

# Geostatistische Analyse der Sprachvariation in deutschen Regionen: Eine Ausarbeitung

Xiwen Feng, Chong Shen

*Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, 70569, Stuttgart*

## Abstract

In dieser Ausarbeitung beschäftigen wir uns mit der geostatistischen Analyse der Sprachvariation in deutschen Sprachregionen. Zunächst geben wir eine Vorstellung für den Hintergrund der Geolinguistik, geostatistische Methoden und aktuelle Forschungsstand der Geolinguistik. Danach folgen ein paar wichtigen Studien, Experimente und ihre Ergebnisse mit entsprechenden Analysen. Zum Schluss diskutieren wir über die Entwicklung der Varianten in China am Beispiel des Dialektes in der Metropole Shanghai.

Diese Ausarbeitung basiert hauptsächlich auf dem Papier von Pickel "Ergebnisse geostatistischer Analysen arealsprachlicher Variation im Deutschen".

## Keywords

Geolinguistik, Dialektologie, Dialektometrie, Varietäten des Deutschen

## 1. Hintergrund

In den 1870en Jahren entdeckte die historisch-vergleichende Sprachwissenschaftler die Regelmäßigkeit der phonetischen Entwicklung. 1876 erhoben August Leskien, ein Angehörige der Junggrammatiker, den Slogan "keine Ausnahmen von den Gesetzen der Phonetik". Mit diesem Slogan deutete er auf die Ausnahmslosigkeit der Lautgesetze, die besagt, dass die Richtung der Lautbewegung in dem Lautwandel bei allen Sprechern einer Einzelsprache oder aller genetisch verwandten Sprachen stets dieselbe ist, sofern keine Dialektspaltung eintritt. Eine weitere Annahme ist, dass die Ausnahmslosigkeit der Lautgesetze mit Hilfe der Dialektstudie bewiesen werden könnte.

1876 versuchte der deutsche Sprachwissenschaftler Georg Wenker, die Ausnahmslosigkeit der Lautgesetze zu beweisen. Dazu schickte er einen Fragebogen mit 40 Sätzen von ca. 300 Wörtern an alle Grundschullehrer im Rheinland und bat sie, diese in lokale Dialekte zu übertragen. Die Übersetzung wies eine klare Abgrenzung zwischen Hoch- und Niederdeutsch im Rheingebiet auf, welche mit seiner Annahme übereinstimmt. Durch Kartierung der Ergebnisse in sechs Dialektkarten erhielt er den Sprachatlas des Deutschen Reichs. Zu seiner Überraschung fand er

---

*Modul: Varietäten des Deutschen, Nummer: 182214100, Prüferin: PD. Dr. Eleonore Brandner, Art: Seminar, WS2021/2022, Leistungspunkte: 6 LP*

 [st152848@stud.uni-stuttgart.de](mailto:st152848@stud.uni-stuttgart.de) (X. Feng); [st143575@stud.uni-stuttgart.de](mailto:st143575@stud.uni-stuttgart.de) (C. Shen)

 3219649 (X. Feng); 3111514 (C. Shen)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

aus dem Sprachatlas, dass sich die Benrather Linien, die die phonologischen Veränderungen in verschiedenen Wörtern widerspiegeln, oft nicht überschneiden. Mit der Tenuisverschiebung [k] → [x] als Beispiel zeigte er, dass die Laute [k] in "maken" und in "ik" nicht übereinstimmen. Daraus stellte er fest, dass die geografische Verteilung einer Lautverschiebung von Wort zu Wort variiert. Dieses Phänomen widerlegt die Theorie der Junggrammatiker, die besagt, dass eine phonologische Veränderung alle Wörter in gleicher Weise betrifft.

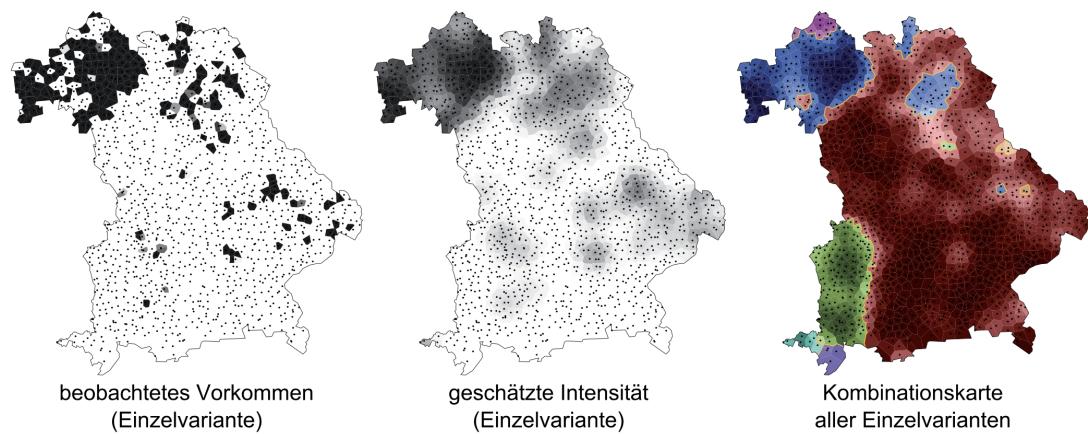
Hugo Schuchardt stimmte sich in der Blütezeit der Junggrammatiker stark gegen die oben genannten Theorie. Jules Gilliéron wurde von seinen Ideen beeinflusst und führte eine Studie mit Edmond Edmont über französische Dialekte durch. Sie erstellten einen Fragebogen mit ca. 2.000 Wörtern und führte eine Feldstudie durch, um die Dialekte in etwa 650 Orten zu untersuchen und zu kartieren. Die Ergebnisse wurden in dem Atlas der französischen Sprache (Atlas Linguistique de la France, 1902-1914 und 1920) veröffentlicht. Daraus entdeckte Gilliéron, dass fast jedes Wort seine eigene, einzigartige Benrather Linie hat. Daher erhob Gilliéron einen neuen Slogan "Jedes Wort hat seine eigene Geschichte" (Lyle Campbell, 1999), um sich gegen den Slogan der Junggrammatiker zu stemmen. Seitdem sind in der Entwicklung der Geolinguistik zahlreiche Sprachatlanten in Europa und Amerika veröffentlicht worden.

## 2. Einführung in die Geostatistik

Die Dialektgeographie hat folgenden Hauptaufgaben: (1). die spezifische Bestandteile oder Strukturen bestimmter Dialekte zu erfassen; (2). eine lokale oder umfassende Beschreibung eines einzelnen Dialekt zu geben; (3). sprachliche Unterschiede in einem Gebiet zu untersuchen; (4). Dialektgebiete abzugrenzen; und (5). die Ursachen der Unterschiede und den Entstehungsprozess der Dialektgebiete zu erklären. Da die räumliche Nähe entscheidend für die Ausbreitung von Sprachvarianten und damit auch für die Ausbildung von Varietäten ist, wurde Geostatistik, die sich zwischen den Methoden im engeren Sinne und den im weiteren Sinne unterscheidet, in Linguistik eingeführt.

