Computerlinguistische Experimente für die schweizerdeutsche Dialektlandschaft: Maschinelle Übersetzung und Dialektometrie

Scherrer, Yves

Computerlinguistische Experimente für die schweizerdeutsche Dialektlandschaft: Maschinelle Übersetzung und Dialektometrie

Yves Scherrer, Université de Genève

1. Einführung

Die hier präsentierte Arbeit soll zwei entgegengesetzte Felder der Sprachwissenschaft zusammenbringen: die Dialektologie und die Computerlinguistik. Obwohl sich diese beiden Forschungsrichtungen grundlegend in Bezug auf Studienobjekt und Methodologie unterscheiden, ergeben sich durch ihre Kombination interessante Fragestellungen, die ich in diesem Artikel vertiefen möchte.

Zum einen können computerlinguistische Hilfsmittel helfen, dialektologische Resultate intuitiver und interaktiver zu präsentieren. Es ist öfters festgestellt worden, dass beispielsweise Dialektwörterbücher und Sprachatlanten für Laien nicht einfach zugänglich sind; informatische Hilfsmittel können diesen Zugang ansprechender gestalten. Zum anderen soll untersucht werden, inwiefern die Eigenheiten von Dialektdaten eine Anpassung computerlinguistischer Methoden erfordern. Die Computerlinguistik arbeitet nämlich typischerweise mit Sprachdaten, die kaum nennenswerte interne Variation (in diatopischer und orthographischer Hinsicht) aufweisen.

Es gibt in der Computerlinguistik zwei grundlegend verschiedene Ansätze. Regelbasierte Systeme beruhen auf der expliziten Modellierung von Sprachphänomenen durch einen kompetenten Linguisten. Im Gegensatz dazu lernen statistische Systeme diese Regelmässigkeiten automatisch aus grossen Datenmengen. Obwohl statistische Systeme in den letzten Jahrzehnten an Bedeutung gewonnen haben, sprechen die verfügbaren Dialektdaten eher für regelbasierte Systeme. Textkorpora sind meist nicht in genügender Grösse und Qualität vorhanden, damit ein statistisches System die Unterschiede zwischen den einzelnen Dialekten eines Sprachareals lernen könnte. Hingegen sind wissenschaftlich aufbereitete Datenquellen (z.B. Wörterbücher, Grammatiken, Atlanten) dank jahrzehntelanger dialektologischer Forschung - zumindest für die Dialektregionen Europas - einfacher verfügbar. Um eine optimale Abdeckung und Präzision zu gewährleisten, sollten computerlinguistische Anwendungen diese Datenquellen nutzen können, was unter dem regelbasierten Ansatz bedeutend einfacher ist.

Das Ziel der vorliegenden Arbeit ist es, standarddeutschen Text automatisch in verschiedene schweizerdeutsche Dialekte umzuwandeln. Es handelt sich hierbei also um eine Instanz von maschineller Übersetzung zwischen nah verwandten Sprachvarietäten, wie sie beispielsweise für die verschiedenen romanischen Sprachen Spaniens (Corbí-Bellot, et al., 2005) oder zwischen Tschechisch und Slowakisch (Hajic, Homola, & Kubon, 2003) zum Einsatz kommt. Die Besonderheit meines Ansatzes liegt darin, dass die Zielsprache in Wirklichkeit aus einem Kontinuum von Zieldialekten

besteht. Es geht mir also nicht nur darum, die Unterschiede zwischen Standarddeutsch und dem Alemannischen zu formalisieren (Forst, 2002), sondern auch kleinräumigere Phänomene der Deutschschweizer Dialektlandschaft zu berücksichtigen.

Im Folgenden stelle ich verschiedene Aspekte dieses Übersetzungssystems vor. In Abschnitt 3 behandle ich die Wort-für-Wort-Übersetzung, wobei die areale Gültigkeit dieser Regeln durch Kartenmaterial aus den Sprachatlas der deutschen Schweiz [SDS] (Hotzenköcherle, Schläpfer, Trüb, & Zinsli, 1962-1997) bestimmt wird. In Abschnitt 4 gehe ich näher auf die Veränderungen der syntaktischen Struktur ein. Dafür werden ebenfalls Transformationsregeln erstellt, die aber auf den Daten des Syntaktischen Atlas der deutschen Schweiz [SADS] (Bucheli & Glaser, 2002) beruhen. Die für die Wort-Übersetzungsregeln notwendige Digitalisierung von SDS-Karten wird in Abschnitt 2 besprochen. Dieses digitalisierte Kartenmaterial kann daraufhin gewinnbringend für andere Zwecke eingesetzt werden; in Abschnitt 5 zeige ich als Illustration einige dialektometrische Untersuchungen.

2. Digitalisierung von SDS-Karten

Der Sprachatlas der deutschen Schweiz (Hotzenköcherle, Schläpfer, Trüb, & Zinsli, 1962-1997) besteht aus acht Bänden und enthält etwa 1500 handgezeichnete Karten zu den Bereichen Phonetik/Phonologie, Morphologie und Lexikon. Jede Karte stellt die verschiedenen dialektalen Realisierungen eines bestimmten sprachlichen Phänomens mittels Punktsymbolen dar. Die den Karten zugrundeliegenen Daten wurden zwischen 1939 und 1958 and 625 Ortspunkten erhoben. Abbildung 1 zeigt eine solche Karte.

