeit wurden drei neue Methoden zur Messung der phonetischen Abstände zwischen den Dialekten einer Sprache vorgestellt: Vektoranalyse, ein auf Bigrammen beruhender Ansatz und Verfahren aus dem Bereich der Informationstheorie. Während die beiden letztgenannten aggregierende Methoden darstellen, handelt es sich bei der Vektoranalyse um eine extrahierende Methode, die die Analyse einzelner Elemente eines Datensatzes erlaubt. Alle genannten Methoden produzieren als Ergebnis eine Distanz- oder Ähnlichkeitsmatrix, die anschließend weiterverarbeitet werden kann.

Werden die oben genannten Methoden im Hinblick auf die jeweilige Unterteilung der zu untersuchenden Dialektdaten betrachtet, so ergeben sich drei mögliche Einteilungen der Dialektdaten in kleinere Einheiten. Die dialektometrischen Methoden lassen sich dementsprechend ihrer Art der Dateneinteilung nach kategorisieren:

- Vektoranalyse: Bevor die Vektorketten durch die Dialektdaten eines Messpunktes gezogen werden können, müssen die einzelnen Wörter des Messpunktes in eine festgelegte Reihenfolge gebracht werden. Für die anschließend zu erstellende Vektorkette bedeutet eine Wortgrenze einen Sprung auf der Y-Achse, weswegen die Vektoranalyse eine *wortbasierte* Methode darstellt. In diese Kategorie gehören auch die Edit-Distanz-basierten Methoden.

- Die Entropie und die von ihr abgeleiteten Messwerte werden auf Basis der gesamten Daten eines Messpunktes errechnet, eine weitere Untergliederung der Daten erfolgt nicht mehr. Diese Methoden sind dementsprechend *messpunktbasiert*.

- Die Information eines Messpunktes wird auf Grundlage der Frequenzen aller Elemente im gesamten Datensatz berechnet. Somit handelt es sich um eine *korpusbasierte* Methode.

Die Bigrammanalyse weist Charakteristika aller drei Paradigmen auf: Sie ist *wortbasiert*, weil die Erstellung der Bigramme sich an den Wortgrenzen orientiert und nicht über die einzelnen Wörter hinausreicht. Anschließend wurden die Bigramme *innerhalb eines Messpunktes* einander gegenübergestellt (Graustufenmatrizen). In einem letzten Schritt schließlich wurden die Bigrammdichten bezogen auf das *gesamte Korpus* in eine Ähnlichkeitsmatrix umgerechnet. Es hat sich gezeigt, dass - bezogen auf den Datensatz des Buldialect-Projekts - die Bigrammanalyse mit die besten Ergebnisse in Bezug auf die Einteilung der Daten in Dialektareale geliefert hat. Hier wirkt sich die Integration der verschiedenen Einheitendimensionen positiv auf die nachfolgenden Analysen wie bspws. das Clustering und die Visualisierung aus.

Unabhängig von der zuvor angewendeten dialektometrischen Methode kommt diesen weitergehenden Analysemethoden der Distanzmatrizen besondere Bedeutung zu. Quantitative Analysen wie Intervallalgorithmen, Clustering, Multidimensional Scaling - sie alle interpretieren die gewonnenen Daten unterschiedlich und erlauben verschiedene Blickwinkel auf die Resultate der dialektometrischen Methoden. Dabei erlauben eine Vielzahl von frei zu setzender Parametern eine Feinjustierung der angewendeten Analysemethode. Aus praktischen Gründen ist es nicht möglich, alle möglichen Variationen von Parametereinstellungen durchzuspielen und zu publizieren. Stattdessen wurde in dieser Arbeit ein fester Satz Parameter (Cluster-Methode WARD, Anzahl der Cluster 12) auf alle Distanzmatrizen angewendet.