Geostatistik im engeren Sinne umfasst die Methoden, mit denen die geographische Anordnung der Belegdaten als Größe direkt in Berechnungen und in Analysen eingeht. Durch Einbezug der räumlichen Distanzen und dialektalen Unterschieden in ein Koordinatensystem erhält Seguy(1971) ein Bild, das sich ein logarithmisches Verhältnis anzeigt und "Seguy-Kurve" genannt wird. Nerbonne(2009) berechnete die Korrelation zwischen der logarithmischen geographischen Distanz und dem akkumulierten phonetischen Abstand aus den Datenpunkten des Phonetischen Atlas Deutschlands. Er nahm den Quadratwert der Korrelationskoeffizient als Maß für den Anteil an rein geographisch konditionierter Sprachvariation. Neben räumlichen Distanzen ist die räumliche Autokorrelation eine weitere Messgröße für den Einfluss der Geographie auf die sprachliche Variation. Sie beschreibt, wie stark ähnliche Werte einer Variablen im Raum geballt sind. Dazu sind globale Autokorrelationsmaße(z.B. Moran's I) und lokale Autokorrelationsmaße(z.B. Getis-Ord Gi\*) zu unterscheiden. Die ersten geben die Gesamtvariation auf einer Karte an, während die letzten für jeden Ort einen Wert der lokalen

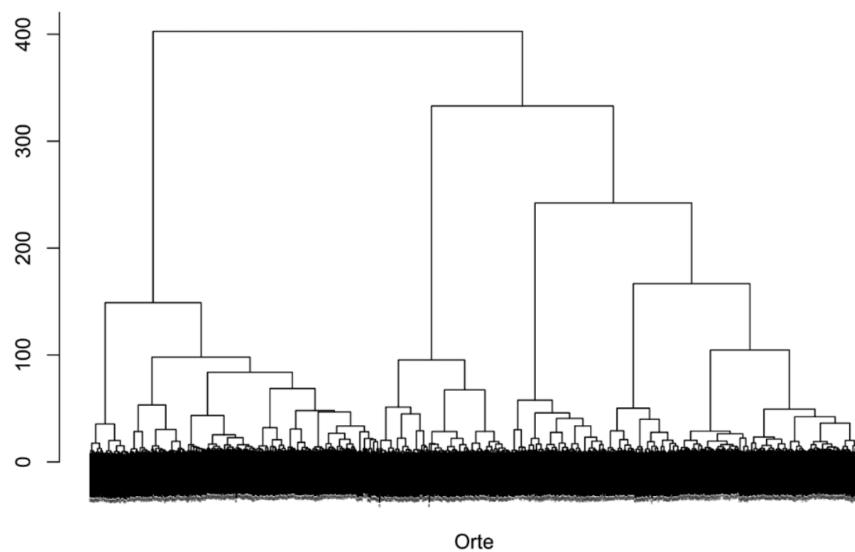
Kohäsion liefern und Hotspots sowie Coldspots in dem Ergebnis angeben. Für Datensätze mit spärlichen Daten, die z.B. Lücken aufweisen oder Granularität nicht feinkörnig genug ist, werden Dichteschätzung bzw. Intensitätsschätzung verwendet. Die Dichteschätzung schätzt die räumliche Ballung von punktuell auftretenden, nicht weiter unterspezifizierten Einheiten, zum Beispiel Varianten. Die Intensitätsschätzung schätzt die bereits graduelle Werte, zum Beispiel die relative Variantenhäufigkeiten an einem Ort, die Grundlage für die Intensitätsberechnung bilden. Die Intensitätsschätzung liefert Intensitätskarten, die jeweils das Intensitätsfeld einer einzelnen Variante angeben. Ein Beispiel für Intensitätskarten ist der Bayerische Sprachatlas (Figure 1). Die Vereinigung der Intensitätsfelder verschiedener Varianten in einer Karte folgt eine kombinierte Intensitätskarte. Wenn man dabei an jedem Punkt die dominante Variante verzeichnet, für die die Farbintensität den Grad der Dominanz angibt, nennt man die resultierende Karte die gradierte Flächenkarte.



**Figure 1:** Intensitätskarten vom Bayerischen Sprachatlas. Links nach rechts: beobachtetes Vorkommen der Einzelvariante, geschätzte Intensität der Einzelvariante, kombinierte Intensitätskarte aller Einzelvarianten.

Geostatistik im weiteren Sinne enthält die Methoden, die geographisch verteilte Daten mehrerer Varianten oder Variablen aggregieren oder korrelieren, ohne sich auf ihre geographische Lage zu stützen. Seguy(1971) führte den Terminus Dialektometrie ein, um Dialektunterschiede zu bestimmen. Goebel(1982) führt die Ort-Ort-Distanzmatrix ein, um die Analyse und Kartierung mit Hilfe der Distanzmatrix zu ermöglichen.

Mit Hilfe der Distanzmatrix für die aggregierte Ähnlichkeit der Orte zu Dialektgebieten kann jedem Ort ein Cluster zugewiesen werden. So ist die Clusteranalyse für Dialekte möglich und ist sie daher zu einem Standardverfahren der Dialektometrie geworden. Davon ist die hierarchische Clusteranalyse am häufigsten verwendet, in der zunächst genauso viele Cluster wie Orte mit je einem enthaltenen Ortspunkt definiert und dann sukzessiv die beiden Cluster mit geringster Distanz verschmolzen werden, bis eine bestimmte Anzahl von Clustern erreicht wird(z.B.

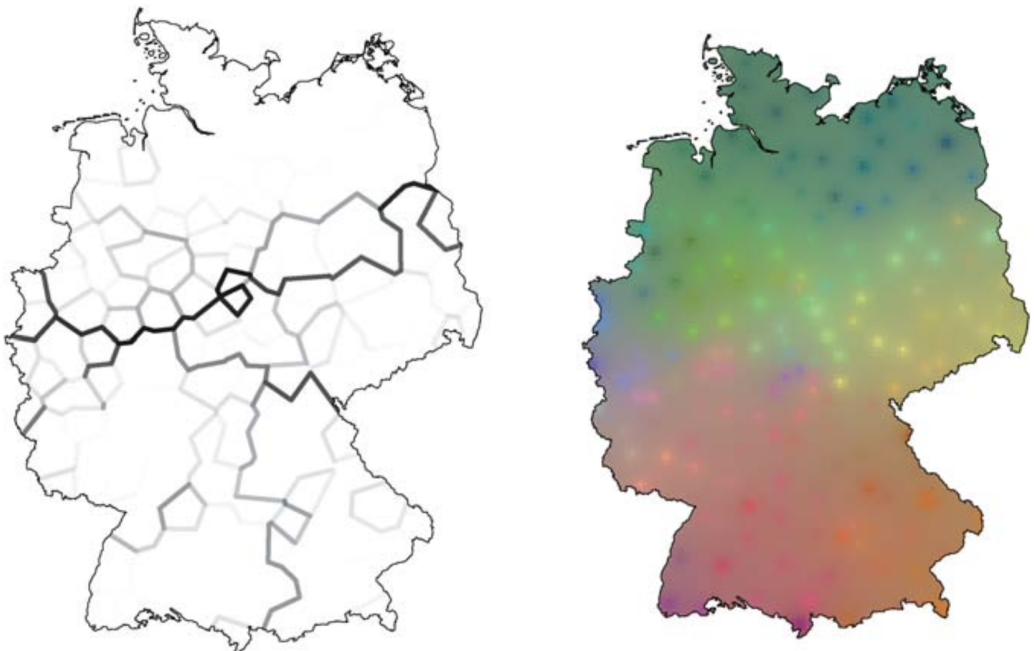


**Figure 2:** Dendrogramm als Ergebnis der hierarchischen Clusteranalyse

1). Die Geschichte der Verschmelzungen ist in einer Hierarchie in Form eines Binärbaums gespeichert. Als Ergebnis bekommt man einen sogenannten Dendrogramm (Figure 2), in dem jeder Cluster ein Dialektgebiet repräsentiert. Die Anzahl von Clustern in dem Ergebnis kann durch Abschnitten auf y-Achse des Dendrogramms geändert werden. Mit verschiedenen Abschnittspositionen kriegt man verschiedenen Anzahl von Clustern. Deswegen sind die Ergebnisse der Clusteranalyse oft instabil. Um die Ergebnisse zu stabilisieren, werden "fuzzy clustering" angewendet, die robustere Ergebnisse durch Wiederholung der Clusteranalyse mit leichten Veränderungen in Ausgangsdaten bekommen. Dazu gehören bootstrapping clustering und noisy clustering, die methodisch unterschiedlich, aber deren Ergebnisse ähnlich sind.