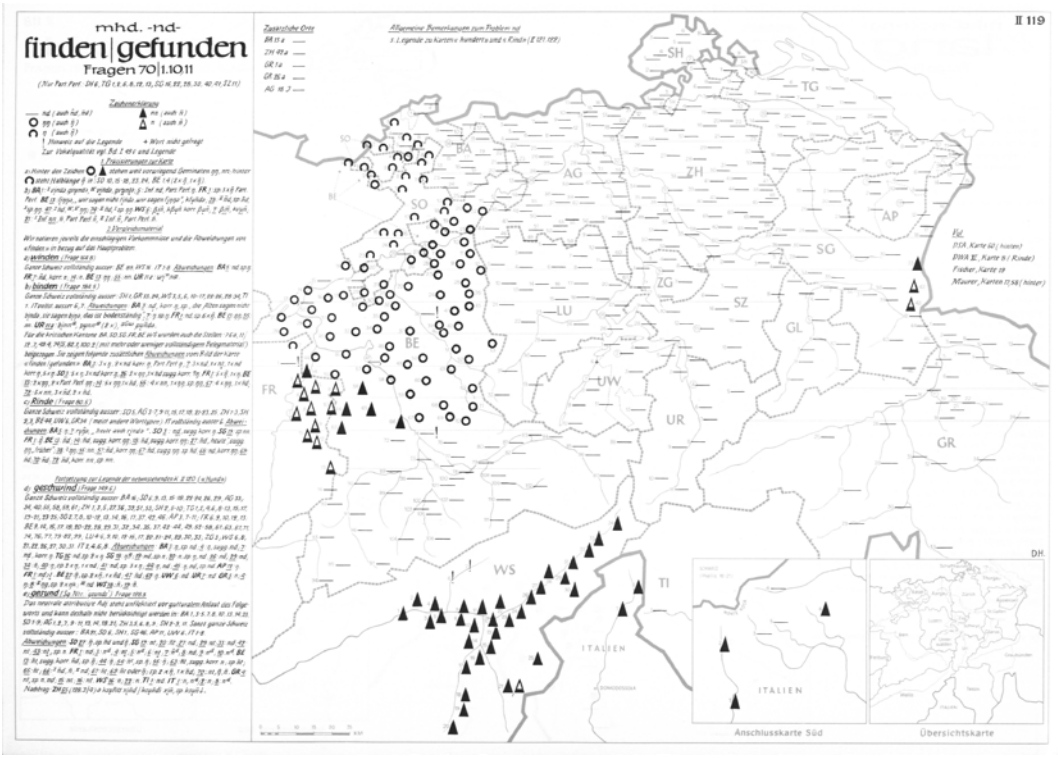

Abbildung 1: Originalkarte aus dem SDS (II/119)

Für die Bedürfnisse des Übersetzungssystems habe ich eine Teilmenge der SDS-Daten zur Digitalisierung ausgewählt, bestehend aus 59 Lautkarten, 110 Karten zur Morphologie und 27

Wortkarten. In einem ersten Schritt werden die Karten gescannt und mit Hilfe eines Geographischen Informationssystems (GIS) digitalisiert. Das Resultat dieses Schrittes ist in Abbildung 2 illustriert. Um diesen Prozess zu erleichtern, mussten einige Vereinfachungen vorgenommen werden. Erstens wurden Varianten, die in weniger als zehn Ortspunkten vorkamen, nicht berücksichtigt. Christen hat gezeigt, dass viele solcher kleinräumigen Varianten seit der SDS-Datenerhebung verdrängt worden sind (Christen, 1998). Zweitens werden diejenigen phonetischen Unterscheidungen zusammengefasst, die in der vom Übersetzungssystem verwendeten Dieth-Schreibung (Dieth, 1986) nicht ausgedrückt werden können.

Abbildung 2: Digitalisierte Punktkarte. *nd*-Werte werden mit grauen Quadraten dargestellt, *ng*-Werte mit schwarzen Dreiecken, *n*-Werte mit weissen Rauten und *nn*-Werte mit schwarzen Kreisen.

Die SDS-Karten, ob handgezeichnet oder digitalisiert, sind Punktkarten. Sie geben die an den Aufnahmeorten erhobenen Daten präzise wieder, liefern aber keine Angaben über die Dialekteigenschaften der umliegenden Orte. Diese Angaben können durch Interpolation geschätzt werden. Deshalb werden in einem zweiten Schritt die digitalisierten Punktkarten interpoliert, so dass für jede Variante eine Flächenkarte entsteht. Als Interpolationsmethode benutze ich die Kerndichteschätzung (Rumpf, Pickl, Elspaß, König, & Schmidt, 2009), die im Vergleich zu herkömmlichen Interpolationsmethoden eine wahrscheinlichkeitstheoretische Interpretation erlaubt und darüber hinaus weniger anfällig für Ausreisser ist, die sich in der Datensammlung oder der Datenaufbereitung einschleichen können. Ein Beispiel wird in Abbildung 3 gezeigt.

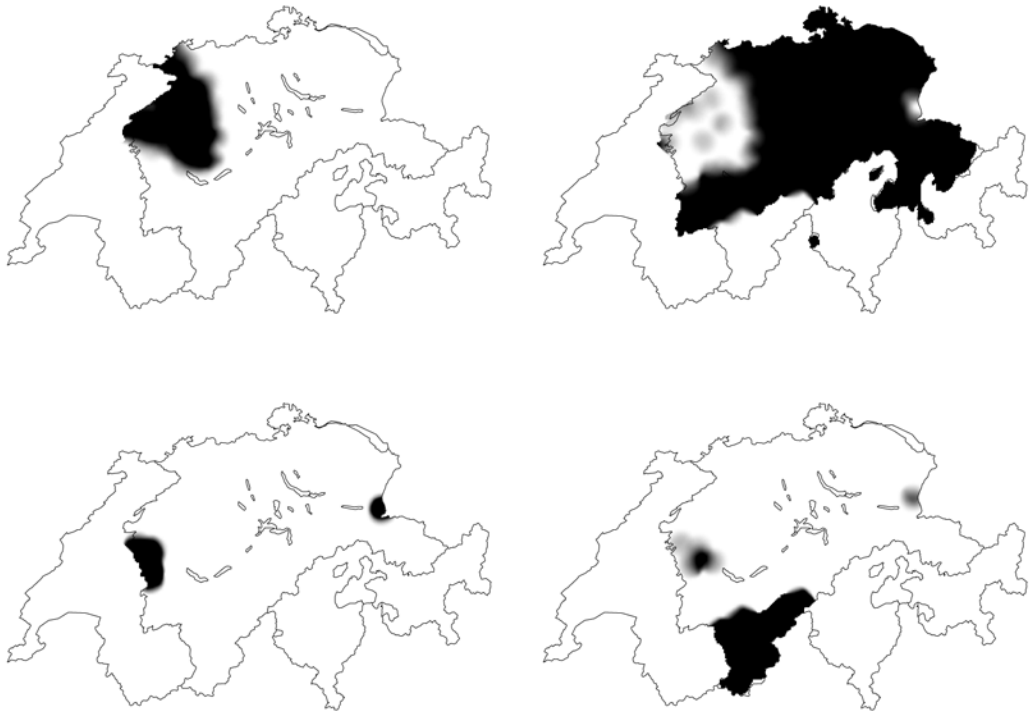

Abbildung 3: Interpolierte Flächenkarten zu den vier Varianten der Karte II/119: *ng* oben links, *nd* oben rechts, *n* unten links, *nn* unten rechts. Die Graustufen entsprechen Wahrscheinlichkeiten (schwarz: sehr wahrscheinlich, weiss: sehr unwahrscheinlich).

