Universität Zürich[UZH]

Masterarbeit
zur Erlangung des akademischen Grades
Master of Arts
der Philosophischen Fakultät der Universität Zürich

Morphologieanalyse und standarddeutsche Lemmatisierung für schweizerdeutsche Alltagstexte mit gewichteten Transduktoren

Verfasser: Reto Flavio Baumgartner
Matrikel-Nr: 09-706-409

Referent: Prof. Dr. Martin Volk

Betreuer: Dr. Simon Clematide

Institut für Computerlinguistik

Abgabedatum: 7.4.2016

Zusammenfassung

In den letzten Jahren hat die schriftliche Verwendung von Schweizerdeutsch zugenommen. Durch die Absenz einer standardisierten Orthographie ist es aber für viele texttechnologische Werkzeuge immer noch unzugänglich. Mit einem System aus gewichteten Transduktoren können schweizerdeutsche Wortformen nun erkannt werden und ein eigens definiertes morphologisches Tagset und standarddeutsche Lemmata ermöglichen eine weitere Verarbeitung. Angesichts der grossen Variation von Schreibungen kann eine Erkennungsrate von 90% auf ausgewählten Texten als Erfolg bezeichnet werden. Die Gewichte stellen sicher, dass für die meisten Wörter die besten Analysen bevorzugt sind. Neben dem Analysesystem ist auch ein morphologisch annotiertes Korpus entstanden, das zum Aufbau einer Nachbehandlung der Analyse verwendet werden kann.

Abstract

In the last years, there was an increase of written use of the Swiss German language. Due to the absence of a standardised orthography, however, many tools for text technology cannot be used to process it. A system based on weighted transducers is now able to recognise Swiss German word forms and further processing is enabled with a specially defined morphological tagset and lemmas in Standard German. With the high variability of possible spellings, a recognition rate of over 90% on certain texts is a success. The weights make sure that for most words, the best analyses are preferred. Besides the system for analysis, the development of a morphologically annotated corpus will help building a post-processing tool for the analyses.

Danksagung

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mich bei meiner Masterarbeit unterstützt und motiviert haben.

Besonderen Dank verdient auch Dr. Simon Clematide, der mich während des Studiums für das Thema begeistert hat und sich als Betreuer angeboten hat. Mit wertvollen Hinweisen stellte er meine Motivation während der ganzen Arbeit sicher. Ebenfalls bedanken möchte ich mich bei Prof. Dr. Martin Volk, der sich als Referent zur Verfügung gestellt hat.

Auch meiner Mitstudentin Janina Fontanive bin ich dankbar für ihr aufmerksames Gegenlesen und Verbesserungsvorschläge.

Danken möchte ich auch meiner Familie, die mich während des Studiums und meiner Masterarbeit unterstützt hat und die Arbeit Korrektur gelesen hat.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

deu	Deutsch (Standarddeutsch) (nach ISO 639-3)
gsw	Schweizerdeutsch (nach ISO 639-3)
mhd.	Mittelhochdeutsch
HK	Häufigkeitsklasse
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
OCR	Optical Character Recognition
STTS	Stuttgart-Tübingen-Tagset

1 Einleitung

In den letzten Jahren hat sich die schriftliche Verwendung von schweizerdeutschen Dialekten in der Schweiz ausgeweitet und aus dem gelegentlichen Gebrauch von geschriebenem Schweizerdeutsch wurde ein alltägliches Mittel zur Kommunikation auf Kanälen wie SMS, Chat und E-Mail und auch auf Postkarten. Im Zuge des Phänomens der neuen Schriftlichkeit wird mehr geschrieben als früher und mit der vermehrten privaten schriftlichen Kommunikation hat sich die Mundart verschriftlicht (Rumjanzewa 2013). Für die Wahl des Schweizerdeutschen in diesen Kommunikationsmitteln spricht vor allem die empfundene Nähe, die für verbal geprägte Kommunikation gewünscht ist (Ueberwasser 2013).

Neben einem neuen Spannungsverhältnis zwischen Dialekt und Standardsprache mit Folgen in der Bildungspolitik (u. a. Volksinitiativen für Mundart in Kindergärten) führt diese Verschriftlichung zu mehr schweizerdeutschen Texten in digitaler Form. Die leichtere Verfügbarkeit dialektaler Texte ebenso wie Wünsche von Konsumenten wie die Spracherkennung (vgl. Bolzern 2015) rücken Schweizerdeutsch in den Fokus der Sprachtechnologie. Einer der zentralen Punkte dabei ist die Analyse der schweizerdeutschen Wortformen in Texten.

1.1 Motivation

Das Fehlen eines Standards und die damit verbundene Breite an Varianten stellen jedoch für viele Anwendungen ein Problem dar. Gleichzeitig verstärkt die Zahl der Varianten den Mangel an ausreichenden Trainingsdaten und macht rein statistische Verfahren impraktikabel. Mit einem schweizerdeutschen Dialekt als Muttersprache ist es mir ein Ziel, diese Breite der Varianten behandeln zu können und die Texte für weitere Anwendungen der Sprachtechnologie zu öffnen.

Tabelle 1 gibt einen Überblick darüber, welche Schreibweisen für ‚Jahr' im Schweizerdeutschen möglich sind. Ob jemand mit ⟨a⟩ oder mit ⟨o⟩ schreibt, ist von der Lautung im jeweiligen Dialekt abhängig. Auch das ausbleibende ⟨r⟩ im Appenzell ist ein typisches Dialektmerkmal. Ob dagegen ein ⟨h⟩ oder ein Doppelvokal

verwendet wird, ist von den Präferenzen des Schreibers abhängig.

Dialekt	Form	Kommentar
Aargau	Johr, Joor, Jahr	z. T. Verdumpfung
Appenzell	Joh	*r*-Ausfall
Basel	Joor	Verdumpfung
Bern	Jahr	
Zürich	Jahr, Jaar	

Tabelle 1: Schreibweisen für ‚Jahr' aus dem NOAH-Korpus nach Dialekt geordnet.

Neben den verschiedenen Lautungen und Schreibungen derselben besteht auch das Problem, dass mit Wortgrenzen nicht einheitlich umgegangen wird. Hollenstein und Aepli (2014) bringen dazu das Beispiel *bruchtmese* ‚braucht man sie'. Die einzelnen Teile davon können auch als separate Wörter geschrieben werden wie in *brucht me se* oder nur zum Teil verbunden wie in *bruchtme se*. Für dieses Problem braucht es eine konsistente linguistische Beschreibung und auch eine Möglichkeit zur Erkennung solcher Formen.

Mit Lemmata in standarddeutscher Sprache und einer auf die Eigenheiten des Schweizerdeutschen abgestimmten Auszeichnung der morphologischen und syntaktischen Kategorien soll Schweizerdeutsch für etablierte Werkzeuge der Sprachtechnologie zugänglich werden.

Vergleichbar mit dem Verhältnis den Varietäten Standarddeutsch und Schweizerdeutsch ist auch dasjenige zwischen Rumantsch Grischun und den nahe verwandten, traditionellen Schriftidiomen. Ein früherer Versuch mit den Idiomen bei der Erstellung eines Morphologieanalysesystems für Rumantsch Grischun (R. Baumgartner et al. 2013) zeigte, dass einfache Finite-State-Maschinen nicht für die Erkennung solcher Varianten ausreichen. Die traditionellen Schriftidiome des Rätoromanischen in der Schweiz und die dazu geschaffene Dachsprache Rumantsch Grischun verwenden zwar ein ähnliches Vokabular, unterscheiden sich aber in der Orthographie teils erheblich. Ohne eine Gewichtung der Transduktoren erwies sich das Finden der Balance zwischen genügender Abdeckung und der Vermeidung von Übergenerierung als schwierig. Von Seiten der Morphologie müssten die Idiome eher als eigene Sprachen betrachtet werden. So weist Rumantsch Grischun beispielsweise deutlich weniger Verbformen als das Idiom Vallader auf und Wortgrenzen werden von den Idiomen (vor allem bei den Präpositionen) unterschiedlich gehandhabt. Das Verhältnis zwischen den Idiomen und Rumantsch Grischun ist also mit demjenigen zwischen den schweizerdeutschen Dialekten und Standard-

deutsch vergleichbar. Die Konsequenz daraus ist einerseits die Verwendung von Gewichten um die Analysen zu ordnen und mit der Übergenerierung umgehen zu können und andererseits der eigenständige Aufbau der Wortformen, wobei Stämme aus der Dachsprache übernommen werden können.

1.2 Forschungsfragen

Aus diesen Überlegungen folgen die Hauptfragen, welche diese Arbeit behandeln soll:

1. Welches Darstellungsformat bei der linguistischen Anayse eignet sich für die Morphologie?
2. Lässt sich die erhöhte Mehrdeutigkeit infolge der Nichtstandardisierung der Rechtschreibung durch gewichtete Transduktoren besser behandeln?

Aus diesen Fragen eröffnen sich folgende weitergehenden Fragen:

3. Können bestehende Werkzeuge für die standarddeutsche Sprache beim Aufbau eines Morphologieanalysesystems genutzt werden?
4. Welchen Nutzen bringt ein manuell erstellter Kern für die Abdeckung der Wortformen?
5. Welchen Nutzen bringt eine separate Behandlung für die Lautformen der jeweiligen Dialekte?
6. Wie fein soll die Auszeichnung der morphologischen Kategorien bei der Analyse sein? Wie soll das Tagset gestaltet werden? (Welche Informationsmenge erweist sich bei der Analyse als angemessen?)
7. Wie lässt sich das Problem der ungenau bestimmten Wortgrenzen behandeln? Sollen Zusammenschreibungen als morphologisches Phänomen behandelt werden und entsprechende morphologische und syntaktische Kategorien definiert werden; oder ist es besser, Zusammenschreibungen zu trennen und entsprechend zu analysieren? Diese Frage ist auch im Zusammenhang mit der Tokenisierung zu betrachten.