Es ist zu beachten, dass die fixen Clustereinteilungen, die das Ergebnis eines Clusteringprozesses darstellen, nicht die Stärke der Grenzen zwischen den einzelnen Clustern widerspiegeln. Aus dem Ergebnis eines Clusteringpro-

zesses lässt sich nicht ablesen, wie eng beieinander bzw. wie weit voneinander entfernt die einzelnen Cluster sind. Methoden wie das Noisy Clustering können hierfür angewendet werden: Sie zeigen auf, wie durchlässig die Grenzen zwischen den einzelnen Clustern sind bzw. wie stabil das Ergebnis des Clustering als Ganzes ist. Aufgrund seines stetigen Charakters fällt diese Problematik beim Multidimensional Scaling weg: Im Gegenzug sind hier die einzelnen Bereiche nicht klar voneinander abgegrenzt und eine automatisierte Einteilung der Datensätze in Gruppen findet nicht statt.

In einem letzten Schritt werden die Ergebnisse der Analysen der Distanzmatrizen interpretiert. Das wichtigste Mittel hierzu sind Visualisierungen: Dendrogramme, Scatterplots und Diagramme verdeutlichen die gewonnenen Zahlenreihen in geometrischer Art und Weise. In Kombination mit dem Einsatz verschiedener Farben können einzelne Aspekte der Visualisierung hervorgehoben werden. Die Darstellung der analysierten Daten auf Voronoi-Karten stellt die wichtigste Visualisierungsmethode im Bereich der Dialektforschung dar. Die Einteilung der Stillen Karte in Polygone repräsentiert die geographische Verteilung der Messpunkte - deren Raumaufteilung ist ein fixer Parameter, der sich nicht ändert. Die Extrusion der Polygone in die dritte Dimension ermöglicht die Einbeziehung eines zusätzlichen Parameters und somit einen weiteren Blickwinkel auf die Daten in derselben Visualisierung. Allgemein lassen sich Visualisierungen durch den Einbezug der dritten Dimension um die Darstellung einer weiteren Datenreihe anreichern, dies kann auch in Form von dreidimensionalen Scatterplots geschehen. Ein Nachteil der dreidimensionalen Darstellungen besteht darin, dass sie sich nur schlecht in herkömmlichen (papiergebundenen) Medien publizieren lassen. Eine zweidimensionale Wiedergabe dreidimensonaler Strukturen kann immer nur einen Schnappschuß aus einem bestimmten Winkel darstellen. Hier sind neue Datenformate und Medien gefragt, die auch im zweidimensionalen Raum einen Blick auf die dritte Dimension erlauben.

Anwendung der Methoden auf den Buldialects Datensatz

Die oben genannten dialektometrischen Methoden, Analysen und Visualisierungen wurden auf den phonetischen Teil des Buldialect Datensatz angewendet. Was die Grobstrukturen der bulgarischen Dialektgebiete anbelangt,

ergeben sich weitgehende Übereinstimmungen zwischen den einzelnen Methoden. Der Verlauf der Yat-Linie und die Abgrenzung der Rhodopen stechen heraus, sie sind nahezu unabhängig von verwendeter dialektometrischer Methode, Analyse oder Visualisierung sichtbar. Weitere Einteilungen der Dialektareale sind nicht in allen Fällen erkennbar: Diese Gebiete weisen höhere Ähnlichkeiten zu den sie umgebenden Dialektarealen auf und werden methodenabhängig diesen zugeschlagen. Hierbei handelt es sich um die Übergangsdialekte zu Serbien sowie eine Enklave westlicher Dialekte am südöstlichen Rand in der Nähe der türkischen Grenze. In diesen Fällen sind extrahierende Methoden im Vorteil, das sie das Augenmerk auf ein bestimmtes Element richten und die anderen Daten ausblenden.

Die Ergebnisse wurden anschließend verglichen mit den Dialektarealen, wie sie von der traditionellen bulgarischen Dialektologie angegeben wurden. Auch hier ergibt sich weitgehende Übereinstimmung. Allerdings erscheint in den dialektometrischen Analysen die Yat-Linie ein Stück weiter nach Osten verschoben. Des Weiteren ist die Enklave der westlichen Dialekte im Südosten nicht vorhanden auf Stojkovs Karte.