3. Wort-Übersetzung

Der Hauptbestandteil eines herkömmlichen Übersetzungssystems ist ein Wörterbuch, das die Wörter der Quellsprache explizit mit Wörtern der Zielsprache in Verbindung setzt. Solche Wörterbücher werden entweder manuell erstellt oder automatisch aus mehrsprachigen Textkorpora extrahiert (Koehn, 2010). Eine der grundlegenden Hypothesen der vorliegenden Arbeit ist es, dass für die Übersetzung von standarddeutschen Wörtern in die schweizerdeutschen Dialekte ein solches Wörterbuch nur in Ausnahmefällen notwendig ist. Aufgrund der nahen etymologischen Verwandtschaft können die meisten Dialektwörter durch phonetische, phonologische und morphologische Transformationen von den standarddeutschen Wörtern hergeleitet werden. Damit muss ein manuell erstelltes Wörterbuch nur die Informationen von etymologisch nicht verwandten Wortpaaren und von unregelmässig hergeleiteten Wortpaaren enthalten. Da auch Unterschiede in den Flexionsparadigmen zu erwarten sind, werden alle flektierten Formen für den jeweiligen Dialekt rekonstruiert.

Die Implementation der Wort-Übersetzungsregeln basiert auf dem Konzept von endlichen Transduktoren (Hopcroft, Motwani, & Ullman, 2006). Diese erlauben es, bestimmte Symbole einer Zeichenkette in Abhängigkeit des Kontextes zu verändern. In der Computerlinguistik werden endliche Transduktoren standardmässig im Bereich der phonologischen und morphologischen Analyse benutzt. In dieser Arbeit verwende ich das Toolkit XFST (Beesley & Karttunen, 2003), das die Erstellung von Transformationsregeln für linguistische Zwecke mittels endlicher Transduktoren erleichtert.

Phonetische Transformationsregeln

Wie oben erläutert, gehe ich davon aus, dass viele Wörter vorhersehbare und regelmässige phonetische Unterschiede zwischen Standarddeutsch und Schweizerdeutsch aufweisen, und dass phonetische Transformationsregeln genügen, um solche Wörter korrekt umzuwandeln.

Zum Beispiel wird die Graphemsequenz *nd* in intervokalischem Kontext (z.B. *finden*, *gestanden*) in den meisten schweizerdeutschen Dialekten beibehalten, während sie im Freiburger Dialekt zu *n* wird, im Walliser Dialekt zu *nn*, und im (nördlichen) Berner Dialekt zu *ng* (SDS II/119). Im XFST-Formalismus wird diese Regel folgendermassen formuliert:

```
define ndVoc [
        d -> [ d | g | n | 0 ]
|| Vowel n _ Vowel ];
```

Die Regel heisst *ndVoc* und wandelt ein *d* in ein *d*, *g* oder *n* um, oder löscht es (*0*). Der Bereich nach den beiden vertikalen Linien definiert den Anwendungskontext: das *d* wird nur umgewandelt, wenn links davon ein Vokal und ein *n* stehen, und rechts davon ein Vokal steht.

In dieser Formulierung wird allerdings nicht klar, welche Variante in welcher Region gültig ist. Dazu wird bei jeder Variante angegeben, welcher Wahrscheinlichkeitskarte sie entspricht. Dies geschieht mit Hilfe von sogenannten *flag diacritics* (Beesley, 1998), das sind Sonderzeichen, die von XFST nicht als linguistisches Material interpretiert werden und deshalb bei der Anwendung von darauffolgenden Regeln nicht stören. In diesen Sonderzeichen können Schlüssel-Wert-Paare gespeichert werden. Die aktualisierte Regel sieht so aus (der Lesbarkeit halber auf mehrere Zeilen verteilt):

```
define ndVoc [
        d -> [    d "@U.2-119.nd@" |
                  g "@U.2-119.ng@" |
                  n "@U.2-119.nn@" |
                  0 "@U.2-119.n@" ]
|| Vowel n _ Vowel ];
```

Flag diacritics werden durch Anführungszeichen und das Symbol @ definiert und bestehen aus einer Aktionsart (hier *U*, um einen neuen Wert hinzuzufügen), einem Schlüssel und einem Wert. In meinen Regeln entspricht der Schlüssel der SDS-Kartennummer (hier 2-119), und die Werte den jeweiligen Varianten. Damit können nach der Anwendung der Regel die entsprechenden Karten konsultiert werden (siehe unten). Regeln, bei denen jede Variante mit einem Verweis zu einer Wahrscheinlichkeitskarte versehen ist, nenne ich **georeferenzierte Regeln**. Von der Idee her entsprechen diese Regeln denjenigen, die Veith im Rahmen seiner generativen Dialektologie vorgeschlagen hat (Veith, 1982).

Wörterbucheinträge für nicht verwandte Wortpaare

Nicht alle Dialektwörter können allein aufgrund von phonetischen Regeln hergeleitet werden. Erstens gibt es häufig Unregelmässigkeiten in der phonetischen Entwicklung von hochfrequenten Wörtern. Zum Beispiel wird *und* im Berner Dialekt zu *u* reduziert, während die phonetischen Regeln

**ung* suggerieren. Zweitens gibt es Fälle, in denen komplett unterschiedliche Lexeme benutzt werden. Ein Beispiel dafür ist *immer*, das einer Vielzahl von Dialektwörtern entspricht (siehe unten). Für solche Fälle wird ein Wörterbuch erstellt, das aber nach denselben Prinzipien wie die phonetischen Regeln implementiert ist. Ein Beispiel ist hier wiedergegeben:

```
define immer [{immer} -> [
                {immer} "@U.6-026-immer.immer@" |
                {gäng} "@U.6-026-immer.gäng@" |
                {geng} "@U.6-026-immer.geng@" |
                {gi} "@U.6-026-immer.gi@" |
                {ging} "@U.6-026-immer.ging@" |
                {alewil} "@U.6-026-immer.alewil@" |
                {all} "@U.6-026-immer.all@" |
                {albig} "@U.6-026-immer.albig@" |
                {eisder} "@U.6-026-immer.eisder@" ]
];
```

Die geschweiften Klammern bedeuten, dass ganze Wörter umgewandelt werden, und nicht wie oben nur einzelne Symbole. Ausserdem kann bei diesen lexikalischen Regeln auf die Angabe des Anwendungskontextes verzichtet werden.

Flexionsaffixe

Die Umwandlung flektierter standarddeutscher Wortformen in flektierte schweizerdeutsche Wortformen geschieht in drei Schritten. Zuerst wird der standarddeutsche Wortstamm ermittelt und gegebenenfalls mittels lexikalischer Regeln an den Zieldialekt angepasst. Danach werden die dialektalen Flexionsaffixe generiert. Schliesslich werden die phonetischen Transformationsregeln auf die flektierten Formen angewandt.[1] Als Beispiel einer Flexionsregel soll hier die schwache Adjektivendung im Nominativ Singular dienen. Die standarddeutsche Endung *-e* (wie in *die schwarz**e** Katze*) wird in westlichen Dialekten zu *-i*, in östlichen Dialekten verschwindet sie. Die folgende Regel erfasst dieses Phänomen:

```
define adj-2-flex [
        ADJA [Nom | Acc] Sg Gender Degree Weak -> [
                0 "@U.3-254-adj-F-Sg.0@" |
                i "@U.3-254-adj-F-Sg.i@" ]
];
```

Diese Regel überführt eine Sequenz von morphosyntaktischen Features[2] in zwei alternative Endungen und deren *flag diacritics*. Die morphosyntaktischen Features werden vor Anwendung der

[1] Die oben erwähnte phonetische Regel zum intervokalischen *nd* illustriert, warum diese zwingend nach den Flexionsregeln angewandt werden: beim Wortstamm *find-* ist das *nd* nicht in intervokalischer Position, bei den flektierten Formen *finde* und *gfunde* (o.ä.) jedoch schon.