Im Zusammenhang mit diesen Fragen entstehen ein Tagset *STTS.gsw* für die morphologische Auszeichnung für Schweizerdeutsch und ein Morphologieanalysesystem *Taggswort* (von ‚Tagge das Wort'), das schweizerdeutschen Wortformen solche Tags zuweist. Als weitere Ressource entsteht im Rahmen der Entwicklung von *STTS.gsw*

und *Taggswort* ein mit Lemmata und Morphologie annotiertes Textkorpus.

1.3 Definitionen

Der Begriff *Schweizerdeutsch* bezieht sich in dieser Arbeit auf die in der Deutschschweiz gesprochenen alemannischen Dialekte. Neben Dialekten des Deutschen sind auch Französisch im Westen, Italienisch und Rätoromanisch im Südosten angestammte Sprachen in der Schweiz und offizielle Landessprachen. Abbildung 1 zeigt das Verbreitungsgebiet des Schweizerdeutschen in der Schweiz im Kontext der oberdeutschen Mundarten, zu denen es als Untergruppe des Alemannischen gehört.

Die schweizerdeutschen Dialekte (gsw) unterscheiden sich untereinander vor allem durch ihre lautlichen Realisierungen, es gibt aber auch Unterschiede in der Lexik und in der Syntax. Trotz dieser Unterschiede werden sie in der ganzen Deutschschweiz als Umgangssprache genutzt und Verständnisprobleme sind selten.

Davon abzugrenzen sind Sprachformen des Deutschen wie Standarddeutsch (deu). Dieses ist in seiner Schriftform durch die normierte Orthographie geprägt, während es sich im Allgemeinen durch den neuhochdeutschen Lautstand und seine Konservativität in der Grammatik auszeichnet. Trotz seiner Normierung weist Standarddeutsch verschiedene nationale und überregionale Ausprägungen auf. Eine davon ist *Schweizer Hochdeutsch* oder *Schriftdeutsch*, das sich vor allem durch sein Vokabular, in geringerem Masse aber auch durch orthographische und grammatische Eigenheiten auszeichnet, die aber immer noch zur Standardsprache gezählt werden können. Diese Arbeit verfolgt den Ansatz einer breiten Definition der Standardsprache und auf eine Unterteilung in nationale oder regionale Varietäten kann verzichtet werden. Mit der Fokussierung auf die Schriftsprache ist in dieser Arbeit mit Standarddeutsch die normierte Schriftsprache in ihrer geschriebenen Form gemeint.

Weitere Bezeichnungen für deutsche Sprachvarietäten sind *Mittelhochdeutsch* und *Neuhochdeutsch*. Unter Mittelhochdeutsch versteht man üblicherweise die deutsche Sprache in ihrer Ausprägung zwischen 1150 und 1350 (siehe Christen et al. 2012, S. 18), während Neuhochdeutsch die heute gebräuchliche Sprache ist und dem auch Standarddeutsch angehört. Beide Begriffe werden hier primär im Hinblick auf das Lautsystem der jeweiligen Varietät verwendet.

Die beiden Hauptaufgaben des erstellten Systems sind die *morphologische Analyse* und die *Lemmatisierung*. Formell ist *morphologische Analyse* die Zuordnung von

Abbildung 1: Verbreitungsgebiet des Schweizerdeutschen (schwarz) im oberdeutschen Sprachraum (grau). Gebiete des Schweizerdeutschen im Ausland sind dunkelgrau eingefärbt. (Bearbeitet vom Verfasser nach der Vorlage von Wikimedia Commons (2015))

Markierungen der Belegungen grammatischer Kategorien zu den Wörtern einer Sprache. Bei der Verbform *isch* ‚ist' würde beispielsweise die dritte Person, der Numerus Singular und das Tempus Präsens identifiziert. Unter *Lemmatisierung* versteht man die Bezeichnung der verschiedenen Realisierungsformen eines Wortes mit einer lexikographischen Standardform. Üblicherweise entspricht dies der Grundform. Als Lemma für *isch* würde also der Infinitiv ‚sein' gewählt. Bei nominalen Wortarten nimmt man meist den Nominativ Singular als Lemma. Mangels einer einheitlichen Norm für Schweizerdeutsch sind das in *Taggswort* standarddeutsche *Lemmata*.

Verwandt aber nicht deckungsgleich damit sind die Begriffe der *Normalisierung* und *Glossierung*. Während bei der *Normalisierung* das Sprachmaterial einer Norm angepasst wird, versteht man unter *Glossen* in der Regel die wörtliche Übersetzung der Wörter in eine andere Varietät.

Für die Erklärung der variablem Wortgrenzen muss auch der Begriff *Klitikon* definiert werden. *Klitika* sind schwach bis nicht betonte Wörter, die sich lautlich an ein anderes Wort anlehnen. Im Schweizerdeutschen werden sie uneinheitlich, manchmal mit anderen Wörtern zusammen und manchmal als unabhängige Wörter geschrieben.

1.4 Aufbau der Arbeit

Nach der Einführung ins Thema durch dieses Kapitel folgt in Kapitel 2 ein Überblick über die bisherigen Auseinandersetzung in der Computerlinguistik mit Schweizerdeutsch sowie mit morphologischer Analyse nicht-standardsprachlicher Varietäten im Allgemeinen.

In Kapitel 3 wird auf die schweizerdeutschen Dialekte eingegangen, das linguistische Vorgehen erklärt und ein Schema für die Annotation festgelegt. Kapitel 4 behandelt dagegen das technische Vorgehen, die Umsetzung als Finite-State-System und wie die Gewichte berechnet werden.

Kapitel 5 beinhaltet eine quantitative Evaluation des Systems und enthält auch qualitative Analysen zur Veranschaulichung der Resultate und Probleme Während sich der erste Teil des Kapitels der Anzahl mit dem System erkennbarer Formen kümmert, widmet sich der zweite Teil der Gewichtung der Analysevorschläge.

2 Stand der Forschung

2.1 Sprachtechnologische Arbeiten für schweizerdeutsche Dialekte

Dadurch, dass die Sprecherzahl des Schweizerdeutschen deutlich kleiner ist als diejenige des Standarddeutschen, sind sprachliche Ressourcen für das Schweizerdeutsche spärlicher als für das Standarddeutsche. Dies verstärkt sich noch dadurch, dass in der Deutschschweiz die meisten Texte in der Standardsprache verfasst werden. Dass die Dialekttexte keinem etablierten Standard folgen, stellt ein Hindernis bei der Erstellung von Werkzeugen für die Verarbeitung dar. Das Fehlen des Standards erschwert zum Beispiel die Digitalisierung existierender Texte, da für *optical character recognition* (OCR) umfangreiche und unter Umständen dialektspezifische Wörterlisten nötig wären. Die Schwierigkeiten, grosse Textsammlungen zu erstellen, wirken sich auch auf die Verfügbarkeit korpusbasierter Werkzeuge aus. Mit der gesteigerten Verwendung des Schweizerdeutschen in der digitalen Kommunikation vereinfacht sich aber der Zugang zu Sprachdaten.

2.1.1 Korpora

Zu den wenigen schriftlichen Korpora für Schweizerdeutsch gehört das *Swiss SMS Corpus* von Stark et al. (2009–2015). Das *Swiss SMS Corpus* besteht aus insgesamt 25 947 Kurznachrichten und umfasst rund 500 000 Tokens in diversen Sprachen. Von diesen stammen 275 000 Tokens aus schweizerdeutschen Dialekten. Die einzelnen Tokens sind mit Glossen, automatischer Lemmatisierung und Wortartenannotation versehen.

Als weitere Ressource ist *NOAH's Corpus of Swiss German Dialects* von Hollenstein und Aepli (2014) zu erwähnen. Dieses umfasst ursprünglich 73 616 manuell mit Wortarten annotierte Tokens und wurde in Release 2.0 auf 115 000 Tokens erweitert. Das Korpus umfasst fünf Teile: *Blick* mit Texten der Dialektausgabe Gratiszeitung *Blick am Abend*, *Blogs* mit diversen Internetblogs auf Schweizer-

deutsch, *Schobinger* mit einem Kriminalroman, *Swatch* mit einem Geschäftsbericht des gleichnamigen Unternehmens und *Wiki* mit Texten aus der alemannischen Wikipedia.

2.1.2 Maschinelle Sprachverarbeitung

Trotz der scheinbar uninteressanten Rolle als lokal begrenzte Sprachform und der dürftigen Datenlage gibt es für Schweizerdeutsch eine kleine Zahl von Anwendungen. Dieser Abschnitt gibt einen Überblick darüber.

Sprachidentifikation

Ein Dialektidentifikationssystem[1] von Scherrer und Rambow (2010) verwendet mangels Trainingskorpora Daten aus dem *Sprachatlas der deutschen Schweiz* von H. Baumgartner et al. (1962-2003) für die Extraktion eines geeigneten Featuresets für die Klassifikation. Auf einer Karte der Schweiz werden die Gebiete eingefärbt, für die ein Sprachmuster typisch ist.