Der lexikalische Teil des Buldialect-Datensatzes lässt sich nicht aggregierend analysieren. Bei der phänomenologischen Darstellung einzelner Arbeitskarten ergeben sich zwar zusammenhängende Dialektareale, die zumindest der dialektalen Einteilung Bulgariens nach phonetischen Gesichtspunkten nicht zuwider läuft. Eine Aggregation dieser einzelnen Arbeitskarten ist allerdings nicht mehr möglich. Eine Anwendung des *Relativen Identitätswerts* auf die gesamten lexikalischen Daten hat keine zusammenhängende Dialektareale erkennen lassen.

Zusammenfassend lässt sich sagen, dass sich die hier vorgestellten dialektometrischen Methoden zur Analyse des Buldialect-Datensatzes bewährt haben. In weiteren Analysen gilt es herauszufinden, ob auch eine Anwendung der Methoden auf Dialektdaten anderer Sprachen erfolgreich sein wird.

Vergleich Levenshtein und RIW

Der Vergleich des Edit Distance-Ansatzes mit der RIW-Analyse anhand der italienischen Dialektdaten lässt eine weitgehende Übereinstimmung der

Ergebnisse erkennen. In zukünftigen Arbeiten wäre eine Ausweitung der analysierten Daten auf die Diakritika beinhaltende Datenreihe des italienischen Datensatzes sowie die Einbeziehung weiterer dialektometrischer Methoden in den Vergleich sinnvoll.

Anhang A

Anhang - Tabellen

	Bulg., Kyrillisch	Deutsch	Wortart
1	агне	Lamm	S
2	аз	ich	PRO
3	бели	weiss pl.	A
4	берат	Aufnehmen, 3. pl.	V
5	беше	War, 3. Person	V
6	брашно	Mehl	S
7	бързо	schnell	A
8	бяхме	sind 1. pl.	V
9	вежда	Augenbraue	S
10	вече	bereits	AV
11	вечер	Abend	S
12	вие	ihr pl.	PRO
13	вино	Wein	S
14	вода	Wasser	S
15	вол	Ochse	S
16	време	Zeit	S
17	връх	Gipfel	S
18	във	in	PRE
19	вълк	Wolf	S
20	вълна	Wolle	S
21	вътре	Das Innere	S
22	вятър	Wind	S
23	глава	Kopf	S
24	гладен	hungrig	A
25	говедо	Rind, Rindfleisch	S
26	горе	auf	PRE
27	гости	Gäste	S
28	градът	Die Stadt	S
29	дадоха	gaben, 3. pl.	V
30	две	zwei	NUM
31	двор	Hof	S
32	ден	Tag	S
33	дера	gerbe, 1. sg.	V
34	десет	zehn	NUM
35	дете	Kind	S
36	джоб	Tasche	S
37	днес	heute	AV
38	долу	abwärts	AV
39	дошъл	ist gekommen, 3. sg.	V
40	дъжд	Regen	S
41	дълбок	tief	A
42	дъно	Boden	S
43	дърво	Baum	S
44	език	Zunge	S
45	желязо	Eisen	S
46	жена	Frau	S
47	жив	lebten, 3. pl.	V
48	жълт	gelb	A
49	жътва	Ernte	S
50	звезда	Stern	S
51	земя	Erde	S
52	зет	Schwiegersohn/Schwager	S
53	й	sie, dativ, Kurzform	PRO
54	им	ihnen, dativ, Kurzform	PRO
55	име	Name	S
56	камък	Stein	S
57	ключ	Schlüssel	S
58	кон	Pferd	S
59	леща	Linse	S
60	майка	Mutter	S