[2] Die Features in der obengenannten Regel besagen folgendes: attributives Adjektiv, Nominativ oder Akkusativ, Singular, beliebiges Genus, beliebiger Steigerungsgrad, schwache Deklination.

Flexionsregeln an den Wortstamm angehängt; dies ist die übliche Vorgehensweise bei Morphologiegeneratoren (Beesley & Karttunen, 2003).

Alle Regeln werden in Kaskade angewandt, so dass für ein gegebenes standarddeutsches Wort am Schluss eine Vielzahl dialektaler Wortformen entsteht. Jede dieser Wortformen enthält eine Reihe von Kartendeskriptoren in den *flag diacritics* (einen pro angewandte Regel). Am Schluss des Transformationsprozesses werden die entsprechenden Wahrscheinlichkeitskarten eingelesen und miteinander multipliziert, wodurch eine einzige Karte pro Dialektwort entsteht. In diesem Schritt werden geographisch inkompatible Herleitungen herausgefiltert - zum Beispiel ein Wortstamm in Basler Dialekt, der mit einer Endung im Walliser Dialekt kombiniert wurde.

Experimente

Die Abdeckung der Transformationsregeln für die Wort-Übersetzung ist in (Scherrer, 2011a) experimentell evaluiert worden. Ich fasse hier den Aufbau und die Resultate dieses Experiments zusammen.[3]

In einem ersten Schritt wird ein multidialektales morphologisches Wörterbuch erstellt. Dazu wird eine Liste standarddeutscher Wörter, lemmatisiert und morphosyntaktisch annotiert, den Transformationsregeln zugeführt. Diese generieren für jedes Quellwort eine Reihe von Dialektwörtern mit den zugehörigen Wahrscheinlichkeitskarten. Ein Eintrag im morphologischen Wörterbuch enthält demnach (i) ein Dialektwort, (ii) die Ausbreitungskarte, (iii) das zugehörige standarddeutsche Lemma, und (iv) die zugehörigen morphosyntaktischen Annotationen.

In einem zweiten Schritt wird gemessen, wieviele Wörter eines Dialektkorpus im morphologischen Wörterbuch vorhanden sind. Als Dialektkorpus verwende ich Texte aus der schweizerdeutschen Wikipedia[4] in fünf Dialekten: Baseldeutsch, Berndeutsch, Ostschweizer Dialekt, Walliserdeutsch und Zürichdeutsch. Die Texte wurden jeweils direkt von den Autoren mit der Dialektkategorie versehen. Pro Dialekt wurden 100 Sätze ausgewählt; diese enthalten etwa 1000 Wort-Types pro Dialekt.

Der Walliser Dialekt schneidet konsistent schlechter ab als die vier Mittellanddialekte. Für letztere werden zwischen 26% und 42% der Wort-Types und zwischen 39% und 57% der Wort-Tokens korrekt erkannt, für den Walliser Dialekt 16% der Types und 28% der Tokens. Die noch relativ häufig auftretenden Fehler sind auf vier hauptsächliche Ursachen zurückzuführen:

1. Unterschiede in der Verschriftlichung. Beispielsweise generiert das System die zürichdeutsche Verbform *bestaat* 'besteht' laut den Dieth-Richtlinien, während der Wikipedia-Text die standardnähere Schreibung *bestaht* enthält. Hier handelt es sich also lediglich um eine orthographische Abweichung.
2. Mangelhafte Abdeckung bezüglich lexikalischer Ausnahmen. Das Übersetzungssystem wandelt *Kirche* in *Chirche* um, im betreffenden Zürcher Dialekt ist aber *Chile* gebräuchlich. Es handelt sich dabei um eine phonetische Unregelmässigkeit, die durch einen separaten Lexikoneintrag *Kirche* → *Chile* ergänzt werden müsste.

[3] Das in (Scherrer, 2011a) beschriebene System basiert auf einer anderen Implementation der Transformationsregeln, was die leicht unterschiedlichen Resultate erklärt. Der Versuchsaufbau bleibt aber gleich.

[4] *http://als.wikipedia.org*

3. Diachronischer Wandel. Die SDS-Daten entsprechen teilweise nicht mehr dem heutigen Sprachgebrauch. Zum Beispiel werden für das Wort *zeigt* im Ostschweizer Dialekt die Varianten *zeigt*, *zäägt* und *zaagt* generiert; diese sind (zumindest für den als Referenz benutzten Stadtsanktgaller Dialekt) veraltet und eher marginal, während das Wikipedia-Korpus die heute gebräuchlichste Variante *zaigt* enthält.
4. Falsche Lokalisierung. Gemäss SDS-Daten dominiert die Öffnung der Vokale *i*, *u* und *ü* (zu *e*, *o* und *ö*) im gesamten Bernbiet. Im Gegensatz zum Aargauer und Luzerner Dialekt wird diese Tatsache aber in der Regel nicht verschriftlicht, so dass im Wikipedia-Text beispielsweise *Süde* und nicht *Söde* 'Süden' vorkommt. Wenn das Übersetzungssystem also die *ö*-Variante vorschlägt, die Referenz aber die *ü*-Variante enthält, wird das als falsch gezählt. Werden solche Lokalisierungsfehler ausser Acht gelassen und die Varianten aller beliebigen Dialekte als korrekt angesehen, dann verbessern sich die Resultate folgendermassen: 42% bis 48% korrekte Types (Wallis 25%), und 59% bis 65% korrekte Tokens (Wallis 43%).

Dieses Experiment bestätigt, dass ein respektabler Prozentsatz an Dialektwörtern mittels phonetischer Regeln von den standarddeutschen Wörtern hergeleitet werden kann, es zeigt aber auch deutlich die Grenzen dieses Ansatzes auf. Während die Abdeckung der lexikalischen Regeln verbessert werden müsste, gibt es für die anderen Fehlerquellen keine einfache Lösung. Die Regeln müssten einerseits fehlertoleranter werden, um besser mit der diachronischen und orthographischen Variation umgehen zu können, dürfen andererseits aber nicht an Präzision einbüssen, wenn die Vorteile der multidialektalen Übersetzung bewahrt werden sollen.