Bilinguale Lexikonerstellung

Scherrer (2007) untersuchte den Einsatz von Ersetzungsregeln für die Erstellung eines bilingualen Lexikons. Die grosse Ähnlichkeit zwischen dem Vokabular des Standarddeutschen und des Schweizerdeutschen ist dabei zentral.

Übersetzung

Ansätze zur maschinellen Übersetzung haben aktuell das Standarddeutsche als Quellen- und das Schweizerdeutsche als Zielsprache. Bei der Wahl dieser Übersetzungsrichtung spielt bestimmt eine Rolle, dass die Quellsprache durch die Standardisierung einfacher zu verarbeiten ist. Auf der Zielseite besteht die Freiheit, einen Standard zu definieren, da diese Seite nicht weiter verarbeitet werden muss.

Aufbauend auf der Methode mit Ersetzungsregeln nutzen Scherrer und Rambow (2010) die Möglichkeit, standarddeutsche Wörter nach Morphologie und Lemma zu analysieren um damit die gewünschten schweizerdeutschen Formen zu generieren (Mehr dazu unter 2.1.3) . Lexikale und sekundär auch phonetische Regeln

[1] Interaktiv verfügbar auf `http://latlntic.unige.ch/~scherrey/prod/dialectID.html` (aufgerufen am 29. März 2016)

übernehmen hierbei die Übersetzung des Lemmas. Bei diesem Schritt und bei der anschliessenden Formengenerierung wird nach Zieldialekten unterschieden. Regeln zur syntaktischen Umstellung enthält dieser wortbasierte Ansatz keine, auch wenn die Autoren diesbezüglich von einer möglichen Verbesserung ausgehen.

Die maschinelle Übersetzung vom Schweizerdeutschen ausgehend gestaltet sich durch die Fülle an nebeneinander verwendeten Formen komplizierter und der Profit durch ein Morphologieanalysewerkzeug wäre gross. In Zusammenhang mit dem Anteil der Deutschsprachigen, die Schweizerdeutsch nicht beherrschen, und im Gedanken an die Bevölkerung der anderssprachigen Landesteile ist aber gerade ein System für diese Übersetzungsrichtung erstrebenswert.

Wortartentagging

Durch das Korpus von Hollenstein und Aepli (2014) existiert Trainingsmaterial für maschinelle Wortartenannotation. Es ist ein Tagging-Modell für den *BTagger* (Gesmundo und Samardžic 2012) verfügbar, das auf diesem Korpus trainiert ist und eine Genauigkeit von 90,62% erreicht.

Parsing

Scherrer und Rambow (2010) skizzieren ein mögliches Vorgehen zum Bau eines syntaktischen Parsers für Schweizerdeutsch. Dabei könnten sowohl Teile der Dialektidentifikation für die Wörter als Erkenntnisse von der Entwicklung maschineller Übersetzung verwendet werden. Die Schwierigkeit ist hier ebenfalls das Fehlen entsprechender Daten für die Entwicklung und für die Evaluation.

In Zusammenhang mit dem Dependenzparser für Schweizerdeutsch von Klaper (2014) ist inzwischen ein Teil des NOAH-Korpus mit gut über 10 000 Tokens mit ungelabelten Dependenzannotationen, welches für das Training verwendet wurde, verfügbar.

2.1.3 Morphologische Analyse

Der wahrscheinlich wichtigste Ansatz für schweizerdeutsche Morphologie in der Sprachtechnologie bisher ist ein System zur Morphologiegenerierung[2] von Scherrer (2011). Mit Hilfe von Ersetzungen ist es ihm weitgehend gelungen, ausgehend

[2]Interaktiv verfügbar auf `http://latlntic.unige.ch/~scherrey/prod/dialect.html` (aufgerufen am 15. Februar 2016)

vom standarddeutschen Lemma die schweizerdeutsche Wortform für fünf ausgewählte Dialektgruppen zu generieren. In einem Testverfahren, welches diese Regeln in Kombination mit einem Lexikon aus möglichen Analyseformen – somit als bidirektionales System – auf schweizerdeutsche Texte anwendet, konnte je nach Dialekt eine korrekte Analyse für 25% bis 45% der Types und für 44% bis 65% der Tokens erreicht werden. Der Fokus auf Formengenerierung macht allerdings ein solches System für die morphologische Analyse ungeeignet, wie auch Scherrer bemerkt (Scherrer 2011, S. 138 – 139). Die bei der Generierung gewünschte Homogenität der Formen und der Schreibweise lassen sich nicht mit der grossen Abdeckung verschiedener Formen vereinbaren, die bei einem Analysewerkzeug erwünscht sind. Sowohl das Generierungswerkzeug als auch die Testdaten folgen einer relativ strengen lautnahen Schreibweise. Bereits kleine Abweichungen von diesem Schema führen dazu, dass solche Wörter nicht erkannt werden.

2.2 Morphologische Analyse für nicht-standardisierte Sprachen

Endliche Transduktoren wurden auch schon von anderen Autoren für die morphologische Analyse nicht-standisierter Sprache eingesetzt. In den Folgenden Beispielen verfügen die Sprachen zwar über einen Standard, dieser wird jedoch nicht konsequent verwendet, da es dafür keine Tradition gibt. Mit Hilfe der Analyse mit diesen Standards nahe verwandter Dialekte und einer möglichen Normalisierung sollen die Ressourcen für Sprachtechnologie in diesen Sprachen erweitert werden.

Ein Projekt, in dem Morphologieanalyse auf nicht-standardisierte Schriftsprache angewandt wird, ist Normalisierung für südliche Quechua-Varietäten von Rios und Mamani (2014). Anlass dazu ist das Vorhandensein sprachtechnologischer Ressourcen für einen Schriftstandard, während für andere Orthographien kaum Werkzeuge existieren. Vergleichbar mit der Situation in der Deutschschweiz gibt es für die südlichen Quechua-Varietäten Vorschläge zu Schriftstandards, denen jedoch nur ein Teil des Sprachmaterials folgt. Für die morphologische Analyse wenden die Autoren eine Kaskade von Transduktoren an, wobei derjenige am Anfang dem Standard am nächsten ist und jeder weitere als Rückgriffssystem fungiert. Die anschliessende Disambiguierung ist der Grundstein für die Generierung des normalisierten Textes.

Auch Hulden et al. (2011) befassen sich mit der Normalisierung von dialektalen Texten. Ihr Artikel vergleicht zwei Ansätze dazu, wie man dialektale Formen des Baskischen an den Standard annähern kann, um sie mit gängigen Werkzeugen

der Sprachtechnologie verarbeiten zu können. Die Unterschiede zwischen den Varietäten sind dabei vor allem lexikalisch und morphophonologisch, während viele Wörter im Standard und in den Varietäten die gleiche Form tragen. Beide Ansätze verwenden Ersetzungsregeln, die aus parallelen Daten extrahiert sind.

3 Linguistische Analyse

Dieses Kapitel beinhaltet eine vertiefte Beschreibung des Schweizerdeutschen. Aus den daraus gewonnenen Erkenntnissen werden ein Schema zur morphologischen Auszeichnung und Regeln zur Lemmatisierung formuliert.

3.1 Schweizerdeutsch

Schweizerdeutsche Dialekte werden primär in Niederalemannisch um Basel, Hochalemannisch im Mittelland und Höchstalemannisch am weitesten im Süden eingeteilt. Zusätzlich gibt es eine Unterscheidung zwischen West und Ost, die sich im Vokabular und in den Formen niederschlägt. Viele dieser Isoglossen entlaufen entlang der Brünig-Napf-Reuss-Linie, an der sich beispielsweise auch Unterschiede im Brauchtum abzeichnen (Christen et al. 2012, S. 29 – 30).

Die schweizerdeutschen Dialekte stehen in der Deutschschweiz in einem Diglossieverhältnis mit dem Standarddeutschen. Während sie prägend für die mündliche Kommunikation sind, wird die schriftliche Kommunikation vom Standarddeutschen beherrscht und nur ein kleiner Anteil findet auf Schweizerdeutsch statt. Meist ist dies in der privaten Kommunikation der Fall und Siebenhaar und Wyler (1997, S. 10) sehen dieses Phänomen vor allem für jüngere Leute als typisch.

Mit der Diglossie geht das Fehlen eines Kontinuums zwischen Dialekt und Standardvarietät einher. Diese Trennung der Varietäten wird aufrechterhalten und Mischformen werden bewusst vermieden (Siebenhaar und Wyler 1997, S. 14).

In der Dialektliteratur verzeichnen besonders diejenigen Gattungen einen Zuwachs, die auf einen effektvollen mündlichen Vortrag ausgerichtet sind und weniger Wert auf die Reinheit des Dialekts legen (Christen et al. 2012, S. 25). „Einen immer wichtigeren Stellenwert nimmt der geschriebene Dialekt in der informellen Schriftlichkeit von meist kurzen Texten ein, bei denen die Nähe zu den Adressaten eine herausragende Rolle spielt. So werden Kartengrüsse, persönliche Briefe, Schreibzüge in der Chat- oder SMS-Kommunikation, Einträge auf Internetplattformen, aber

auch Kontaktanzeigen in den Printmedien gerne im Dialekt geschrieben“ (Christen et al. 2012, S. 25). Auch hier steht nicht eine möglichst lautgetreue Verschriftlichung im Zentrum. Als wichtiges Kriterium sehen Christen et al. (2012, S. 25), dass das Geschriebene wie mündliche Kommunikation aufgefasst werden soll.