Ein alternativer Ansatz für hierarchische Clusteranalyse ist die multidimensionale Skalierung (MDS), die ebenfalls auf Distanzmatrizen basiert, jedoch keine scharfe Dialektgebiete, sondern Kontinua liefert. Die Ortspunkte werden in einem zweidimensionalen Raum abgebildet, so dass die resultierenden Distanzen den paarweisen Distanzen in der Distanzmatrix möglichst gut approximieren. Um mehr Anteil der Variation erklären zu können, bildeten Nerbonne et al. (1999) die Ortspunkte in einen dreidimensionalen RGB-Farbraum ab. Infolgedessen sind in ihren resultierenden Karten sprachlich nähre Orte mit ähnlichen Farben eingefärbt (Figure 3 rechts). Damit haben sie die Abweichungen von den originalen Distanzen trotz relativ wenigen Dimensionen gedämpft und etwas höheren Anteil der Variation erklärt. Ein weiterer Vorteil ist, dass die Bruchkanten in der eingefärbten Karte als Dialektgrenzen interpretiert werden können.

Die eben genannte Clusteranalyse und MDS verwenden beide Distanzmatrizen. Im Gegen-



**Figure 3:** Abgrenzung der aggregierten Phonation(links) und Hauptdimensionen der entsprechenden MDS(rechts)

satz dazu gibt es zwei weitere Verfahren, nämlich Hauptkomponentenanalyse(PCA) und Faktorenanalyse(FA), die sich nicht auf Distanzmatrizen basieren, sondern auf den originalen unaggregierten Daten. Die beiden Verfahren kombinieren die Vorteile von Clusteranalyse und von MDS, in dem sie sowohl kategoriale Unterschiede als auch Gradualitäten erfassen. Ein weiterer Vorteil besteht darin, dass sie sowohl dominante Strukturen als auch nicht-dominante, latente Muster erfassen, die relativ schwer sichtbar sind. Analog zu Clusteranalyse und MDS muss bei diesen beiden Verfahren auch die Anzahl der Kategorien bzw. Dimensionen gewählt werden, jedoch haben sie ab bestimmten Werten geringen Einfluss auf dem gesamten Ergebnis. Die Ergebnisse sind überlappende, unscharfe und mitunter diskontinuierliche Dialektgebiete(Figure 5).

Die Tabelle in Figure 4 gibt eine Zusammenfassung der oben genannten Verfahren.

Clusteranalyse	MDS	PCA & FA
Klare Grenzen	Kontinuierlich, mit Übergangsbereich	Mitunter diskontinuierlich, mit Übergangsbereich
Distanzmatrix	Distanzmatrix	Keine Distanzmatrix
Ergebnis: nicht stabil	Ergebnis: stabil	Ergebnis: stabil

**Figure 4:** Zusammenfassung verschiedener Verfahren

### 3. Studien zu Teilregionen des deutschsprachigen Raums

Nun werden Ergebnisse ausgewählter Studien zu dem aktuellen Forschungsstand über den sprachräumlichen Struktur des deutschsprachigen Gebiets vorgestellt. Diese Studien lassen sich in zwei Teilbereiche unterteilen: Studien zu Teilregionen des deutschsprachigen Raums und Länder-/gesamtsprachgebietsweite Studien. Danach erzählen wir unsere eigene Analyse und Ansicht zu den Ergebnissen dieser Studien.

#### 3.1. Studie 1: Analyse des Alemannisch-bairischen Kontaktraums

##### 3.1.1. Experimente

Pickl(2013) führte zunächst eine detaillierte Untersuchung auf den kombinierten Intensitätsskarten unter Rückgriff auf raumstatistische Kennwerte für Komplexität, Kompaktheit und Homogenität. Dann machte er eine Faktorenanalyse der Verteilungen von 12k lexikalischen Formen in 735 SBS-Karten und versuchte, die räumliche Strukturen zu finden, die von der Verteilung der Dialekte abhängig sind. Die Ergebnisse wurden auf den SBS-Karten bezeichnet.

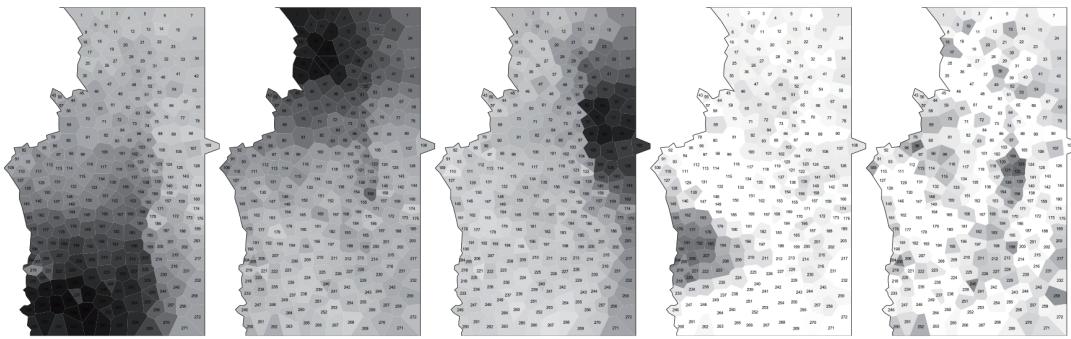
Pröll(2015) führte eine Untersuchung und Faktorenanalyse auf 2155 SBS-Karten durch, um die räumlichen Verhältnisse der Varianten bzgl. Lexik, Phonologie und Morphologie zu vergleichen. Als Kenngröße nutzte er auch die Kennwerte für Komplexität, Kompaktheit und Homogenität der Karten.

##### 3.1.2. Ergebnisse

Pickl(2013):

Durch der Untersuchung und Analyse der kombinierten Intensitätsskarten wurde auffällige

Groß- oder Kleinräumigkeit der Variantengebiete bestimmter semantischen Bereiche gefunden. Jedoch zeigten sich die Ergebnisse keinen Zusammenhang zwischen Gebrauchsfrequenz der Lexeme und der Größe ihrer Variantengebiete. Das Ergebnis der Faktorenanalyse zeigte sich nicht nur die dominanten Dialektgebiete wie Allgäuerisch, Nordostschwäbisch und Mittelbairisch, sondern auch erstmals schwächere, nicht dominante latente Strukturen, und zwar sozioökonomischen Bezugsräume und urbane Zentren, die in den letzten beiden Intensitätskarten in Figure 5 dunkel gefärbt sind. Das heißt, dass die Verteilung der sprachlichen Varianten nicht nur von der geographischen Unterteilung der Sprachgebiete beeinflusst wird, sondern auch von den in den Daten latent vorhandenen Strukturen. Aufgrund der dunklen Färbung von der Region des Markortes und ihre Umgebung lässt sich sagen, dass die früheren Einzugsgebiete der Markorte einen Effekt auf die Verbreitung der sprachlichen Formen hatten.