4. Umwandlung der syntaktischen Struktur

Die syntaktischen Unterschiede zwischen Standarddeutsch und den schweizerdeutschen Dialekten sind annäherungsweise auf zwei Faktoren zurückzuführen, einen soziolinguistischen und einen dialektologischen. Einige Unterschiede sind charakteristisch für die hauptsächlich gesprochene Gebrauchsweise der schweizerdeutschen Dialekte; sie weisen kaum interdialektale Variation auf und kommen auch in anderen gesprochenen Varietäten des Deutschen vor. Als Beispiele können hier das Fehlen des Präteritums und des Genitivs angeführt werden, sowie den Gebrauch von Artikeln vor Personennamen. Andere Unterschiede sind dialektologischer Art; sie sind auf bestimmte Untergruppen der schweizerdeutschen Dialekte beschränkt und erscheinen meist nicht ausserhalb der alemannischen Dialekte. Letztere Unterschiede sind in den vergangenen Jahren im Rahmen des Forschungsprojekts *Dialektsyntax des Schweizerdeutschen* (Glaser, 2003) untersucht worden. Beispiele dieses Typs sind die Wortreihenfolge in Verbclustern, die präpositionale Dativmarkierung, die Flexion von prädikativen und koprädikativen Adjektiven, oder die Einleitung finaler Infinitivsätze.

Zuallererst wollte ich bestimmen, wie häufig die obengenannten syntaktischen Konstruktionen im Sprachgebrauch zu erwarten sind. Dafür habe ich die standarddeutsche Baumbank TIGER (Brants, Dipper, Hansen, Lezius, & Smith, 2002), bestehend aus 40'000 syntaktisch annotierten Sätzen, nach den relevanten Konstruktionen durchsucht. Tabelle 1 zeigt die Resultate dieser Vorstudie. Mehr als ein Drittel aller Sätze enthalten eine Nominalphrase im Genitiv oder ein Verb im Präteritum; die entsprechenden Transformationsregeln sind daher unverzichtbar. Eine Reihe weiterer Phänomene treten in 7% bis 14% aller TIGER-Sätze auf. Eine dritte Gruppe von Konstruktionen ist eher selten und erscheint in weniger als 2% der Sätze. Bei diesen Zahlen muss allerdings beachtet werden, dass das

TIGER-Korpus aus Zeitungstext besteht und deshalb nicht repräsentativ für die Nutzung der schweizerdeutschen Dialekte ist.

Total Sätze in TIGER	**40'000**	**100%**
Genitiv	15'351	38%
Präteritum	13'439	34%
Eigennamen	5410	14%
Relativpronomen	4619	12%
Verbcluster	3246	8%
Prädikative Adjektive	2784	7%
Präpositionale Dativmarkierung	2708	7%
Finalsätze	629	2%
W-Wort-Verdoppelung	478	1%
Artikelverdoppelung/-umstellung	61	<1%
Pronomenfolgen	6	<1%

Tabelle 1: Häufigkeit der für die Dialektübersetzung relevanten syntaktischen Konstruktionen

Analog zu den im vorhergehenden Abschnitt vorgestellten Wortübersetzungsregeln wurden Transformationsregeln für syntaktische Konstruktionen entwickelt. Diese Regeln suchen bestimmte Strukturen in einem syntaktisch annotierten Text und verändern diese. Die syntaktische Annotation des Quelltextes basiert dabei auf dependenzgrammatischen Grundsätzen (Buchholz & Marsi, 2006). Wie die Wortübersetzungsregeln sind die Syntaxregeln georeferenziert. Sie beruhen auf den für den SADS erhobenen Daten (Bucheli & Glaser, 2002), aus denen ebenfalls durch Interpolation Flächenkarten erstellt werden. Das resultierende Übersetzungssystem enthält 29 Regeln mit 65 dialektalen Varianten, die 14 syntaktische Konstruktionen abdecken.

Die Genauigkeit dieser Transformationsregeln wurde ermittelt, indem für jede Regel 100 Sätze mit der entsprechenden Struktur aus dem TIGER-Korpus extrahiert und transformiert wurden. Die Ausgabe des Regelsystems wurde manuell auf Fehler überprüft. Die Genauigkeit der Regeln variiert zwischen 85% und 100%; nur für zwei Regelvarianten liegt sie darunter. Details zur Implementation der Syntaxregeln, zur Evaluation und zur Fehleranalyse finden sich in (Scherrer, 2011b).

5. Dialektometrie

Für die georeferenzierten Transformationsregeln wurden fast 200 Karten aus dem SDS gescannt und digitalisiert. Dieses Material kann nun für andere Zwecke weiterverwendet werden, beispielsweise für dialektometrische Studien. Die Dialektometrie hat sich in den letzten Jahrzehnten als ein Feld der Sprachwissenschaft etabliert, das die Anwendung statistischer und mathematischer Methoden in der Dialektforschung untersucht. Auch als quantitative Dialektologie bekannt, gilt das Interesse der Dialektometrie vorwiegend der regionalen Verteilung von Dialektähnlichkeiten. Als Datenbasis dienen dazu aggregierte Daten, wie sie beispielsweise in Sprachatlanten vorkommen. Die untersuchten Dialekte können dabei in Klassen eingeteilt und entsprechend visualisiert werden (Goebl, 2002).

In diesem Artikel wird auf zwei weit verbreitete Analyse- und Visualisierungsmethoden eingegangen: die hierarchische Clusteranalyse und die multidimensionale Skalierung (MDS). Kelle hat die schweizerdeutsche Dialektlandschaft mit Hilfe der Clusteranalyse dialektometrisch untersucht und mit früheren, nicht dialektometrischen Gliederungsversuchen verglichen (Kelle, 2001). Er benutzte

dafür Material aus 170 SDS-Karten, deren Ortsnetz allerdings aus arbeitstechnischen Gründen auf 100 Punkte reduziert wurde. Die hier verwendeten digitalisierten SDS-Karten decken knapp 200 Phänomene ab; das Ortsnetz umfasst alle SDS-Punkte (ca. 600) mit Ausnahme der acht italienischen Orte.

Clusteranalyse

Bei der hierarchischen Clusteranalyse werden die einzelnen Ortspunkte aufgrund ihrer dialektalen Ähnlichkeit schrittweise zusammengefasst, so dass eine Baumstruktur, ein sogenanntes Dendrogramm, entsteht (Bacher, Pöge, & Wenzig, 2010) (Nerbonne & Siedle, 2005). Es sind verschiedene Algorithmen der Clusteranalyse entwickelt worden, die sich in der Berechnungsmethode der Ähnlichkeit zweier Teilbäume unterscheiden. Dementsprechend können die Resultate der Clusteranalyse je nach benutztem Algorithmus relativ stark variieren.