	Bulg., Kyrillisch	Deutsch	Wortart
61	месец	Monat	S
62	месо	Fleisch	S
63	много	viel	A
64	мъж	Mann	S
65	мъжът	der Mann	S
66	неделя	Sonntag	S
67	нещe	will nicht, 3. sg.	V
68	нея	sie	PRO
69	ние	wir	PRO
70	носят	tragen, 3. pl.	V
71	нощ	Nacht	S
72	няма	da ist nicht, will nicht	AV
73	овца	Schaf	S
74	овце	Schafe	S
75	овчар	Hirte	S
76	овчари	Hirten	S
77	огън	Feuer	S
78	орех	Walnuss	S
79	пепел	Asche	S
80	петел	Hahn	S
81	понеделник	Montag	S
82	пръч	Ziegenbock	S
83	първият	Erster – er	S
84	път	Strasse	S
85	пясък	Sand	S
86	река	Fluss	S
87	ръка	Hand	S
88	ръце	Hände	S
89	се	jemandes	PRO
90	сега	jetzt	AV
91	седя	sitze, 1. sg.	V
92	сестра	Schwester	S
93	сирене	Käse	S
94	сол	Salz	S
95	старец	alter Mann	S
96	страх	Angst	S
97	сух	trocken	A
98	събота	Samstag	S
99	сърп	Sichel	S
100	със	mit	PRE
101	такъв	derart	A
102	твой	dein	PRO
103	това	dieses	ART
104	тогава	dann	PRE
105	тънко	dünn	A
106	ухо	Ohr	S
107	хляб	Brot	S
108	хоро	Kettentanz	S
109	хубаво	hübsch	A
110	цял	ganz	PAR
111	червен	rot, mask.	A
112	черен	schwarz, mask.	A
113	чешма	Brunnen	S
114	човек	Person	S
115	ще	sollen	V
116	я	sie akkusativ, Kurzform	PRO
117	ябълка	Apfel	S
118	яйца	Eier	S
119	яйце	Ei	S

Abbildung A.1: Die in den Analysen verwendeten 119 Wörter in kyrillischer Schreibweise mit deutscher Übersetzung und Angabe der Wortart

Xsampa-Code	IPA	Anzahl
"	ˈ	22049
e	e	10671
A	ɑ	7016
o	o	5424
r	r	5362
_j	ʲ	5294
i	i	4842
n	n	4774
7	ɤ	4720
v	v	4716
d	d	4550
t	t	4401
s	s	4236
l	l	4028
@	ə	3854
u	u	3804
k	k	3122
m	m	2813
j	j	2158
p	p	2142
g	g	2053
b	b	1994
S	ʃ	1694
t_s	t͡s	1650
t_S	t͡ʃ	1565
Z	ʒ	1278
f	f	1170
z	z	1075

Xsampa-Code	IPA	Anzahl
E	ɛ	593
c	c	468
x	x	437
r_=	ɻ	350
w	w	305
t_s\		262
U	ʊ	260
d_Z		256
s\	ɕ	248
O	ɔ	243
h	h	230
z\	ʑ	203
n`	ɳ	196
J\	ɟ	181
d_z		158
a	a	143
l_=		140
:	ː	123
l	l	68
p\	ɸ	42
C	ç	35
1	ɪ	28
V	ʌ	25
y	y	21
d_z\		19
Q	ɒ	19
M\	ɯ̹	12
?		0

Abbildung A.2: Die im bulgarischen Datensatz verwendeten XSampa Codes, sortiert nach Frequenz. Bei Einträgen mit Liaisonbogen (_) handelt es sich um zwei Laute, die zu einem zusammengefaßt werden

Anhang B

Anhang - Datenpublikation

Abbildung B.1: Die eSciDoc-Solution der Buldialects-Daten

Im Rahmen des Projektes BW-eSci(T) an der Universität Tübingen[1] wurde der phonetische Teil des Buldialects Datensatz in die eScience-Plattform eSciDoc (https://www.escidoc.org/) eingespielt[2]. Hierfur wurde

[1] http://www.bwescit.uni-tuebingen.de/

[2] Arbeiten von Niko Schenk

eine eSciDoc-Solution[3] als graphische Benutzerschnittstelle entworfen. Diese ermöglicht direkten Zugriff auf die Primärdaten sowie erste statistische Analysen und die zugehörigen Visualisierungen der Ergebnisse (Abbildung B.1).