Während bis vor kurzem SMS eine wichtige Domäne für geschriebenes Schweizerdeutsch darstellte (dazu auch das *Swiss SMS Corpus*), so nimmt der Gebrauch von Smartphone-Textkommunikations-Applikationen wie *WhatsApp*[1] in den letzten Jahren zu. Mit dem Wegfall von Textlängenbeschränkungen und der Verbesserung von Eingabemethoden könnte man eine standardnähere Sprache erwarten. Dem wirkt aber in der Regel die hohe Schreibgeschwindigkeit entgegen, da *WhatsApp* eher wie ein Chat funktioniert (Dürscheid und Frick 2014).

3.1.1 Verhältnis zum Standarddeutschen

Schweizerdeutsche Dialekte und Standarddeutsch zeigen vor allem bei der Syntax und beim Vokabular mehr Gemeinsamkeiten als mit anderen westgermanischen Sprachen wie Englisch oder Niederländisch (siehe Siebenhaar und Wyler 1997, S. 33). Im Lautsystem fällt auf, dass die schweizerdeutschen Vokale ab der Zeit des Mittelhochdeutschen eine andere Entwicklung durchgemacht haben als die standarddeutschen. Das Vokalsystem steht dem des Mittelhochdeutschen nahe (Siebenhaar und Wyler 1997, S. 37).

In der Flexion zeigen sich hauptsächlich Unterschiede bei den Zeitformen und den Fällen. Das Präteritum fehlt bei den Verben vollständig, was eine Bildung des Plusquamperfekts analog zum Standarddeutschen verunmöglicht. Das stärkste Unterscheidungsmerkmal bei der Deklination ist das Fehlen des Genitivs, aber auch zwischen Akkusativ und Nominativ kann die Abgrenzung schwierig werden. Eindeutig unterscheidbar sind Nominativ und Akkusativ jedoch bei den Personalpronomina (vgl. Siebenhaar und Wyler 1997, S. 37). Die Syntax des Schweizerdeutschen ist mündlich geprägt, was durchaus auch auf umgangssprachliches Deutsch ausserhalb der Schweiz zutrifft (Siebenhaar und Wyler 1997, S. 37).

3.1.2 Ansätze für Schriftstandards

Bei der privaten Kommunikation bedienen sich die Schreiber in der Regel ihrer individuellen Orthographie (Siebenhaar und Wyler 1997, S. 10). Daneben existieren

[1] `https://www.whatsapp.com/` (aufgerufen am 15. Februar 2016)

aber Ansätze einer Systematisierung der Schreibungen, die eher in der Mundartliteratur und Sprachwissenschaft gebräuchlich sind. Es handelt sich dabei nicht um Standardsprachen, sondern um Richtlinien, wie die Laute des individuellen Dialekts verschriftlicht werden sollen.

Dieth-Schreibung

Die Schreibung nach Erwin Dieth (1986) ist der wahrscheinlich wichtigste Ansatz für eine geregelte Dialektschreibung. Ihr Ziel ist es, die Lautung der Ortsdialekte möglichst genau abzubilden. Begründet wird dies damit, dass auch Sprecher anderer Dialekte einen Text vorlesen können sollen, was bei standardnahen Schreibungen nicht möglich wäre (Dieth 1986, S. 14). Dabei soll aber auf sinnentstellende Assimilationen verzichtet werden, und gewisse Diakritika, wie Gravis für geschlossene Vokale, können weggelassen werden, wodurch eine *weite* Schreibung erreicht wird. Bei langen Vokalen wird aber eine konsequente Doppelschreibung verlangt. Weiter typisch ist ⟨scht⟩ für [ʃt] im Wortinneren (z. B. in *Schwöschter* ‚Schwester', gegenüber ⟨st⟩ am Wortanfang), sowie der Gebrauch von ⟨v⟩, ⟨ck⟩ und ⟨tz⟩ wie im Standarddeutschen (z. B. *voll* ‚voll', *Stuck* ‚Stück', *Netz* ‚Netz'). Der Gebrauch von ⟨h⟩ hingegen ist nur bei entsprechendem Laut erlaubt (*Huus* [huːs] ‚Haus' vs. *faare* [fɑːrə] ‚fahren').

Assimilationen und Zusammenzüge zwischen Wörtern sollen nur eingeschränkt in die Schrift übernommen werden, das heisst, dass die einzelnen Wörter ein möglichst einheitliches Schriftbild zeigen sollen. Erlaubt ist die Zusammenschreibung zwischen Verbformen und den Pronomina ‚es' und ‚wir' und zwischen Präpositionen und Artikeln in den Formen, die kein *d* (von ‚die') mehr enthalten (Dieth 1986, S. 43 – 46).

Bärndütschi Schrybwys

Einen anderen Ansatz als Dieth verfolgt Werner Marti mit der *Bärndütschen Schrybwys* 1985. In Hinblick auf die existierende Dialektliteratur und auf den Umstand, dass die meisten Leser in Standarddeutsch geübt sind, lässt er standardnahe Schreibungen zu. Konsequent lautgetreue Schreibungen sieht Marti (1985a, S. 24 – 25) gar als Hindernis für ein flüssiges Lesen.

Wichtige Unterschiede zur Schreibung von Dieth sind, dass die geschlossenen Vokale markiert (statt der offenen) und dass auch die Vokallängen nicht besonders gekennzeichnet werden ausser bei möglichen Verständnisschwierigkeiten. Standard-

sprachliches ⟨h⟩ zur Markierung langer Vokale darf übernommen werden. Nur in Zweifelsfällen ist die Verdoppelung von Vokalbuchstaben für die Markierung einer langen Aussprache vorgeschrieben (Marti 1985a, S. 35 – 45). Entsprechend ist Marti auch bei der Gemination der Konsonanten weniger konsequent.

Beim Zusammenzug von Präpositionen mit Artikeln ist Marti noch restriktiver als Dieth (1986), der eine gewisse Wahl offen lässt. Das Pronomen ‚es' kann als ⟨'s⟩ an finite Verbformen angehängt werden.

Beiden Schreibweisen gemeinsam ist ⟨y⟩ für [i(ː)], was eine lange Tradition hinter sich hat und in Familien- und Ortsnamen wie *Wyss* oder *Schwyz* allgegenwärtig ist. Gerade bei Dieth muss diese Tradition als besonders wichtig betrachtet worden sein, da der Gebrauch dieses Buchstabens sich nicht gut ins System mit Verdoppelungen und Diakritika einfügt.

Grammatiken

Zu den Ansätzen für Schriftstandards gibt es auch eine Fülle von Grammatiken, welche die jeweiligen Lokalgrammatiken beschreiben. Für die Formen wurden im vorliegenden Projekt eine Grammatik für Baseldeutsch (Suter 1992), eine für Berndeutsch (Marti 1985b) und eine für Zürichdeutsch (Weber und Bund Schwyzertütsch 1948) konsultiert.

3.1.3 Charakterisierung des Sprachmaterials

Im alltäglichen Gebrauch verfolgen die meisten Schreiber bei der Wahl ihrer Schreibweisen pragmatische Ziele. In Kurzmitteilungen ist Knappheit ein Mittel zur Verhinderung, dass Mitteilungen mit entsprechenden Folgekosten aufgeteilt werden. Ausserdem gibt es Schreiber, die offen erklären, dass sie beim Schreiben Merkmale, die für ihre lokalen Dialekte typisch sind, unterdrücken, um die Verständlichkeit zu erhöhen (siehe Ueberwasser 2013). Diese Verflachung steht in Kontrast zum Grundgedanken der Schreibung nach Dieth, die für den Ausdruck ortsspezifischer Züge die Möglichkeit bietet und so den Verlust der Diversität verhindern will.

Konventionen aus dem Standarddeutschen werden gerne befolgt, jedoch gibt es – vor allem bei den Vokalen – bewusste Abweichungen zur Markierung bestimmter Dialektmerkmale (siehe Ueberwasser 2013).

Auch die bei der Entwicklung von *Taggswort* verwendeten Daten aus dem NOAH-

Korpus folgen nicht konsequent den Anweisungen von Dieth oder Marti.

Beispiel 3.1 aus dem Teil *Blick* zeigt eine Schreibung, die dem Standarddeutschen sehr nahe steht. Nach Dieth müsste *da* als ⟨daa⟩ und *zumindest* als ⟨zumindescht⟩ geschrieben werden. Obwohl der Satz eher auf einen Lautstand wie in der Gegend um Zürich hinweist, ist die Schreibung mit der *Bärndütschen Schrybwys* vergleichbar.

(3.1) Da sind sich zumindest d Lüt in Kanada sicher .
Da sind sich zumindest die Leute in Kanada sicher .

Schreibweisen, die völlig von den Traditionen abweichen, finden sich in den *Blogs*. In Beispiel 3.2 fallen darunter *leidr* und *ish*, wo der Schwundvokal vor ⟨r⟩ nicht geschrieben ist, beziehungsweise /ʃ/ durch ⟨sh⟩ dargestellt wird. Von Dieth empfohlene Schreibweisen wären dagegen ⟨läider⟩ und ⟨isch⟩. Die Vokallänge ist nur im betonten Wort *uu* markiert.

(3.2) leidr muessi momentan da na id Schuel und es ish uu
Leider muss ich momentan hier noch in die Schule und es ist sehr
stressig .
stressig .