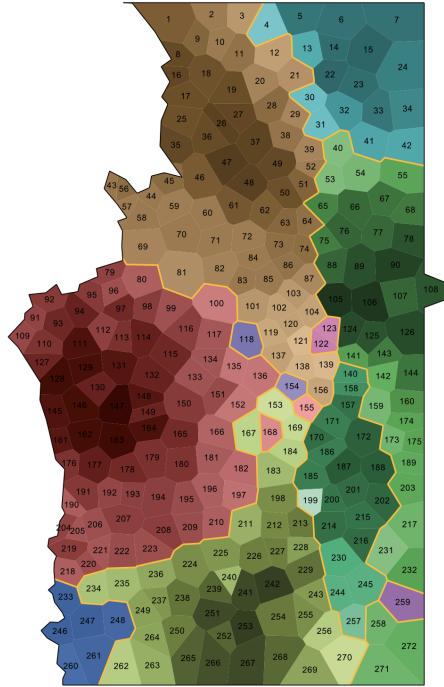
**Figure 5:** Fünf Faktoren der Wortgeographie des SBS. Von links nach rechts: Allgäuerisch, Nordostschwäbisch, Mittelbairisch, Einzugsgebiet Markt Memmingen, Urbanität. Die Prozentwerte repräsentieren den durch den jeweiligen Faktor erklärten Anteil an der Gesamtvariation

#### Pröll(2015):

Dabei wurde raumstatistische Abweichung zwischen Lexikkarten gefunden, während es keinen signifikanten Unterschied zwischen Phonologie bzw. Morphologie aufwies. Durch Faktorenanalyse fand Pröll feinere Untergliederungen und schärfere oder breitere Übergangsgebiete von einzelnen Teilsystemen. Ein Beispiel für scharferes Übergangsgebiet ist der Übergang von dem blauen Gebiet zu dem roten Gebiet in dem unteren Bereich in Figure 6. Es weist einen rasanten Wechsel von einer Variante zu einer anderen Variante. Ein breiter Übergangsgebiet ist beispielsweise der Übergangsbereich oben rechts in Figure 6 vom dunklen Grün zum hellen Grün und dann zum hellen Türkis und danach zum dunklen Türkis.

#### **3.1.3. Analyse und Diskussion**

Die Ergebnisse der Studie von Pickl(2013) zeigt, dass die Verteilung der sprachlichen Varianten nicht nur von der diatopischen Faktoren der Sprachgebiete beeinflusst wird, sondern auch



**Figure 6:** Dominante Faktoren des SBS

von sozialen Faktoren wie zum Beispiel die soziale Schicht und Bildung. Das stimmt mit der Intuition überein, dass Händler aus verschiedenen Orten verschiedenen Dialekte bei dem Handel in Markttorten mitbringen und durch Kommunikation neue Varianten dort einführen. Weitere bestätigte Intuition ist, dass Menschen aus wirtschaftlichen Gründen in Großstädte umsiedeln und damit die Varianten aus ihren Heimatstädten dorthin mitbringen. Durch Verwendung verschiedener Varianten in den urbanen Zentren mit hoher Bevölkerungsdichte lässt sich im Laufe der Zeit neue Varianten entwickeln, die spezielle urbane Charakter aufweisen. Dies entspricht dem in der Vorlesung erwähnten diastratischen Sprachwandel und dem Schichtenmodell. Dasselbe Phänomen existiert auch in der Entwicklung der chinesischen Dialekte. Detaillierte Diskussion folgt am Ende dieses Essays.

Im Vergleich zu der Studie von Pickl(2013) hat die Studie von Pröll(2015) einen Fortschritt gemacht, in dem die Abgrenzung verschiedener Dialektegebiete sowie die Intensität von dem erklärten Anteil durch Färbung übersichtlich gemacht werden und damit die Karten feinere Strukturen aufweisen. In Figure 6 hat jeder Cluster einen zentralen Bereich mit sehr dunkler Farbe, während die von dem Zentrum entfernte Bereiche sukzessiv heller gefärbt sind. Meiner Meinung nach beschreibt diese innere Heterogenität auch das diastratische Effekt in dem jeweiligen Sprachgebiet. Ein auffälliges Detail ist, dass es ein paar Sprachinseln in der Karte gibt. Davon unterscheiden sich zwei Arten von Sprachinseln, nämlich die Inseln mit einzigartigen

Farben, zum Beispiel die lila gefärbte Insel unten rechts(beziffert mit 259), und die Inseln mit gleicher Farbe eines großen Gebietes wie zum Beispiel die zwei kleine rot gefärbte Inseln in der Mitte der Karte(beziffert mit 168 und 155). Nach unserer Vermutung ist die Entstehung der ersten Art von Inseln vermutlich auf den Purismus zurückzuführen. Purismus ist die Manifestation des Wunsches einer Sprachgemeinschaft, ihre Sprache vor (vermeintlichen) fremden oder unerwünschten Elementen zu bewahren. Am Beispiel des lila gefärbten Gebiets unten rechts lässt sich vermuten, dass die Bewohner dort die Sprache von dem grünen Gebiet nicht annehmen wollen und ihre lokale Sprache stets behalten. Das ist wahrscheinlich den Grund dafür, dass diese Gebiete einzigartige Farbe besitzen, und dass es keine Nachbargebiete mit gleicher oder ähnlicher Farbe in der Nähe gibt. Dagegen konnte die zweite Art von Inseln ursprünglich mit dem großen roten Bereich verbunden werden und war an der Grenze zu dem grünen Bereich. Jedoch hat es sich im Laufe der Zeit irgendwie von dem großen roten Bereich abgesplittet. Allerdings wollen die Bewohner die Sprache von dem grünen Gebiet auch nicht annehmen und lassen sich somit zu einer Insel geworden. Das ist also vielleicht ein diaphasischer Wandel der Sprachvariante des roten Bereichs mit puristischer Motivation.

### **3.2. Studie 2:**

#### **3.2.1. Experimente**

Mathussek (2014) wählte 517 Items aus der Sprachatlas von Mittelfranken (SMF) aus und verwendete Clusteranalyse mittels des GabMap(Nerbonne et al. 2011), um die traditionelle dialektgeographische Raumaufteilung mit den wahrgenommenen dialektlogischen und dialektonometrischen Raumauftteilungen im mittelfränkischen Untersuchungsgebiet zu vergleichen. Zunächst einmal wurde in der Clusteranalyse Diakritika in der Teuthonista verwendet. Allerdings wurde damit das eigentliche Ziel dieser Studie, und zwar die Untersuchung der räumlichen Variation von den Dialekten, verhindert. Deswegen wurde die Clusteranalyse unter Elimination der meisten Diakritika wiederholt.

#### **3.2.2. Ergebnisse**

Das Experiment mit Diakritika zeigte, dass die erzeugenden Cluster mit den von den einzelnen Exploratoren untersuchten Gebieten stark abhängig sind. Davon stellte er sich heraus, dass sich in den Daten des SMF diejenigen Mundarten insgesamt am ähnlichsten sind, die von einem Explorator erhoben wurden. Es ist jedoch relativ unabhängig davon, ob sich die Ortsmundarten tatsächlich auch sprachlich besonders ähnlich sind oder nicht.