[3]Nähere Informationen zur eSciDoc-Software finden sich auf https://www.escidoc.org/

Literaturverzeichnis

[Alewijnse u. a. 2007] Alewijnse, Bart ; Nerbonne, John ; Veen, Lolke van d. ; Manni, Franz: A Computational Analysis of Gabon Varieties. In: *Proceedings of the International Workshop Computational Phonology (RANLP 2007)*, 2007, S. 3–12

[Alexander 2002] Alexander, Ronelle: The Scope of Double Accent in Bulgarian Dialects. In: *GAIA Books* (2002)

[Altmann u. Ziegenhain 2010] Altmann, Hans ; Ziegenhain, Ute: *Prüfungswissen Phonetik, Phonologie und Graphemik*. Vandenhoeck und Ruprecht, UTB, 2010

[Appel 1968] Appel, Arthur: Some techniques for shading machine renderings of solids. In: *Proceedings of the April 30–May 2, 1968, spring joint computer conference* (1968)

[Baehr u. Kabelac 2009] Baehr, Hans D. ; Kabelac, Stephan: *Thermodynamik: Grundlagen und technische Anwendungen*. Springer Verlag, 2009

[Bamberg u. Baur 2008] Bamberg, Günter ; Baur, Franz: *Statistik*. Oldenbourg Wissenschaftsverlag GmbH, 2008

[Barbiers u. a. 2005] Barbiers, Sief ; Bennis, Hans ; Vogelaer, Gunther de: *Syntactic Atlas of the Dutch Dialects*. Amsterdam University, 2005 `http://www.meertens.knaw.nl/projecten/sand/sandeng.html`

[Bauer 2003] Bauer, Roland: *Dialektometrische Analyse des Sprachatlasses des Dolomitenladinischen und angrenzender Dialekte (ALD-I)*. Salzburg (Institut für Romanistik), 2003

[Brillouin 1962] Brillouin, Leon: *Science and Information Theory.* Academic Press, New York, 1962

[Bußmann 2002] Bussmann, Hadumod: *Lexikon der Sprachwissenschaft.* Körner, 2002

[Chambers u. Trudgill 1980] Chambers, J. K. ; Trudgill, Peter: *Dialectology.* Cambridge University Press, 1980

[Cover u. Thomas 2006] Cover, Thomas M. ; Thomas, Joy A.: *Elements of Information Theory.* Wiley-Interscience, 2006

[Covington 1996] Covington, Michael A.: An algorithm to align words for historical comparison. In: *Computational Linguistics* (1996)

[Deichsel u. Trampisch 1985] Deichsel, G. ; Trampisch, H.J.: *Clusteranalyse und Diskriminanzanalyse.* Spektrum Akademischer Verlag, 1985

[Dulicenko 2002] Dulicenko, Aleksandr: Banater Bulgarisch. In: *Wieser Enzyklopädie des europäischen Ostens 10* (2002), 203-208. `http://www.uni-klu.ac.at/eeo/BanaterBulgarisch.pdf`

[Frahm 2003] Frahm, Eckart et. a.: *Renaissance des Dialekts?* Tübinger Vereinigung für Volkskunde, 2003

[Glaser u. Bucheli Berger 2000] Glaser, Elvira ; Bucheli Berger, Claudia: *Dialektsyntax des Schweizerdeutschen. Syntaktischer Atlas der Deutschen Schweiz.* Universität Zürich, Deutsches Seminar, 2000-2010 `http://www.ds.uzh.ch/dialektsyntax/pro_beschrieb.html`

[Goebl 1982] Goebl, Hans: *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF, Band 1.* Max Niemeyer Verlag, Tübingen, 1982

[Goebl 1998] Goebl, Hans (Hrsg.): *ALD-I Sprachatlas des Dolomitenladinischen und angrenzender Dialekte.* Dr. Ludwig Reichert Verlag, Wiesbaden, 1998

[Goebl 2004] Goebl, Hans: Sprache, Sprecher und Raum: Eine kurze Darstellung der Dialektometrie. Das Fallbeispiel Frankreich. In: *Mitteilungen der österreichischen Geographischen Gesellschaft* 146 (2004), S. 247–286

[Goebl 2006] Goebl, Hans: Recent Advances in Salzburg Dialectometry. In: *Literary and Linguistic Computing* 21, No. 4 (2006), S. 411–435

[Goebl 2007a] Goebl, Hans: Dialektometrische Streifzüge durch das Netz des Sprachatlasses AIS. In: *Ladinia XXXI* (2007), S. 187–272

[Goebl 2007b] Goebl, Hans: Kurzvorstellung der Korrelativen Dialektometrie. In: *Quantitative Linguistics* 62 (2007), S. 165–180