Die bei den literarischen Texten im Teil *Schobinger* verwendete Schreibung folgt dagegen den Anleitungen von Dieth. Typisch ist hier auch die phonetische Schreibung von Fremdwörtern wie *Grã-scheff* in Beispiel 3.3, die Markierung von Langvokalen durch Verdoppelung und die Verwendung von ⟨è⟩.

(3.3) Em Grã-scheff isch es unaaggnèèm .
Dem Grand-Chef ist es unangenehm .

Einen Zwischenweg zeigen die Daten von *Swatch*. In Beispiel 3.4 fällt die Schreibung mit Doppelvokalen in Wörtern wie *ströömt* und *aasteckends* auf. Bei *Kommunikation* stimmt die Schreibung mit der standarddeutschen überein. Damit folgt die Schreibung im Grossen und Ganzen derjenigen, die von Marti empfohlen wird, wo die Nähe zum Standard in unmissverständlichen Fällen empfohlen wird.

(3.4) Es aasteckends Gfüeu wo us aune Produkt ströömt u o
Ein ansteckendes Gefühl , das aus allen Produkten strömt und auch
ir Kommunikation vor Marggä verankeret isch .
in der Kommunikation von der Marke verankert ist .

Auch bei den Texten im Teil *Wiki* ist die Schreibung eher standardnah, wie Beispiel 3.5 zeigt. Wie im Beispiel 3.4 bei *Gfüeu* ist die Vokalisierung von *l* durch ⟨u⟩

dargestellt, während Dieth und Marti auch Möglichkeiten wie ⟨w⟩ oder ⟨l̥⟩ vorschlagen. Was ausserdem auffällt ist die Kombination *we uf* ohne hiatustilgenden Konsonanten ⟨n⟩. Ohne die Aussprache dieses Satzes durch den Schreiber können keine vollständigen Schlüsse gezogen werden, doch es ist möglich, dass der Schreiber ein /n/ sprechen würde (also in [ʊɛnuf]) und einem einheitlichen Schriftbild zuliebe hier nicht schreibt.

(3.5) d Vokalisierig findet nid statt , we uf ds l e Vokau chunnt .
Die Vokalisierung findet nicht statt , wenn auf das *l* ein Vokal kommt .

Zusammengefasst kann also gesagt werden, dass bei literarischen Texten wie derjenigen von Schobinger die Schreibung bewusst den Empfehlungen folgt. Einen Mittelweg nehmen die Texte von Blick, Swatch und Wikipedia ein, wo die Schreiber teils den Empfehlungen folgen, andererseits aber nicht gleich stark an deren Einhaltung interessiert sind und oft standardnahe Schreibungen verwenden. Die Schreibung, die am meisten von den Empfehlungen von Dieth oder Marti abweichen, finden sich in den Blogs. Diesen Schreibern sind die Schreibweise-Empfehlungen entweder unbekannt oder sie legen keinen grossen Wert auf deren Einhaltung.

Für ein Morphologieanalysesystem, das auf Ersetzungen basiert, sind die standardnahen Formen aus naheliegenden Gründen die einfachsten. Auch eine konsequente Schreibung ohne Abhängigkeit vom Lautlichen wie bei *we uf* vereinfacht die Verarbeitung. Auf der anderen Seite der Skala sind Schreibungen wie *Grã-scheff*, die weit vom standardsprachlichen ‚Grand-Chef' entfernt liegen. Eine Erfassung solcher lautnaher Schreibungen nach der Art eines Spracherkennungssystems würde den Rahmen dieser Arbeit sprengen und es muss deshalb auf die Erkennung dieser Formen verzichtet werden.

3.2 Klassifikationsstandards für Wortarten und Morphologie (Tagsets)

Die Hauptwortarten *Substantive*, *Verben*, *Pronomina* etc. (vgl. Teufel und Stöckert 1996, S. 10; Schiller et al. 1999, S. 4) ermöglichen lediglich eine relativ grobe Einteilung, die oft durch die Definition von Untergruppen ergänzt wird. Hinzu kommen meistens noch Labels für Satzzeichen, Sonderzeichen und andere Elemente, die nicht durch die Hauptwortarten abgedeckt sind, aber dennoch nötig sind, wenn ein Text lückenlos annotiert werden soll.

Für Standarddeutsch fungiert das *Stuttgart-Tübingen-Tagset* (STTS) von Schil-

ler et al. (1999) als weit akzeptierter Standard. Dieses beinhaltet zum einen ein einfaches Wortartentagset mit 11 Tags, das sich vor allem für lexikalische Beobachtungen eignet und zum anderen ein weit verbreitetes Tagset mit 54 Tags, das auf lexikalische Eigenheiten wie Untergruppen einer Wortart (wie Voll-, Hilfs- und Modalverben) eingeht und morphologische Informationen wie die Finitheit der Verben kennzeichnet. Durch die bereits enthaltenen morphologischen Informationen in den Wortartentags sind die Felder für die morphologischen Features in der Regel schon klar definiert, die in Kombination mit den feinen Wortartentags eine dritte Möglichkeit darstellen.

Neben modernem Standarddeutsch wird das STTS in angepasster Form auch für andere Varietäten der deutschen Sprache eingesetzt. Ein solches Gebiet ist Kiezdeutsch, eine Form der urbanen Jugendsprache in Deutschland. Rehbein und Schalowski (2013) gehen dabei vor allem auf Phänomene der gesprochenen Sprache ein und ergänzen das Tagset um Tags für Partikeln in der gesprochenen Sprache, Tags für diskursbedingte Besonderheiten sowie Möglichkeiten, Probleme bei der Transkription zu markieren. Das Tagset bildet die Grundlage eines Taggers, den sie auf Kiezdeutsch trainiert haben.

Im *Referenzkorpus Althochdeutsch* wird ein vom STTS abgeleitetes Tagset *Deutsch Diachron Digital Tagset* (DDDTS) verwendet (Linde 2011). Das Tagset folgt im System den Richtlinien von Schiller et al. (1999), ist aber durchaus als selbständiges Tagset zu betrachten.

3.2.1 Wortartentags

Mit dem *Swiss SMS Corpus* von Stark et al. (2009–2015) sind bereits zwei unterschiedliche Tagsets für Wortartenannotation für Schweizerdeutsch in Gebrauch. Genau genommen handelt es sich dabei um Schemata, die für die Annotation von standarddeutschen Glossen für schweizerdeutsche Texte angepasst wurden. Eines davon entspricht dem STTS und wurde bei der Annotation des Korpus (Ueberwasser 2015b) mit *TreeTagger* (Schmid 1995) verwendet.

Das andere Schema, das im *Swiss SMS Corpus* in Verwendung ist, ist das *EAGLES*-Tagset von Teufel und Stöckert (1996), welches bei der Annotation mit dem *RFTagger* (Schmid und Laws 2008) verwendet wurde. Teile, die beim STTS im Wortartentag enthalten sind, werden dabei als sonstige Merkmale aufgeführt. Die beiden Systeme sind aber kompatibel zueinander definiert (Schmid und Laws 2008, S. 781). Mit den Wortarten und den morphologischen Kategorien lassen sich für dieses Tagset über 700 Tags kombinieren.

Beide Sets wurden durch ein Tag für Infinitivpartikeln ergänzt, für die es keine standarddeutsche Entsprechung gibt und die schweizerdeutsche Form als Glosse übernommen werden musste. Das Tag für diese Wortart ist *PTKINF* beim STTS und *PART.INF* beim *EAGLES*-Tagset. Infinitivpartikeln tragen dort die Bezeichnung *PART.INF*. Eine solche Infinitivpartikel ist in Beispiel 3.6 mit *go* enthalten.

(3.6)

Mir	gönd	den	ame	au	ali	mitenand	go	Zmittag
Wir	gehen	dann	jeweils	auch	alle	miteinander	*go*	Mittag
PPER	VVFIN	ADV	ADV	ADV	PIS	ADV	PTKINF	NN
esse	..							
essen	...							
VVINF	$.							

Beim NOAH-Korpus von Hollenstein und Aepli (2014) ist eine weitere Variante des STTS in Verwendung, die aber direkt für Tagging schweizerdeutscher Texte gebraucht wird. Da im Gegensatz zu den Glossen Wörter mit Klitika (z. B. *bruchtmese*) ein Token sind, müssen die Wortarten der Bestandteile kombiniert werden. Als Verbindungselement bedienen Hollenstein und Aepli sich des „+“-Zeichens, dessen Gebrauch im nächsten Abschnitt erläutert wird. Auch in diesem Korpus werden Infinitivpartikeln mit *PTKINF* gekennzeichnet.

Mein Projekt orientiert sich am Gebrauch des STTS wie im NOAH-Korpus. Grund dafür ist, dass mit dem NOAH-Korpus aufwendig annotierte Daten vorliegen und eine Kompatibilität damit angestrebt wird. Beim *Swiss SMS Corpus* liegt der Wert dagegen in der Glossierung und die Wortartenannotation ist bloss die Ausgabe eines Taggers.

Klitika und Kontraktionen

Die Tags für klitisierte oder kontrahierte Wörter bedürfen tieferer Betrachtung. Es handelt sich dabei eigentlich um zwei verschiedene Arten, wie zusammengezogenen Wörtern Wortarten zugewiesen werden können.