Das Experiment ohne Diakritika erfolgt acht Cluster, wobei sieben davon keine unmittelbar erkennbaren Exploratoreneffekte aufweisen. Matussek stellte sich fest, dass sich die Übereinstimmungen zwischen den dialektmetrischen und den traditionellen bzw. den wahrnehmungsdialektologischen Raumdifferenzierungen überraschend hoch war. Für eine synoptische Gesamtdarstellung griff er aber trotzdem meist auf die traditionellen Grenzen

zurück.

### **3.2.3. Analyse und Diskussion**

Die Datenerhebung für dieses Experiment wurde jedoch manuell durchgeführt und verdeckte damit die tatsächliche Wahrnehmung der Befragten. Die dialektometrische Grenzen unterliegen daher dem Einfluss des Explorator und sind somit teilweise unsicher.

## **4. Länder-/gesamtsprachgebietseite Studien**

### **4.1. Studie 3:**

#### **4.1.1. Die Fundamentale Dialektologische Hypothese**

In der Dialektologie gibt es eine fundamentale Hypothese(engl. Fundamental Dialectological Postulate), die besagt, dass geographisch nahliegende Sprachvarianten tendenziell ähnlicher sind als die geographisch entfernte.

#### **4.1.2. Experiment**

Jeszenszky & Weibel verwendeten die Daten aus dem Syntaktischen Atlas der Deutschen Schweiz(SADS, vgl. Glaser 2006), um die eben genannte Hypothese der Korrelation zwischen linguistischer und geographischer Distanz zu prüfen. In ihrer weiteren Studie verwendeten sie Trend Surface Anaslysis und Regressionsanalyse, um den Übergangsbereich von Einzelvarianten zu messen und zu modellieren.

#### **4.1.3. Ergebnisse**

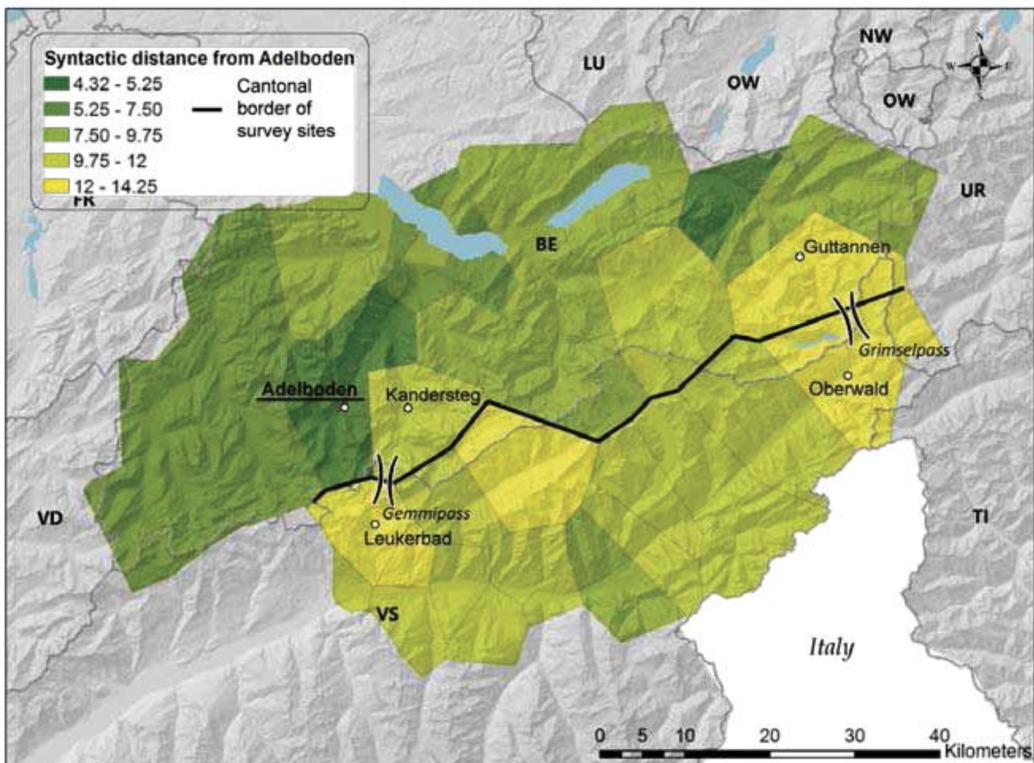
Die Ergebnisse zeigten, dass es regionale Unterschiede in der Korrelation zwischen linguistischer und geographischer Distanz gibt. Alle Korrelationskoeffizienten waren wesentlich signifikant auf der globalen Ebene, während es auf der regionalen Ebene niedrigere Korrelationskoeffizienten aufwies. Die Reisezeit war ein wesentlich besserer Prädiktor für die linguistische Distanz zwischen zwei Orten, während die euklidische Distanz zwischen zwei Orten unterschätzte die Unterschiede der Dialekte in kleinen Räumen.

Durch die Messung und Modellierung des Übergangsbereichs wurden Unterschiede zwischen scharfen Isoglossen und sanften Übergangszonen, die zwischen den Verbreitungsgebiete von Varianten auftreten, gefunden.

#### 4.1.4. Analyse und Diskussion

Das Ergebnis entspricht unserer Wahrnehmung, dass in der Alpenregion, obwohl die beiden Punkte auf der Karte auf einer geraden Linie nahe beieinander liegen, jedoch einen großen Höhenunterschied bestehen kann, der tatsächlich zu einer viel größeren Entfernung führt.

Eine interessante Entdeckung ist es, dass es je einen Bridgingeffekt in zwei Gebirgspässen aufweist, und zwar in Gemmipass und Grimselpass. Der Gemmipass, zum Beispiel, kann heute nicht direkt durchqueren lassen, war aber ein Hauptpass im Mittelalter, in dem sich meisten Dialekte entwickelten.(Figure 7)



**Figure 7:** Bridgingeffekt in Gemmipass und Grimselpass

Was wir methodisch innovativ finden ist die Verwendung einer neuen Methode für die Ermittlung der linguistischen Distanz. Typischerweise ist die linguistische Ähnlichkeit zwischen zwei Varianten mittels Editierdistanz(engl. edit distance) berechnet, zum Beispiel Levenshtein-Distanz. Jedoch haben die Autoren gemeldet, dass pro Untersuchungsort mehrere Varianten in dem SADS existieren können. Deshalb ermittelten sie stattdessen die syntaktische Distanz zwischen je zwei Untersuchungsorten, in dem sie den Anteil der Varianten für

eine gegebene Variable in zwei Untersuchungsorten und damit die Differenz des Anteils berechneten. Die syntaktische Distanz erfolgt dann durch die Summe der Differenzen für alle Untersuchungsorten. Danach wurden die Korrelationskoeffizienten zwischen den linguistischen und geographischen Distanzen ermittelt.

## 4.2. Studie 4:

Scherrer und Stöckle(2016) gaben einen Überblick über die Verteilung von Sprachvarianten in deutschsprachiger Schweiz und verglichen sie auf verschiedenen linguistischen Ebenen.

### 4.2.1. Experiment

Das Experiment besteht aus zwei Subtasks: die Repräsentation der Daten und die dialektometrische Analyse.