[Gooskens 2007] Gooskens, Charlotte: The Contribution of Linguistic Factors to the Intelligibility of Closely Related Languages. In: *Journal of multilingual and multicultural development* 28 (2007), 445-467. `http://www.let.rug.nl/ gooskens/pdf/JMMD.pdf`

[Gordon 2005] Gordon, Raymond C. (Hrsg.): *Ethnologue: Languages of the World.* 15. SIL, Dallas, 2005

[Grey Thomason 1999] Grey Thomason, Sarah: Linguistic Areas and Language History. In: *Journal of Language Contact* (1999)

[Gutschmidt 2002] Gutschmidt, Karl: Bulgarisch. In: *Wieser Enzyklopädie des europäischen Ostens 10* (2002), 219-234. `https://claroline.uni-klu.ac.at/eeo/Bulgarisch.pdf`

[Haas 1990] Haas, Walter: *Jacob Grimm und die deutschen Mundarten.* Steiner Franz Verlag, 1990

[Hartley 1927] Hartley, R. V. L.: Transmission of Information. In: *International Congress of Telegraphy and Telephony, Lake Como, Italy* (1927), S. 535 ff.

[Heeringa 2004] Heeringa, Wilbert: *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen, University of Groningen, Diss., 2004

[Hinrichs u. a. 2005] Hinrichs, Erhard ; Gerdemann, Dale ; Nerbonne, John: *Measuring linguistic unity and diversity in Europe.* 2005. – Project Buldialects proposal

[IPA 1999] IPA, diverse: *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* 8th. Department of Linguistics, University of Victoria : Cambridge University Press, 1999

[Kessler 1995] Kessler, Brett: Computational Dialectology in Irish Gaelic. In: *Proceedings of the European ACL* (1995), S. 60–67

[Klimant u. a. 2006] Klimant, Herbert ; Piotraschke, Rudi ; Schönfeld, Dagmar: *Informations- und Kodierungstheorie.* Teubner Verlag, Wiesbaden, 2006

[Kondrak 2000] Kondrak, Grzegorz: A New Algorithm for the Alignment of Phonetic Sequences. In: *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000).* Seattle : Janyce Wiebe, April 2000, S. 288–295

[König 2005] König, Werner: *dtv-Atlas Deutsche Sprache.* Deutscher Taschenbuch Verlag, München, 2005

[König u. Renn 2007] König, Werner ; Renn, Manfred: *Kleiner Sprachatlas von Bayerisch-Schwaben.* Wissner Verlag, Augsburg, 2007

[Küpfmüller 1954] Küpfmüller, Karl: Die Entropie der deutschen Sprache. In: *Fernmeldetechnische Zeitschrift Nr. 7* (1954), S. 265–272

[Levenshtein 1965] Levenshtein, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. In: *Doklady Akademii Nauk SSSR. 163, Nr. 4, Englische Übersetzung in: Soviet Physics Doklady, 10(8) S. 707-710* (1965)

[Li 2005] Li, Haiyan: *Data Visualization of Asymmetric Data Using Sammon Mapping and Applications of Self-Organizing Maps*, University of Maryland, Diss., 2005

[Lyre 2002] Lyre, Holger: *Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung.* UTB, 2002

[Löffler 2003] Löffler, Heinrich: *Dialektologie. Eine Einführung.* Gunter Narr Verlag Tübingen, 2003

[Nerbonne 2005] Nerbonne, John: Various Variation Aggregates in the LAMSAS South. In: *Accepted to appear in Catherine Davis and Michael Picone (eds.) Language Variety in the South III. Tuscaloosa: University of Alabama Press* (2005)

[Nerbonne u. a. 1999] Nerbonne, John ; Heeringa, Wilbert ; Kleiweg, Peter: Edit Distance and Dialect Proximity. In: *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* CSLI Press, 1999, S. v–xv.