Kelle benutzt den Complete-Linkage-Algorithmus, der in der neueren Dialektometrieforschung allerdings kaum mehr angewandt wird. Zu Vergleichszwecken zeige ich in Abbildung 4 eine Clusterkarte, die auf diesem Algorithmus beruht. Abbildung 5 und Abbildung 6 stellen die Resultate mit den beliebteren Weighted-Average- und Ward-Algorithmen dar.[5] In allen Abbildungen werden zehn Cluster, den zehn grössten sich nicht überlappenden Teilbäumen des Dendrogramms entsprechend, farblich unterschieden. Diese Anzahl Cluster wurde arbiträr festgesetzt.

In der Gegenüberstellung dieser drei Karten können einige stabile und einige fluktuierende Dialektgrenzen ausgemacht werden. Stabil sind die Grenze zwischen dem Wallis und Bern, die Brünig-Napf-Linie zwischen Bern und Luzern, das nordwestliche Dialektareal um Basel, das zentralschweizerische Dialektareal, sowie die Abgrenzung zwischen dem Aargau und der Region Zürich. Fluktuierend sind die Grenze zwischen dem Berner Oberland/Freiburg und dem nördlichen Teil des Kantons Bern, die Grenze zwischen Zürich und der Ostschweiz, die Abgrenzung des Zürcher Dialektgebiets nach Südosten hin, sowie die Zweiteilung Graubündens in die Churerrheintaler und Walser-Dialekte. Diese Dialektgrenzen entsprechen ziemlich präzise den früher vorgeschlagenen Gliederungen, beispielsweise derjenigen von Hotzenköcherle (Hotzenköcherle, 1984).

Im Vergleich mit der Arbeit von Kelle treten auch einige Parallelen hervor, obwohl er seine Analyse bei sechs Dialektgruppen beendet. Die Abgrenzung der Walliser Dialekte ergibt sich in beiden Studien sehr klar. Die bei Kelle deutlich sichtbare Abgrenzung der italienischen Südorte konnte hier nicht reproduziert werden, weil die entsprechenden Daten nicht digitalisiert worden sind. Die restlichen vier von Kelle gefundenen Dialektregionen erscheinen ebenfalls weitgehend in meinen Bildern: Der Kanton Bern ist zweigeteilt und wird nach Osten hin durch die Brünig-Napf-Reuss-Linie (Haas, 2000) begrenzt. Die südöstlichen Dialekte werden auf einer von Luzern zum Walensee reichenden Linie von den nordöstlichen abgegrenzt.

[5] Die Clusterkarten in Farbe können unter *http://archive-ouverte.unige.ch/unige:22904* konsultiert werden.