Für die Datenrepräsentation kombinierten sie die syntaktischen Daten aus SADS und die lexikalischen, phonetischen und morphologischen Daten aus dem Sprachatlas der deutschen Schweiz(SDS) und bildeten sie damit ein gemeinsames Ortsnetz. Die Idee ist, Information aus möglichst vielen Karten zu aggregieren. Dafür wurden 5 Datenmatrizen durch Aggregation der Information aus den Karten generiert, wobei eine für den gesamten Datensatz und die anderen für je eine linguistische Ebene. Jeder Eintrag in der Datenmatrix repräsentiert eine Variante, die in einem gegebenen inquiry Point für eine gegebene Phänomen verwendet wird. Dann wurden die Datenmatrizen zu Ähnlichkeitsmatrizen konvertiert, die Paare von inquiry points enthalten. Jeder Eintrag der Ähnlichkeitsmatrix ist der Ähnlichkeitswert der entsprechenden Zeilen in der Datenmatrix, die mittels des RSVJaccard-Algorithmus ermittelt wird. Danach wurden die Ähnlichkeitsmatrizen weiter zu niedrigdimensionalen Wertematrizen “kondensiert”, wobei jede Zeile genau einen Wert besitzt. Aspekte der Raumverteilung von Dialekten wurden durch entsprechenden Werte repräsentiert. Hierbei werden inquiry points mit ähnlichen Ähnlichkeitswerten mittels hierarchischer Clustering geclustert. Um die Daten zu visualisieren, wurde Jenk's natural breaks classification method verwendet, die Zuweisungen der Farben für jeden Cluster aus Wertematrizen beschaffen.

Bei der dialektometrischen Analyse wurde zunächst Konsistenz und räumliche Autokorrelation der Daten untersucht. Dafür wurde jeder Datensatz bzgl. der Datenmatrix verglichen, in dem Cronbach's alpha berechnet wurde. Cronbach's alpha ist eine Koeffizient für Konsistenz, die beschreibt, wie weit verschiedene Variablen einer Datenmatrix dieselbe Verteilung aufweisen. Der Wert von Cronbach's alpha liegt zwischen 0(alle Variablen weisen verschiedenen geographische Verteilung auf) und 1(alle Variablen weisen dieselbe geographische Verteilung auf). Ein typischer Schwellenwert für gute Konsistenz ist 0,7.

Bei dem Vergleichen der Datensätze bzgl. der räumlichen Autokorrelation wurden zum einen die Pearson's Korrelationskoeffizient zwischen den linguistischen Ähnlichkeitsmatrizen und

geographischen Distanzmatrizen berechnet, zum anderen das local incoherence scores. Je kleiner das local incoherence scores, desto besser ist die Messung der Dialekte.

Darüber hinaus führten sie jeweils eine hierarchische Clusteranalyse, sowohl auf den gesamten Daten als auch den vier linguistischen Ebenen, und generierten sie die entsprechenden Dendrogramme.

#### 4.2.2. Ergebnisse

##### Subtask 1: Repräsentation der Daten

Nach der Visualisierung wurden 10 Klassen für die Sprachvarianten generiert, je mit einer Farbe vom Rot zum Blau auf der Karte eingefärbt.

##### Subtask 2: Dialektometrische Analyse

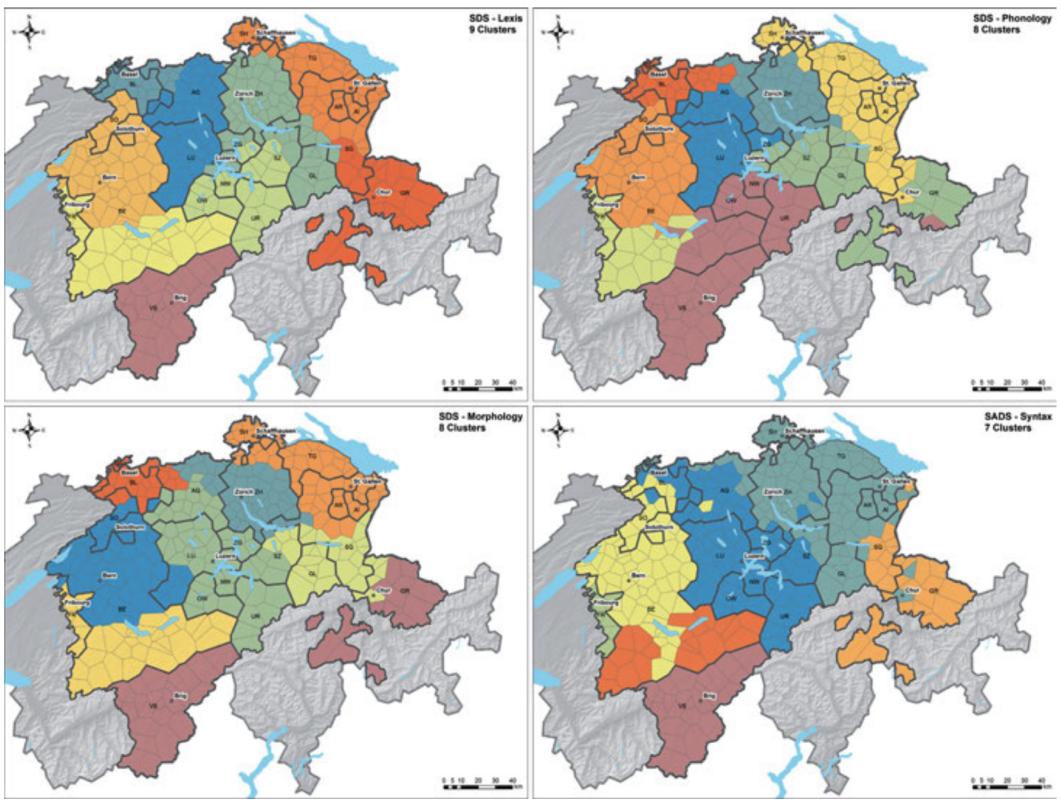
Die Werte für Cronbach's alpha für alle fünf Datenmatrizen lagen über den Schwellenwert 0,7. Das heißt, dass sowohl die gesamten Daten als auch die Teilmenge der Daten auf den vier linguistischen Ebenen konsistent sind. Auch die am wenigsten konsistenten Daten, nämlich die syntaktischen Daten, besitzen einen hohen Wert von 0,81.

Linguistic Level	Cronbach's Alpha
Morphology (118 variables)	0.93
Phonology (100 variables)	0.89
Lexis (64 variables)	0.89
Syntax (68 variables)	0.81
All levels (350 variables)	0.97

**Figure 8:** Tabelle für Cronbach's alpha auf verschiedenen linguistischen Ebenen

Die hierarchischen Clusteranalyse auf den gesamten Daten bestätigte die Annahme, dass es eine Ost-West-Unterteilung(oben links & rechts in Figure 9) und eine Nord-Süd-Unterteilung(unten links in Figure 9) der Deutschschweiz gäbe. Die Clustergrenzen entsprechen ungefähr den Kantongrenzen.

Die hierarchischen Clusteranalysen auf den vier linguistischen Ebenen erwiesen Unterschiede im Raummuster zwischen Syntax und die anderen drei Ebenen. In den Karte für Lexik,



**Figure 9:** Hierarchische Clusteranalysen zur Deutschschweiz bzgl. Lexik(oberen links), Phonologie(oberen rechts), Morphologie(unten links) und Syntax(unten rechts), aus Scherrer & Stöckle(2016)

Phonologie und Morphologie wurden das Berner Oberland und den Kanton Freiburg in einem Cluster aggregiert, nämlich der gelbe Cluster. Jedoch wurden sie in der Syntaxkarte in zwei verschiedenen Clustern aggregiert, also Kanton Freiburg in dem grünen und Berner Oberland in dem orangen. Dieser Unterschied kann auf den Spezifika der syntaktischen Variation zurückzuführen.