[Nerbonne u. a. 2008] Nerbonne, John ; Kleiweg, Peter ; Heeringa, Wilbert ; Manni, Franz: Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: *Proc. of the 31st Annual Meeting of the German Classification Society* (2008)

[Osenova u. Simov 2005] Osenova, Petya ; Simov, Kiril: An Infrastructure for Storing and Processing Dialect Data. In: *Bulgarian islands on the linguistic map of Balkans* (2005)

[Prokic u. a. 2009] Prokic, Jelena ; Nerbonne, John ; Zhobov, Vladimir ; Osenova, Petya ; Simov, Kiril ; Zastrow, Thomas ; Hinrichs, Erhard: The Computational Analysis of Bulgarian Dialect Pronunciation. In: *Serdica Journal of Computing 3* (2009), S. 269–298

[Ruoff 1984] Ruoff, Arno: *Idiomatica, Band 10 und 11: Alltagstexte 1 und 2.* Max Niemeyer Verlag, Tübingen, 1984

[Schonlau 2002] Schonlau, Matthias: The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses. In: *The Stata Journal 3* (2002), S. 316–327

[Schwartz u. a. 1960] Schwartz, Manuel ; Green, Simon ; Rutledge, W.A.: *Vector Analysis*. Harper and Brothers, 1960

[Seguy 1971] Seguy, J.: La relation entre la distance spatiale et la distance lexicale. In: *Revue de Linguistique Romane 35* (1971), S. 335–357

[Seguy 1973] Seguy, J.: La dialectometrie dans l'atlas linguistique de Gascogne. In: *Revue de Linguistique Romane 37* (1973), S. 1–24

[Shannon 1947] Shannon, Claude E.: A Mathematical Theory of Communication. In: *Bell System Technical Journal, Vol. 27* (1947), S. 379–423

[Shannon 1951] Shannon, Claude E.: Prediction and Entropy of Printed English. In: *Bell Systems Technical Journal, Vol. 30* (1951), S. 50–64

[Sobolev 1998] Sobolev, A.N.: *Sprachatlas Ostserbiens und Westbulgariens: Texte*. Biblion, 1998 (Sprachatlas Ostserbiens und Westbulgariens). `http://books.google.de/books?id=suEoAQAAIAAJ`

[Spruit 2006] Spruit, Marco R.: Measuring Syntactic Variation in Dutch Dialects. In: *Literary and Linguistic Computing* 21 No. 4 (2006)

[Stojkov 1962] Stojkov, Stojko: *Bulgarian Dialectology (in cyrilic language)*. Sofia, 1962

[Stojkov 1964] Stojkov, Stojko: *Bulgarski dialecten atlas*. Izdatelstvo na Bulgarskata Akademija na Naukite, 1964-81

[Veith u. a. 1984] Veith, Werner ; Putschke, Wolfgang ; Hummel, Lutz: *Kleiner Deutscher Sprachatlas*. Max Niemeyer Verlag, Tübingen, 1984 - 1999

[Schulte im Walde 2003] Walde, Sabine Schulte i.: *Experiments on the Automatic Induction of German Semantic Verb Classes*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Diss., 2003

[Weigand 1909] Weigand, G.L.: *Linguistischer Atlas des dacorumänischen Sprachgebietes*. J.A. Barth, 1909 `http://books.google.de/books?id=xNDYYgEACAAJ`

[Weiss 2003] Weiss, Helmut: Vom Nutzen der Dialektsyntax. In: *Morphologie und Syntax deutscher Dialekte und Historische Dialektologie des Deutschen*, 2003, S. 21–41

[Wells 1995] Wells, J.C.: *Computer-coding the IPA: a proposed extension of SAMPA*. `http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf`. Version: 1995. – Web Document

[Wintgens 1982] Wintgens, Leo: *Grundlagen der Sprachgeschichte im Bereich des Herzogtums Limburg. Beitrag zum Studium der Sprachlandschaft zwischen Maas und Rhein*. Grenz-Echo-Verlag, 1982

[Zhobov 2006] Zhobov, Vladimir: *Description of the Sources for the Pronunciation Data*. 2006. – Unpublished manuscript. Department of Slavic Philologies, University of Sofia

[Zhobov u. a. 2004] Zhobov, Vladimir ; Alexander, Ronelle ; Kolev, Georgi: Hierarchies of Stress Assignment in Bulgarian Dialects. In: *GAIA Books* (2004)