Bei der ersten Art gehören alle Bestandteile des Wortes zu einem Token und auf Seiten der Wortartentags ist die Anwesenheit von Klitika markiert. Dabei sind die Tags mit Hilfe von Pluszeichen erweitert. Als Beispiel führen Hollenstein und Aepli das Wort *bruchtmese* ‚braucht man sie‘ auf, für das sie das Tag *VVFIN+PIS+PPER* vorsehen. Im NOAH-Korpus ist dieses System in vereinfachter Art eingesetzt und das Tag nach dem ersten Pluszeichen abgetrennt. Das Beispiel trägt im Korpus also nur das Tag *VVFIN+* (hier als Beispiel 3.7). Hollenstein und Aepli begründen diesen Entscheid damit, dass nur wenige fixierte Muster als Kom-

binationen von Tags vorkommen und der Informationsverlust klein ist. Durch die Auszeichnung klitisierter Formen mit einem „+“ steigt die Anzahl Tags zumindest theoretisch auf etwa 100 an, wobei aber nur 30 Tags mit „+“ erweitert im Korpus vorkommen. Von 1652 solcher Tags sind 976 Erweiterungen von finiten Verbformen (*VVFIN+*, *VAFIN+* oder *VMFIN+*). Weitere 214 sind Erweiterungen von satzeinleitenden unterordnenden Konjunktionen (*KOUS+*) und 143 sind Relativpronomina (*PRELS+*). Danach nehmen die Zahlen schnell ab, sodass nur 14 der erweiterten Tags zehn Mal oder öfter vorkommen.

Bei der zweiten Art wird nicht das Token mitsamt seinen Klitika, sondern nur seine Bestandteile annotiert. Die Beispiele 3.8 (mit *TreeTagger*) und 3.9 (mit *RFTagger*) zeigen, wie dies im *Swiss SMS Corpus* gemacht wird (vgl. Ueberwasser 2013, Kap. 3.2). Die manuelle Glossierung erlaubt dabei eine Aufteilung in beliebig viele Wörter auf der Annotationsebene, wobei das Token auf der Oberfläche unverändert bleibt.

Während die Verkürzung des Tags bei 3.7 einen Informationsverlust darstellt, handelt es sich zwischen *VVFIN+PIS+PPER* und dem Schema in 3.8 bloss um eine Darstellungsfrage, die allerdings auch Auswirkungen auf die Darstellung in XML-Strukturen haben kann. Das Schema in 3.9 und 3.8 unterscheidet sich dagegen nur in der Feinheit des Tagsets.

(3.7)
bruchtmese
VVFIN+

(3.8)
bruchtmese
braucht man sie
VVFIN PIS PPER

(3.9)
bruchtmese
braucht man
VFIN.Full.3.Sg.Pres.Ind PRO.Indef.Subst.Nom.Sg.*

sie
PRO.Pers.Subst.3.Acc.Sg.Fem

Die Schaffung neuer Tags nach dem Muster *APPRART* (aus *APPR* und *ART*), wie sie im Standarddeutschen für Wörter wie *im* verwendet werden, lehnen Hollenstein und Aepli ab, um die Zahl der möglichen Tags nicht zu erhöhen.

Der Gebrauch von *APPRART* für Kontraktionen aus Artikeln und Präpositionen ist aber gerade ein Punkt für Uneinigkeit. Während Hollenstein und Aepli im NOAH-Korpus konsequent das Tag *APPRART* verwenden, wird dieses Tag im *Swiss SMS Corpus* von Stark et al. (2009–2015) nur für die Kontraktionen ange-

wandt, die auch im Standarddeutschen existieren. Die konsequente Verwendung von *APPRART* im NOAH-Korpus entspricht dem Sprachgebrauch des Schweizerdeutschen, wo andere Konstruktionen existieren und eine unterschiedliche Behandlung nicht sinnvoll erscheint. Die Orientierung am Standarddeutschen wie im SMS-Korpus erlaubt dagegen eine Generierung von Glossen aus der Annotation ohne Zwischenschritte. Als dritte Möglichkeit könnte man auch komplett auf das Tag *APPRART* verzichten und alle passenden Wörter mit *APPR+ART*[2] annotieren und damit wie mit anderen Kontraktionen umzugehen. Der konsequente Gebrauch von *APPRART* in dieser Arbeit begründe ich mit dessen Verwendung im NOAH-Korpus und dem Wunsch nach Kompatibilität. In Kombination aller morphologischen Merkmale für Artikel lassen sich andere Formate erstellen.

Beispiel 3.10 weist zwei solche Kontraktionen aus Artikeln und Präpositionen auf. Davon ist *ar* nur im Schweizerdeutschen möglich, während im Standarddeutschen für ‚an der' zwei Tokens notwendig sind. Die Auszeichnung als *APPRART* folgt dem Gebrauch bei Hollenstein und Aepli. Bei Stark et al. dagegen sind die Glossen separat mit *APPR* und *ART* bezeichnet. Das zweite Vorkommen für eine solche Kontraktion ist *am*, das auch im Standarddeutschen existiert und für das unumstritten das Tag *APPRART* verwendet wird. Weitere Beispiele wie *vodä Modäu* ‚von den Modellen' oder *vomnä Produkt* ‚von einem Produkt' zeigen, dass Kontraktionen aus Präpositionen und Artikeln im Schweizerdeutschen auch im Plural oder mit unbestimmten Artikeln möglich sind.

(3.10)

Mir	probiere	nis	ender	chli	ar	dütsche	
Wir	probieren	uns	eher	bisschen	an der	deutschen	
PPER	VVFIN	PRF	ADV	PIS	APPRART	ADJA	
Schribwiss	aazpasse	,	aus	am	rein	Phonetische	
Schreibweise	anzupassen	,	als	am	rein	Phonetischen	
NN	VVIZU	$,	KOKON	APPRART	ADJD	NN	
.							
.							
$.							

Ein weiterer Punkt, wo im Standarddeutschen Wörter kontrahiert werden, ist bei Partikelverben mit der Infinitivpartikel ‚zu', markiert mit dem Tag *VVIZU*. Bei der Morphologiegenerierung von Scherrer[3] kann dieses Tag auch für Verben ohne Partikeln (allerdings nur für Vollverben) angewendet werden und *zu* wird direkt

[2]Der Unterschied dieser beiden Möglichkeiten wird in Kombination mit morphologischen Merkmalen offensichtlich. *am* würde in ersterem Fall mit ‚an/APPRART.ddsm' und in letzterem Fall mit ‚an/APPR.d+die/ART.ddsm' annotiert.

[3]Interaktiv verfügbar `http://latlntic.unige.ch/~scherrey/prod/dialectmorpho.html` (aufgerufen am 15. Februar 2016)

angefügt. Bei diesem Ansatz ergibt sich der Vorteil, dass die entsprechende Form für alle Verben definiert ist und das bedeutungstragende Lemma an erster Stelle steht. Als Nachteil ergibt sich, dass auf standarddeutscher Seite keine Entsprechung existiert – was zwar behoben werden kann – und im Vergleich zur Präpositionen ‚zu' oder zur Partikel ‚zu' vor Adverbien oder Adjektiven eine Inkonsequenz entsteht. Eine Verwendung analog zum Standarddeutschen mit *VVIZU* erscheint am angemessensten.

Die Zusammenschreibung von Verbformen und Partikeln ohne ‚zu' dagegen gilt im STTS als eigenes Wort und eine Markierung auf der Wortartenebene bleibt aus.

Während der Annotation der Trainingsdaten stellte sich heraus, dass für *am* im sogenannten *am-Progressiv* die Auszeichnung als Präposition mit Artikel unzureichend ist. Gemäss Duden ist *am* in der Standardsprache als Teil der Verlaufsform unauflösbar, trotzdem wird es als Präposition mit Artikel bezeichnet. Analog gilt auch *am* bei Superlativen als unauflösbare Verschmelzung aus Präposition und Artikel (Duden online 2016). Letzteres trägt im STTS das Tag *PTKA*, während bei der Verlaufsform nur *APPRART* zur Verfügung steht.

Während die von Duden empfohlene Interpretation des Infinitivs als Substantivierung bei einfachen Konstruktionen möglich ist, ist dies nicht der Fall, wenn zusammen mit dem Verb Objekte stehen. Ein vergleichbarer Fall liegt mit Beispiel 3.11 vor, wo das Prädikativ vor *am* steht. Das Wörtchen *am* trennt dabei die Verbalphrase *herzigi chlini Fellbölle werde* auf, was die Interpretation von *am* als Präposition mit Artikel, die ja die ganze Phrase einleiten müsste, ausschliesst. Die enthaltene Nominalphrase *herzigi chlini Fellbölle* kann nur ohne Einleitung stehen, wenn *werde* ein Infinitiv ist. Wäre *werde* ein Substantiv, müsste die Nominalphrase mit einer Präposition eingebunden werden.

(3.11)

d'	Babybüsis	entwicklet	sich	prächtig	,	sind	herzigi	chlini
Die	Babykatzen	entwickeln	sich	prächtig	,	sind	herzige	kleine
ART	NN	VVFIN	PRF	ADJD	$,	VAFIN	ADJA	ADJA
Fellbölle	am	werde	.					
Feldbälle	am	werden	.					
NN	PTKAM	VAINF	$.					

Auch in der Forschung zur Grammatikalisierung (z. B. Van Pottelberge 2005, S. 183) werden diese Argumente als Hinweise auf eine mögliche „Reanalyse des *am*-Progressivs als periphrastische Verbform" beachtet. Zu den Regionen, in denen Akkusativobjekte vom *am*-Progressiv regiert werden können, zählt insbesondere auch die Schweiz (Van Pottelberge 2005, S. 183 – 184).