#### 4.2.3. Diskussion

Nach entsprechender Untersuchung sind wir die Meinung, dass die unterschiedliche Erhebungszeit und Erhebungsweise von SDS und SADS auch den Unterschied verursacht haben können. Das SDS wurde zwischen 1938 und 1958 gesammelt und repräsentiert den linguistischen Status in den früheren 20. Jahrhundert. Im Vergleich dazu wurde das SADS zwischen 2000 und 2002 gesammelt, welche ca. 50 Jahre später ist. Dazwischen kann bestimmt Sprachwandel auftreten.

Goebl(2010) hat einen einfacheren Algorithmus für die Berechnung der Ähnlichkeitsmatrix vorgeschlagen, der einen sogenannten Relative Identity Value(RIV) für jedes Paar von inquiry points ermittelt. Dabei wurden Daten durch on-site Interviews mit Fragen zu Phänomen bzgl. der Varianten auf verschiedenen linguistischen Ebenen gesammelt. Allerdings funktioniert dieser Algorithmus nur für Datensätze mit eindeutiger Antwort. Die SDS Karte enthält für eine Frage jedoch mehrere Antworten. Interessanterweise haben Scherrer und Stöckle in ihren alternativen Ansatz die Jaccardmatrix eingeführt, die in Computer Vision und Natural Language Processing häufig als sogenanntes "Intersection over Union(IoU)" verwendet wird.

### **4.3. Studie 5:**

#### **4.3.1. Experiment**

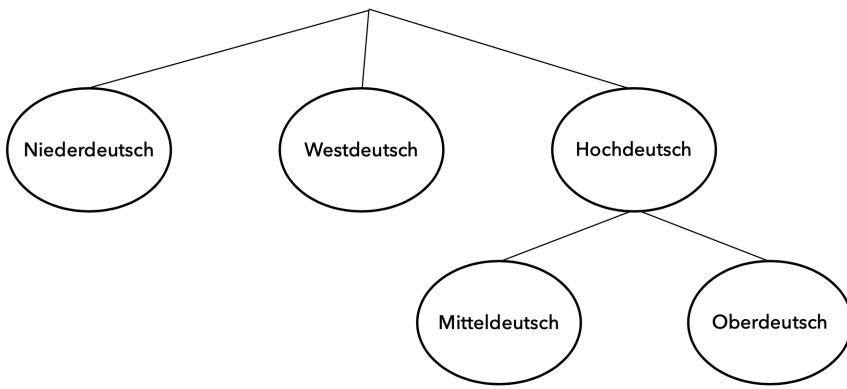
Lameli(2013) führte eine dialektometrische Auswertung der historischen Dialekte auf heutigen Landkreisen Deutschlands durch, um sprachliche Beziehungen unter heutigen Landkreisen in Deutschland zu untersuchen. Die Daten stammten aus dem Sprachatlas des Deutschen Reichs. Für die Auswertung ermittelte er Ähnlichkeitsmatrix, in dem er die Landkreise anhand der sprachlichen Merkmalen paarweise verglich. In einer Karte der “globalen Similarität” wurde pro Landkreis der durchschnittliche Ähnlichkeitswert zu allen anderen Landkreisen abgebildet. Diese durchschnittliche Ähnlichkeit wurde als “kommunikative Reichweite” interpretiert. Außerdem versuchte er, die Ergebnisse durch wiederholter Clusteranalyse mit leicht veränderten Daten zu stabilisieren.

#### **4.3.2. Ergebnisse**

Durch die Untersuchung fand Lameli, dass die Dialekte im Nordwesten Deutschlands und schwäbischen Raum relativ unähnlich wie Dialekte in allen anderen Orten waren. Dagegen teilten Dialekte im ostmitteldeutschen, bairischen und rheinfränkischen Raum viele Gemeinsamkeiten. Die oben genannte “stabile Clusteranalyse” lieferte neue Gliederungen der Dialektlandschaft Deutschlands. Es gliederte sich zunächst in Niederdeutsch, Westdeutsch und Hochdeutsch, wobei das Hochdeutsch sich wieder in Mitteldeutsch und Oberdeutsch gliederte.

#### **4.3.3. Diskussion**

Wir stimmen mit dem Autor überein, dass die Entstehung und Entwicklung der Dialekträume nicht nur mit diatopischen Faktoren, sondern auch mit diastratischen und diaphasischen Faktoren beeinflusst werden. Insbesondere spielen die rezenten Binnenmigration, die Gehaltsstruktur, die Konfession und politischen Standpunkt große Rollen. Aufgrund der Binnenmigration in dem Land wurden Varianten von einem Ort nach anderen Orten mitgebracht, dort mit der lokalen Varianten oder mit anderen ebenfalls mitgebrachten Varianten gemischt und damit sich neue Varianten entwickelt. Menschen mit verschiedenen Gehalt und



**Figure 10:** Neue Gliederungen der Dialektlandschaft Deutschlands, nach der Untersuchung von Lameli(2013)

Bildung bilden sich verschiedenen Schichte, deren Varianten sich variieren können. Es ist auch einfach zu verstehen, dass Menschen mit verschiedenen Konfessionen und politische Standpunkten unterschiedlichen Lexik, Syntax etc. nutzen. Es kann sogar Purismus zwischen verschiedenen Konfessions-bzw. Politikgruppen geben und somit Sprachinseln verursachen.

## 5. Kritischer Punkt über Clusteranalyse

In den oben erwähnten Studien werden hierarchische Clusteranalyse häufig verwendet und damit gute Ergebnisse erzielt. Jedoch hat dieser Methode auch Nachteile. Einerseits hat sie einen hohen Zeitaufwand in Höhe von  $O(n^2 \log(n))$ , andererseits benötigt sie quadratischen Speicherplatz  $O(n^2)$ . Das heißt, die hierarchische Clusteranalyse ist weder zeitlich noch räumlich effizient. Es lohnt sich, in der zukünftige Studien den Algorithmus zu verbessern oder effizientere Methoden zu entwickeln.

## 6. Diskussion über die Sprachvariation in China im Hinblick auf Geolinguistik

In Studie 1 von Pickl(2013) und Pröll(2015) haben wir über den diastratischen und diaphasischen Abgrenzung der Dialekte und das Schichtmodell diskutiert. Die Ergebnisse der Studien von Pickl und Pröll zeigte, dass die Verteilung der sprachlichen Varianten von geographischen(diatopischen), sozialen(diastratischen) und zeitlichen(diaphasischen) Faktoren beeinflusst wurde. Dasselbe Phänomen erkennen wir auch in der Entstehung, Entwicklung und Abgrenzung der chinesischen Dialekte. Wir zeigen das mit dem Beispiel des Dialektes in

Shanghai.