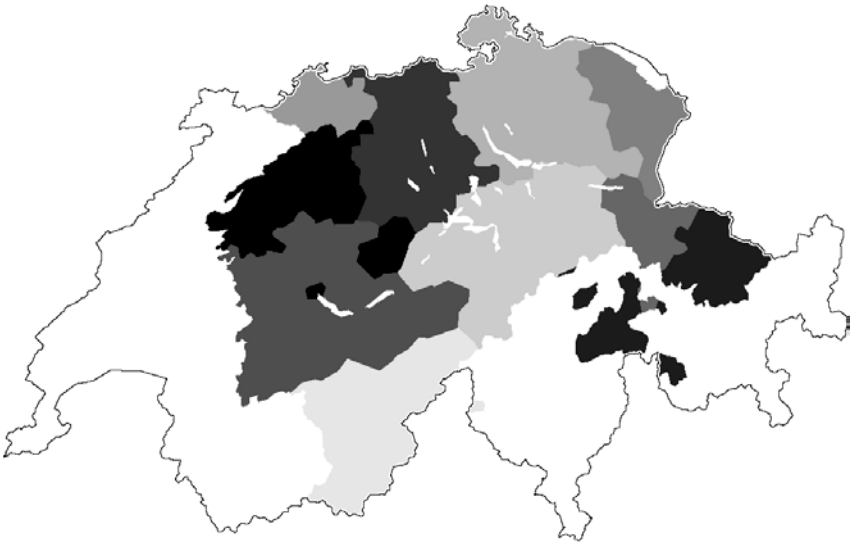

Abbildung 4: Clusterkarte basierend auf dem Complete-Linkage-Algorithmus

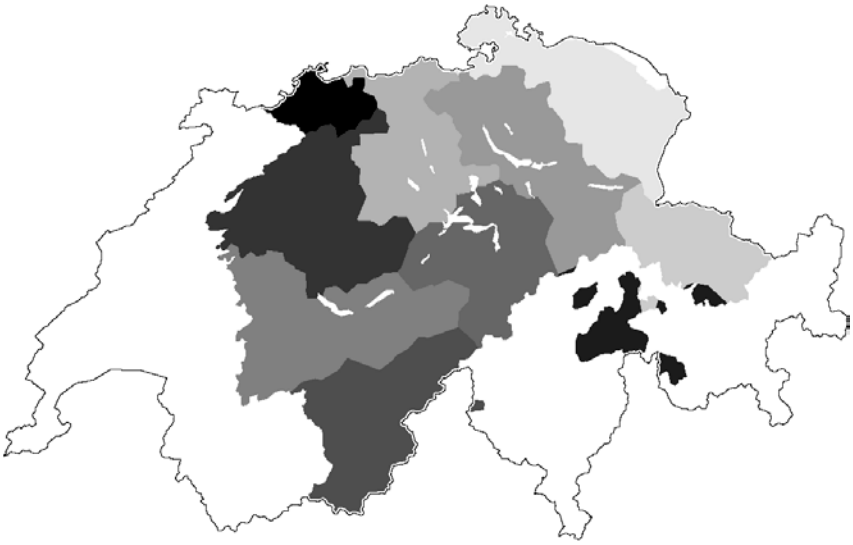

Abbildung 5: Clusterkarte basierend auf dem Weighted-Average-Linkage-Algorithmus

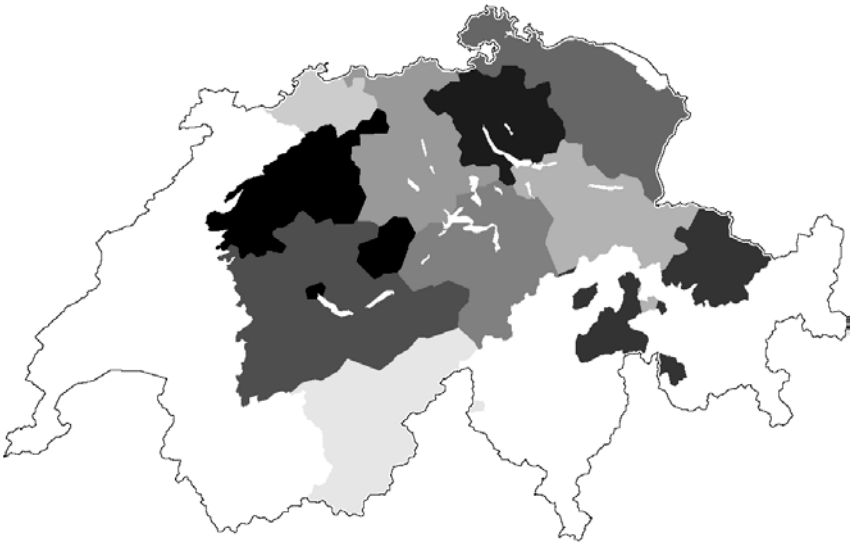

Abbildung 6: Clusterkarte basierend auf dem Ward-Algorithmus

Multidimensionale Skalierung

Die Clusterkarten suggerieren scharfe Grenzen zwischen den Dialektgebieten, was nicht unbedingt der Realität des meist graduellen Dialektwandels entspricht. Darüber hinaus haftet der Clusteranalyse der Ruf der Instabilität an: geringe Unterschiede in den Daten können grosse Veränderungen in der Gruppenbildung nach sich ziehen. Ebenso hat die Wahl des Algorithmus einen nicht zu unterschätzenden Einfluss auf das Endresultat, wie in den obigen Abbildungen illustriert worden ist. Daher wurde als alternative Datenanalysemethode die multidimensionale Skalierung vorgeschlagen (Embleton, 1993).

Um die Ähnlichkeiten zwischen zwei Ortspunkten zu bestimmen, wird in der Dialektometrie auf eine Vielzahl von Merkmalen zurückgegriffen. Im hier vorliegenden Fall entspricht jedes Merkmal einem sprachlichen Phänomen, das in einer SDS-Karte dargestellt ist. Das bedeutet, dass die linguistische Ähnlichkeit zwischen zwei Ortspunkten in einem vieldimensionalen Raum gemessen wird: jede Dimension entspricht einem Merkmal. Um diese Ähnlichkeiten intuitiv darstellen zu können, werden sie mit Hilfe von MDS umgerechnet und auf drei Dimensionen reduziert. Jeder dieser drei Dimensionen wird als Graustufenkarte dargestellt (Abbildung 7, Abbildung 8, Abbildung 9). Darüber hinaus können die drei Dimensionen in einer Farbkarte zusammengefasst werden, indem die drei Werte als Komponenten im RGB-Farbraum interpretiert werden; diese Farbkarten werden "Regenbogenkarte" genannt (Nerbonne & Siedle, 2005).[6]

Die MDS-Karten zeigen ein nuancierteres Bild als die oben diskutierte Clusterkarten, geben aber sowohl die Ost-West-Staffelung (in der ersten Dimension) als auch die Nord-Süd-Staffelung (in der zweiten Dimension) erstaunlich klar wieder. In der dritten Dimension werden mit der Region Berner Oberland/Freiburg und der Ostschweiz weitere Dialektgebiete identifiziert und abgegrenzt.

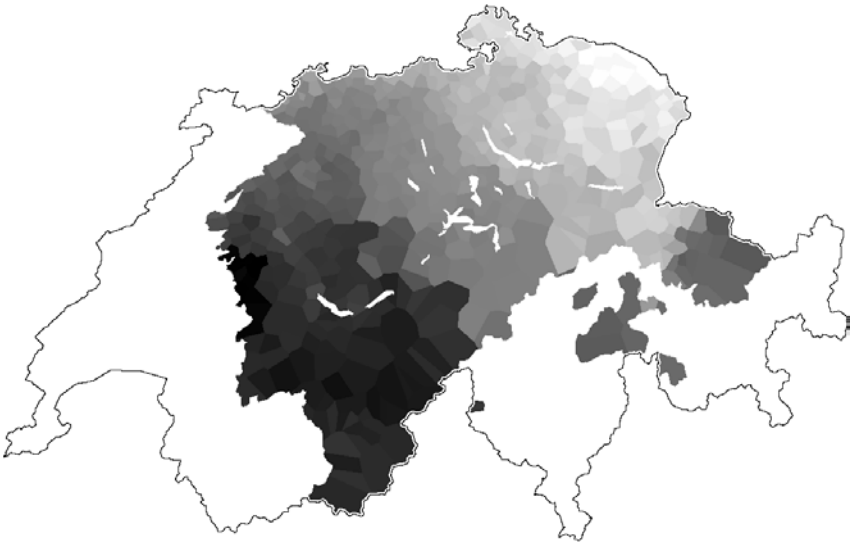

Abbildung 7: MDS-Karte für die erste Dimension.

[6] Aus drucktechnischen Gründen können hier keine Farbkarten wiedergegeben werden. Die Regenbogenkarte kann im elektronischen Anhang unter *http://archive-ouverte.unige.ch/unige:22904* betrachtet werden.

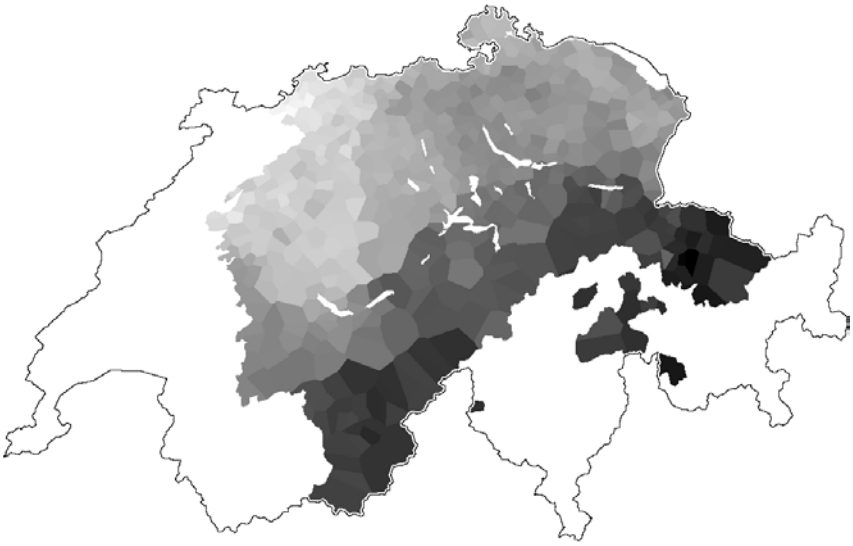

Abbildung 8: MDS-Karte für die zweite Dimension.