Bezüglich der Position der Objekte (aber nicht der Verbzusätze) ist auch das ‚zu' vor Infinitiven vergleichbar. Eine Analyse als Partikel scheint also sinnvoll und ein spezifisches Tag *PTKAM* für die Verlaufsform ist unter anderem für die Erforschung der Grammatikalisierung von Vorteil.

Ähnlich wie *am* in der Verlaufsform verhält sich *zum* als unterordnende Konjunktion mit Infinitiv. Wie in Beispiel 3.12 ersichtlich ist, verhält sich *zum* wie eine Konjunktion und nicht wie eine Präposition mit Artikel, zumal es sich um einen Infinitiv mit ‚zu' handelt. Eine Ausweitung des Tags *KOUI* liegt auf der Hand, wenn man Beispiele wie 3.13 beachtet.

(3.12)

zum	es	massgschniiderets	Konzept	vo	attraktive	und	
um	ein	massgeschneidertes	Konzept	von	attraktiven	und	
KOUI	ART	ADJA	NN	APPR	ADJA	KON	
innovative	Gschäft	azbüüte	.				
innovativen	Geschäften	anzubieten	.				
ADJA	NN	VVIZU	$.				

(3.13)

für	uf	di	groossi	Nachfrag	nach	Schwiizer	Qualitätsuhrä
um	auf	die	grosse	Nachfrage	nach	Schweizer	Qualitätsuhren
KOUI	APPR	ART	ADJA	NN	APPR	ADJA	NN
z'	reagierä	.					
zu	reagieren	.					
PTKZU	VVINF	$.					

Für eine abschliessende Übersicht sind die Wortarten, die in dieser Arbeit für das Schweizerdeutsche zum STTS hinzugefügt worden sind oder die für andere Wörter als im Standarddeutschen verwendet werden, in Tabelle 2 gesammelt. Eine vollständige Liste (Tabelle 23) befindet sich im Anhang.

3.2.2 Morphologische Tags

Bisher scheint es kein spezifisches, zum STTS gehöriges Schema für die morphologische Annotation für Schweizerdeutsch zu geben, das weitherum in Verwendung wäre. Auch für Standarddeutsch gibt es verschiedene Ansätze, die zum Teil nur in der Bezeichnung der Tags, zum Teil aber in ihrer Reihenfolge und sogar den Kategorien voneinander abweichen.

Schiller et al. (1999, S. 8) definieren zwar zusätzlich zum Tagset für die Wortarten ein Tagset für die morphologischen Features, doch scheint dieses keine weite Verbreitung zu geniessen, da an dessen Stelle das kompatible Tagset EAGLES

Tag	Wortart	Kommentar
APPRART	Präposition mit Artikel	Auch für Plural und unbestimmte Artikel
KOUI	Konjunktion mit Infinitiv	Auch für Konjunktionen wie *für* oder *zum*
KOUS	unterordnende Konjunktion	Zusätzliche Analyse für *wo* ‚als' (zeitlich)
PRELS	Relativpronomen *wo*	
(PRELAT)	attribuierendes Relativpronomen	Verzicht
PTKAM	Partikel *am* bei Verlaufsform	Neuschöpfung
PTKINF	Infinitivpartikeln *go, la, cho*	Übernahme von Stark et al. (2009–2015)

Tabelle 2: Erweiterungen des STTS für Schweizerdeutsch. Die aufgeführten Wortartentags wurden zum STTS hinzugefügt oder weichen in deren Verwendung ab.

(Teufel und Stöckert 1996) aus dem gleichen Haus verwendet wird. An die Erweiterung zum STTS von Schiller et al. angelehnt definieren Crysmann et al. (2005) das *TIGER Morphologie-Annotationsschema*. Es weicht jedoch insbesondere dadurch vom Schema von Schiller et al. ab, dass es keine Angaben über die starke oder schwache Flexion bei den Adjektiven (und Substantiven) und die Definitheit bei Artikeln macht. Andererseits führen Crysmann et al. (2005) Tags für die finiten Verbformen auf, die mit dem grossen STTS redundant, dafür aber mit den kurzen Tags kompatibel sind. Zu beachten ist auch, dass die sich Reihenfolge der Kategorien von jener bei Schiller et al. unterscheidet.

Ein System, das einen komplett verschiedenen Ansatz verfolgt, wird bei der *Tübingen Treebank of Written German TüBa-D/Z* verwendet (Telljohann et al. 2015, S. 23 – 24). Die morphologischen Angaben bestehen dort aus kurzen Buchstabengruppen bei denen jeder Buchstabe für eine Kategorie steht. Angaben zur Steigerung der Adjektive fehlen, stattdessen sind auch gesteigerte Formen als Lemmata zugelassen.

Beim EAGLES-Tagset (Teufel und Stöckert 1996), welches der *RFTagger* (Schmid und Laws 2008) verwendet, ist die Grenze zwischen Wortartentags und morphologischen Tags mehrheitlich aufgehoben. Oft entspricht der erste Teil dem Wortartentag im STTS, in einigen Fällen sind auch die ersten zwei oder drei Teile einem STTS-Tag äquivalent. Bei den rein morphologischen Tags weist das EAGLES-Tagset eine andere Reihenfolge auf als von Schiller et al. für das STTS vorgeschlagen.

Das von Scherrer (2011) bei der Formengenerierung verwendete System folgt einem Schema, das dem von Schiller et al. am nächsten steht. Es unterscheidet sich zwar in der Anordnung, aber nicht in der Wahl der grammatischen Kategorien. Auch

dieses System ist dem standarddeutschen System noch sehr nahe.

Der beste Weg für meine eigene Arbeit besteht nun darin, ein Schema zu entwickeln, das mit den bestehenden Standards der Lemmatisierung (mehr dazu in Kapitel 3.3) kompatibel ist und das sich möglichst eindeutig und einfach in andere Standards überführen lässt. Dies kann einerseits in einem Nachbearbeitungsschritt erfolgen und andererseits permanent durch Umschreibung der Software erreicht werden. Dazu müssen die Tags sowohl eindeutig und einfach suchbar, als auch möglichst zentral definiert sein. Mit einbuchstabigen Tags mit allfälligen Begrenzungszeichen scheint mir diese Bedingung erfüllt. Die Kürze der Tags ist auch eine Erleichterung beim Annotieren. Tabelle 3 zeigt für dieses Tagset, welchen Wortarten welche Kategorien zugeordnet sind und wie die entsprechenden Werte zu belegen sind.

Bei den möglichen Belegungen der morphologischen Kategorien kann grösstenteils wie im Standarddeutschen verfahren werden. Wie es in gewissen Beschreibungen für Standarddeutsch verbreitet ist, kann auf die gemischte Flexion der Adjektive verzichtet werden. Die Beispiele 3.14 und 3.15 (typisch für Zürich beziehungsweise für Basel) zeigen den uneinheitlichen Gebrauch der Flexion nach Possessivpronomina. Der Vergleich mit der gemischten Flexion (nach unbestimmtem Artikel) in den Beispielen 3.16 für Zürich und 3.17 für Basel zeigt auch, dass sich nicht bloss die Endungen regional unterscheiden, sondern dass auch die Flexionstypen unterschiedlich verwendet werden. Demgegenüber ist die Annotation der Flexionstypen eindeutig, wenn lediglich die starke und die schwache Flexion zugelassen ist. Auch für syntaktische Untersuchungen eignet sich diese Einteilung besser.

(3.14) mis rooti Huus
mein rot-*schwach* Haus
‚mein rotes Haus‘

(3.15) mi roots Huus
mein rot-*stark* Haus
‚mein rotes Haus‘

(3.16) es roots Huus
mein rot-*stark* Haus
‚ein rotes Haus‘

(3.17) e roots Huus
mein rot-*stark* Haus
‚ein rotes Haus‘

Besondere Aufmerksamkeit verdienen auch die Kasus, darunter besonders Nomi-

nativ und Akkusativ. Reste einer Kasusflexion bei Substantiven gibt es nur noch bei der starken Deklination im Dativ Plural (*de Lüüt* oder *de Lüüte* für ‚den Leuten'). Diese Markierung kann aber kaum als obligatorisch bezeichnet werden und ist leicht mit der schwachen Pluralendung (*Zaal* ‚Zahl' vs. *Zaale* ‚Zahlen') zu verwechseln.

Ganz anders sieht es bei den Adjektiven aus, die für den Dativ noch klare eigenständige Formen zeigen (*gueti Sach* ‚gute Sache' und *gueter Sach* ‚guter Sache'). Andererseits lassen sich Nominativ und Akkusativ nicht an der Form unterscheiden (*guete Raat* für ‚guter Rat' und ‚guten Rat'), was ebenfalls für Artikel und viele Pronomina gilt. Im *Swiss SMS Corpus* sollen solche Formen zu Nominativ normalisiert werden (Ueberwasser 2015a). Für die morphologische Analyse hat das aber zur Folge, dass Präpositionen, die üblicherweise auch mit Kasus markiert werden, und die entsprechende Nominalphrase unterschiedliche Kasus tragen. Ein vollständiger Verzicht auf den Akkusativ ist ausgeschlossen, weil beispielsweise bei den Personalpronomina klar zwischen drei Kasus unterschieden wird. Dort kann Nominativ nicht nach einer Präposition stehen. Eine Analyse zum Akkusativ in Zweifelsfällen führte zwar einheitliche Präpositionalphrasen, widerspricht aber den sprachlichen Traditionen. Als Mittelweg soll hier ein Tag eingeführt werden, das sowohl für Nominativ als auch für Akkusativ stehen kann.