Der Dialekt in Shanghai(e.g. Shanghainesisch) gehört zur Wu-Dialektgruppe des Chinesischen, die um die Mündung des Flusses Jangtse herum gesprochen wird. Unter Verfeinerung gehört er zu der Taihu-Wu Subgruppe, die in Süd-Jiangsu dominant ist. Im Vergleich zu anderen Dialekte in der Wu-Dialektgruppe ist die Entwicklung von Shanghainesisch von verschiedenen Dialekten stark beeinflusst. Historisch gesehen war Shanghai ursprünglich(1277 n.Chr.) ein Teil des Landkreises Huating, bevor es 1290 n.Chr. zu einem unabhängigen Landkreis geworden ist. Aufgrund der geographischen Lage können die dort verwendeten Sprachvarianten auf dem Wu-Dialekt zurückgreifen und sich von Dialekten anderen Region unterscheiden. Damit entsteht die diatopische Abgrenzung der Dialekt in Shanghai.

Damals galt Shanghai als ein abgelegener Ort und es gab relativ wenige wirtschaftliche Entwicklung. 1860 wurde die Provinzen Jiangsu und Zhejiang, die sich neben heutigem Shanghai befinden und wirtschaftlich dominant waren, wegen des Kriegs zerstört. Daraus folgt Flüchtlingswellen reicher Bewohner aus den beiden Provinzen nach Shanghai, die die Wirtschaft Shanghais erstmal vorantrieben. Aufgrund der Öffnung von Shanghai als ein Handelshafen lebten dort immer mehr ausländische Immigranten. Die zunehmend häufigere Handelsaktivitäten führten wieder zu mehrere Einwanderungswellen nach Shanghai. Seitdem ist Shanghai ein internationale Metropole geworden und erfuhr dauerhaften Aufschwung, die in den 1930en Jahren einen Höhepunkt erreichte. Seit 1990er Jahren erfuhr Shanghai aufgrund der Extension und Entwicklung des Landkreises neuen Aufschwung, in dem Migranten wieder einmal aus verschiedenen Region Chinas und aus anderen Ländern eingewandert sind.

Durch die beiden Migrationswellen wurden Varianten aus verschiedenen Region Chinas in die Stadt mitgebracht und neue Varianten dort sich entwickelt. Da die erste Einwanderungsgruppe hauptsächlich aus Bewohner aus Provinzen Jiangsu und Zhejiang besteht, enthält der damals sich entwickelte Dialekt in Shanghai viele Elemente in den Dialekten in den beiden Provinzen. Im Laufe der Zeit sind Menschen aus anderen Region eingewandert und die Charakter der Varianten aus ihren Heimat in die lokale Varianten in Shanghai eingeführt. Da die in verschiedenen Epochen Menschen verschiedenen Lebensumstände haben und unterschiedliches Alter haben, weist es einen diaphasische Abgrenzung auf.

Durch die dauerhafte Einwanderung und Entwicklung der Stadt entstanden innere Grenzen zwischen Bezirken in Shanghai, sowohl geographisch als auch sozio-und ökonomisch. Wirtschaftlich stärkere Menschen erhielten bessere Bildungs-und Sozialressourcen. Im Laufe der Zeit entstanden unterschiedliche soziale Schichte, deren verwendeten Varianten sich variieren. Dies erfolgt schließlich diastratische Abgrenzungen zwischen den lokalen Varianten. Die Einwanderung ausländischer Immigranten führte neue Varianten auf der lexikalischen, phonologischen, morphologischen und syntaktischen Ebenen. Insbesondere hat sich neue Wörter entwickelt, die sich total auf Lehnwörter zurückgreifen und keine chinesische Wurzel haben. Diese Varianten wurden von bestimmten Gruppen der Bewohner akzeptiert, jedoch von anderen aufgrund puristischer Motivation total oder teilweise abgelehnt. Dadurch entstanden Sprachinseln in bestimmten Region der Stadt. Dasselbe Phänomen existiert auch durch

Binnenmigration zwischen verschiedenen Bezirke der Stadt. Es lässt sich daraus feststellen, dass der Neologismus und Purismus für die Entwicklung der Sprachvarianten in Shanghai gemeinsam große Rolle spielen. Im Hinblick auf die oben genannte Geschichte lässt sich sagen, das heutige Raumverhältnis des Dialektes erweist ein kulturelles Gedächtnis, das historische Einflüsse bewahrt und bis heute nachwirken kann.

## **Würdigung**

Hierbei möchten wir uns bei Frau PD. Dr. Eleonore Brandner für die Betreuung und Unterstützung für das Referat bzw. diese Ausarbeitung bedanken.

## **Referenzliste**

1. August Leskien, Die Declination im Slavisch-Litauischen und Germanischen, Preisschrift der Societas Jablonoviana. Leipzig 1876. (Digitalisat und Volltext im Deutschen Textarchiv)
2. Lyle Campbell(1999), Historical linguistics. An introduction. Edinburgh Univ. Press: Edinburgh.
3. Georg Wenker, Das rheinische Platt, den Lehrern des Rheinlandes gewidmet. Düsseldorf 1877.
4. Georg Wenker: Schriften zum „Sprachatlas des Deutschen Reichs“. Gesamtausgabe. Herausgegeben und bearbeitet von Alfred Lameli. Unter Mitarbeit von Johanna Heil und Constanze Wellendorf. 3 Bände. Hildesheim, New York, Zürich 2013f.
5. Johann-Mattis List, "Wörter mit Vergangenheit," in Von Wörtern und Bäumen, 05/08/2018, <https://wub.hypotheses.org/367>.
6. Hugo Schuchardt, Über die Lautgesetze. Gegen die Junggrammatiker (1885) (Digitalisat und Volltext im Deutschen Textarchiv)
7. Jules Gilliéron, Atlas Linguistique de la France (1902–1910) (with Edmond Edmont), Paris: E. Champion.
8. Manuela Lanwermeyer (2019), Sprachwandel und Kognition, <https://elibrary.steiner-verlag.de/book/99.105010/9783515120241>
9. Simon Pickl, Simon Pröll(2019), Geolinguistische Querschnitte und Tiefenbohrungen in Bayern und darüber hinaus.
10. A.K. Jain, M.N. Murty, P.J. Flynn(1999), Data Clustering: A Review

11. Simon Pickl(2013), Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben
12. Simon Pröll(2015), Raumvariation zwischen Muster und Zufall: Geostatistische Analysen am Beispiel des Sprachatlas von Bayerisch-Schwaben
13. Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen(2011), Gabmap—a web application for dialectology. *Dialectologia*, Special Issue II. 65–89.
14. Jeszenszky, Péter & Robert Weibel(2014), Correlating morphosyntactic dialect variation with geographic distance: Local beats global. In Kathleen Steward, Edzer Pebesma, Gerhard Navratil, Paolo Fogliaroni & Matt Duckham (Hrsg.), *Geoinfo 40: Extended Abstract Proceedings of the GIScience 2014*, 186–191. Wien: Department of Geodesy and Geoinformation, Vienna University of Technology.
15. Chen, Yiya & Gussenhoven, Carlos(2015), Shanghai Chinese, *Journal of the International Phonetic Association*, 45(3): 321–327.

## **Internetquellen**

- Wikipedia Seite für Junggrammatiker,
- Wikipedia Seite für Johann Heinrich August Leskien,
- Wikipedia Seite für Deutsches Textarchiv,
- Wikipedia Seite für Georg Wenker,
- Von Wörtern und Bäumen: Historische Sprachwissenschaft nach der quantitativen Wende,
- Wikipedia Seite für Hugo Schuchardt,
- Wikipedia Seite für Jules Gilli,
- Wikipedia Seite für Benrather Linie,
- Wikipedia Seite für Atlas linguistique de la France,
- Wikipedia Seite für Shanghai,
- Wikipedia Seite für Shanghainese,
- Wikipedia Seite für Wu-Chinese.