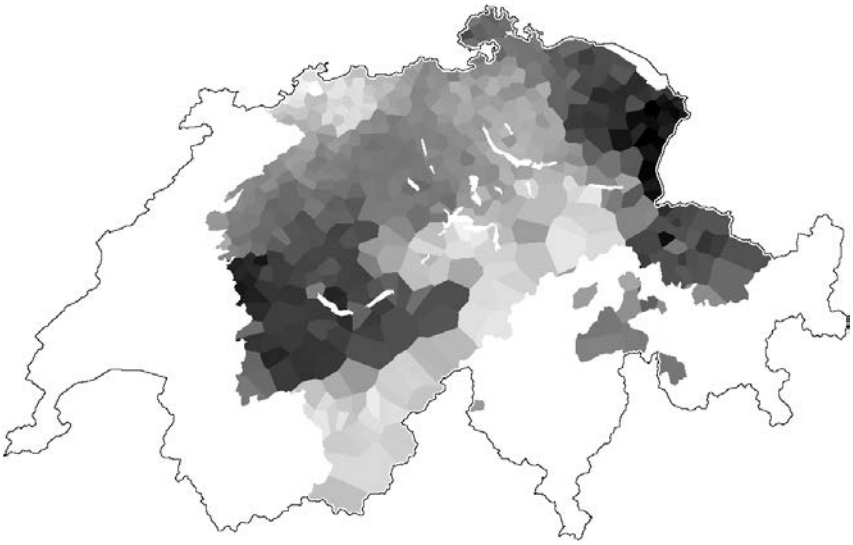

Abbildung 9: MDS-Karte für die dritte Dimension.

6. Ausblick

In diesem Artikel wurden verschiedene Aspekte der maschinellen Übersetzung für schweizerdeutsche Dialekte beleuchtet. Dabei habe ich das Konzept der georeferenzierten Transformationsregel eingeführt und für die Nutzung von phonetisch-graphemischen Transformationen argumentiert. In ersten Experimenten hat sich dieser Ansatz als vielversprechend erwiesen, obgleich die Abdeckung der Regeln noch stark erhöht werden sollte. In Zukunft möchte ich diese Regeln bidirektional auszugestalten, um damit Dialekttexte ins Standarddeutsche übertragen zu können.

Darüber hinaus habe ich einige dialektometrische Versuche gezeigt, die durch die Digitalisierung von SDS-Karten ermöglicht worden sind. Die verschiedenen Visualisierungsmethoden erlauben es, die wichtigsten Dialektgebiete der deutschen Schweiz zu identifizieren und die Gebietsgrenzen zu

charakterisieren. Dieser dialektometrische Ansatz scheint mir grosses Potenzial für künftige Arbeiten aufzuweisen.

7. Literaturverzeichnis

Bacher, J., Pöge, A., & Wenzig, K. (2010). *Clusteranalyse.* München: Oldenbourg.

Beesley, K. R. (1998). Constraining separated morphotactic dependencies in finite-state grammars. *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing.* Ankara.

Beesley, K. R., & Karttunen, L. (2003). *Finite State Morphology.* Stanford: CSLI Publications.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER Treebank. *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, (S. 24-41). Sozopol.

Bucheli, C., & Glaser, E. (2002). The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems. In S. Barbiers, L. Cornips, & S. van der Kleij, *Syntactic Microvariation.* Amsterdam: Meertens Institute Electronic Publications in Linguistics.

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on Multilingual Dependency Parsing. *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*, (S. 149-164). New York City.

Christen, H. (1998). *Dialekt im Alltag: eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten.* Tübingen: Niemeyer.

Corbí-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., et al. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT 2005)*, (S. 79-86). Budapest.

Dieth, E. (1986). *Schwyzertütschi Dialäktschrift* (2. Ausg.). (C. Schmid-Cadalbert, Hrsg.) Aarau: Sauerländer.

Embleton, S. (1993). Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project. In R. Köhler, & B. B. Rieger (Hrsg.), *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics (QUALICO)* (S. 267-276). Dordrecht/Boston/London: Kluwer Academic Publishers.

Forst, M. (2002). La traduction automatique dans le cadre formel de la LFG - Un système de traduction entre l'allemand standard et le zurichois. In *Publications du Centre de traduction littéraire de Lausanne (CTL)* (Bd. 41). Université de Lausanne.

Glaser, E. (2003). Schweizerdeutsche Syntax. Phänomene und Entwicklungen. In *Gömmer MiGro? Veränderungen und Entwicklungen im heutigen SchweizerDeutschen* (S. 39-66). Freiburg, Schweiz: Universitätsverlag Freiburg.

Goebl, H. (2002). Dialektometrie. In G. Altmann, D. Bagheri, H. Goebl, R. Köhler, & C. Prün (Hrsg.), *Einführung in die quantitative Lexikologie* (S. 179-219). Göttingen: Peust & Gutschmidt.

Haas, W. (2000). Die deutschsprachige Schweiz. In H. Bickel, & R. Schläpfer, *Die viersprachige Schweiz* (S. 71-160). Aarau: Sauerländer.

Hajic, J., Homola, P., & Kubon, V. (2003). A simple multilingual machine translation system. *Proceedings of the Machine Translation Summit XI*, (S. 157-164). New Orleans.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2006). *Introduction to Automata Theory, Languages and Computation* (3. Ausg.). Boston: Pearson Addison-Wesley.

Hotzenköcherle, R. (1984). *Die Sprachlandschaften der deutschen Schweiz.* (N. Bigler, & R. Schläpfer, Hrsg.) Aarau: Sauerländer.

Hotzenköcherle, R., Schläpfer, R., Trüb, R., & Zinsli, P. (Hrsg.). (1962-1997). *Sprachatlas der deutschen Schweiz.* Bern: Francke.

Kelle, B. (2001). Zur Typologie der Dialekte in der deutschsprachigen Schweiz: Ein dialektometrischer Versuch. *Dialectologia et Geolinguistica , 9*, 9-34.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Nerbonne, J., & Siedle, C. (2005). Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik , 72* (5), 129-147.

Rumpf, J., Pickl, S., Elspaß, S., König, W., & Schmidt, V. (2009). Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik , 76* (3), 280-308.

Scherrer, Y. (2011a). Morphology generation for Swiss German Dialects. In C. Mahlow, & M. Piotrowski (Hrsg.), *Systems and Frameworks for Computational Morphology, Second International Workshop (SFCM 2011)* (S. 130-140). Berlin, Heidelberg: Springer.

Scherrer, Y. (2011b). Syntactic transformations for Swiss German dialects. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, (S. 30-38). Edinburgh.

Veith, W. H. (1982). Theorieansätze einer generativen Dialektologie. In W. Besch, U. Knoop, W. Putschke, & H. E. Wiegand (Hrsg.), *Dialektologie - Ein Handbuch zur deutschen und allgemeinen Dialektforschung* (S. 277-295). Berlin, New York: De Gruyter.