3.3 Lemmatisierung

Die Zuweisung standarddeutscher Lemmata für schweizerdeutsche Tokens sollte nach einem konsistenten und verständlichen Schema erfolgen. Schwierigkeiten ergeben sich aber, sobald für ein schweizerdeutsches Wort keine direkte standarddeutsche Entsprechung existiert. Ein grosser Teil dieser Fälle wird bereits durch die Richtlinien zur Glossierung von Ueberwasser (2013, Kap. 3.2) geregelt. Deren Anweisung, möglichst die direkt verwandten Wörter zu übernehmen und bei den Präpositionen rein auf die Form und nicht auf die Bedeutung zu achten, kommt dem hier gewählten Ansatz mit Lautentsprechungen sehr entgegen.

Die Richtlinien für die Glossen lassen sich für die Lemmatisierung komplett übernehmen. Das gilt vor allem für die Regeln, dass Lemmata in ihrer Form möglichst nahe bleiben sollen, keine Wörter geschaffen werden sollen und der semantische Gehalt gewahrt sein soll.

Was Gross- und Kleinschreibung und die als Lemma gültige Form bei Wörtern der funktionellen Wortarten angeht, ist durch die im Korpus verwendeten Tagger

Morph. Kategorie	Wortarten	Morphologische Tags
Grad	ADJA, ADJD	p: Positiv c: Komparativ s: Superlativ
Person	PPER, PRF, VAFIN, VMFIN, VVFIN	1: erste Person 2: zweite Person 3: dritte Person
Kasus	ADJA, APPRART, ART, (NN), PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRF, PWAT, PWS	n: Nominativ a: Akkusativ d: Dativ r: Nominativ/Akkusativ
Numerus	ADJA, APPRART, ART, NN, NE, PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRF, PWAT, PWS, VAFIN, VMFIN, VVFIN, VAIMP VVIMP	s: Singular p: Plural
Genus	ADJA, APPRART, ART, NN, NE, PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRF, PWAT, PWS	m: Maskulinum f: Femininum n: Neutrum
Modus	VAFIN, VMFIN, VVFIN	i: Indikativ Präsens j: Konjunktiv I k: Konjunktiv II
Flexion	ADJA, (NN)	s: stark w: schwach
Definitheit	APPRART, ART	i: indefinit d: definit

Tabelle 3: Morphologisches Featureset. Zusätzlich zu den aufgeführten Tags kann „*“ verwendet werden, wenn die Kategorie nicht bestimmbar ist. Im Plural steht für das Genus immer „*“ ausser bei NN und NE. NN tragen nur Werte für die Flexion und den Kasus, wenn sie wie Adjektive flektiert sind.

schon ein Usus festgesetzt. Für die Artikel (ART) wird hier die Lemmatisierung als ‚die' oder ‚eine' übernommen.

Bei Präpositionen mit Artikel (APPRART) wie *am* werde ich dagegen bloss die Präposition als Lemma verwenden (als ‚an' etc.). Grund dafür ist die Verwendung von *APPRART* auch in Fällen, für die eine direkte Entsprechung im Standarddeutschen fehlt.

Die Personalpronomina werden jeweils nach ihrer Nominativform (‚ich, du' etc.) lemmatisiert. Die Reflexivpronomina dagegen tragen unabhängig von Person und Zahl das Lemma ‚sich'. Anhand ihrer Stämme (‚mein, ihr' etc.) werden die Possessivpronomina lemmatisiert. Bei den übrigen Pronomina gilt, dass sie – falls sie wie *vill* in Beispiel 3.18 endungslos sind – mit ihrem Stamm lemmatisiert werden. Tragen sie demgegenüber eine Endung wie *viles* in Beispiel 3.19, werden sie als dekliniertes Pronomen annotiert und das Lemma trägt die Endung ⟨-e⟩. Einer kurzen Erwähnung bedürfen noch die Formen wie *wem* oder *wasem*, die wie die Nominativformen als ‚wer' bzw. ‚was' lemmatisiert werden.

(3.18) de Agathias het scho
die/ART.drsm Agathias/NE.sm haben/VAFIN.3si schon/ADV
vill gwüsst .
viel/PIS.*** wissen/VVPP ./$.
‚Agathias wusste schon viel.'

(3.19) de Agathias het scho
die/ART.drsm Agathias/NE.sm haben/VAFIN.3si schon/ADV
viles gwüsst .
viele/PIS.rsn wissen/VVPP ./$.
‚Agathias wusste schon vieles.'

Die Infinitivpartikeln stellen auch hier einen Spezialfall dar, da eine standarddeutsche Entsprechung fehlt. Während Ueberwasser (2013, Kap. 3.2) für das *Swiss SMS Corpus* zwar *go* für alle von ‚gehen' abgeleiteten Formen festlegt, folgt das Korpus diesem Ansatz nicht. Für die morphologische Analyse werde ich hier ein einheitliches Lemma wie *go* verwenden.

Fremdsprachiges Material und unklare Wörter wird vom Morphologiesystem *Taggswort* nicht berücksichtigt. Bei der manuellen Annotation dürfen auch standarddeutsche Sätze und Phrasen als fremdsprachiges Material annotiert werden, ausser es handle sich um im Schweizerdeutschen etablierte Lehnwörter. Für unklare Wörter muss in jedem Fall separat entschieden werden. Bei der Annotation traten solche Unklarheiten aber selten auf.

3.4 Überblick über das Annotationsschema

An dieser Stelle sollen die Prinzipien der Annotation noch einmal kurz aufgezeigt werden. Die Annotation der Wortarten folgt in den Grundsätzen dem Schema des STTS (Schiller et al. 1999). Adaptionen für das Schweizerdeutsche sind die Tags *PTKINF* für Infinitivpartikeln und *PTKAM* für die Partikel *am* in der Verlaufsform. Tabelle 2 gibt einen Überblick zu den wichtigsten Unterschieden zum Tagset für Standarddeutsch.

Für die morphologischen Merkmale wurde ein eigenes Schema mit einbuchstabigen Tags definiert. Besonders ist dabei das Tag „r“ zu erwähnen, das für Nominativ und Akkusativ stehen kann und in den meisten Wortarten ausser den Personal- und Reziprokpronomina sowie den Präpositionen verwendet wird. Bei Ausbleiben der Merkmale einer Kategorie kann das Tag „*“ gesetzt werden. Tabelle 3 gibt einen Überblick über die morphologischen Tags.

Die Lemmatisierung kann nach den gleichen Kriterien erfolgen wie die Glossierung der schweizerdeutschen Teile im *Swiss SMS Corpus* von Stark et al. (2009–2015). Darunter fällt die Wahl ähnlicher Lemmata, der Verzicht auf Neuschöpfungen und die Wahrung des semantischen Gehalts. Für die Lemmata soll grundsätzlich die Grundform wie in Wörterbüchern verwendet werden. Einzig bei Pronomina soll die Verwendung der Endung ⟨-e⟩ flektierte von nicht-flektierten Lemmata unterscheiden.

Die syntaktischen Tags sollen mit Hilfe eines Punktes mit den morphologischen Tags verbunden werden. Dieses somit feine morphologische Tagset soll entsprechend der Form seiner Tags als *STTS.gsw* bezeichnet werden. Lemmata und Tags sollen durch den Schrägstrich „/“ verbunden werden. Für Tokens mit Klitika, die somit mehrere Lemmata und Tags aufweisen, sollen diese Teile mit dem Pluszeichen „+“ verbunden werden. Mit dem oben definierten Tagset wird *bruchtmese* schliesslich wie in 3.20 annotiert und lemmatisiert.

(3.20) bruchtmese
brauchen/VVFIN.3si+man/PIS.ns*+sie/PPER.3sfa
‚braucht man sie‘

Mit der festgelegten Adaption des STTS sind nun 55 Tags in Verwendung. Zusammen mit den morphologischen Angaben ergeben sich etwa 350 Kombinationen. In *Taggswort* sind rund 160 Kombinationen aus morphologischen Tags bekannt.

4 Material und Methoden

Dieses Kapitel widmet sich dem computergestützten Teil des vorgestellten Projekts. Neben der Annotation der Entwicklungs- und Testdaten wird in die Finite-State-Methode eingeführt und das praktische Vorgehen mit diesen Werkzeugen erklärt.

4.1 Korpus

Für die Entwicklung und Evaluation wurde das NOAH-Korpus (Hollenstein und Aepli 2014) verwendet. Dieses wurde in acht Teile mit etwa gleichem Umfang aufgeteilt, von denen einer für die Entwicklung und einer für die Evaluation des Morphologieanalysesystems *Taggswort* ausgewählt wurden. Bei der Aufteilung wurde einerseits auf ein ausgeglichenes Resultat geachtet und andererseits jeweils sechs aufeinanderfolgende Sätze zusammengehalten, um den Zusammenhang zu einem gewissen Grad zu bewahren.

In Kapitel 3.1.3 wurden bereits die fünf nach Quelle beziehungsweise Textgattungen erstellten Teile vorgestellt. Nun gibt Tabelle 4 einen Überblick, nach welchen Kategorien sich das Korpus aufteilen lässt und wie gross die einzelnen Teile sind.

4.1.1 Annotation

Um den Aufwand zu begrenzen, wurden die Sprachdaten mit *Taggswort* im Entwicklungsstadium vorannotiert und manuell die richtigen Analysen ausgewählt respektive von Hand ergänzt. Während der Annotation des Entwicklungssets wurde dieses auch als Inspiration für die Morphologie verwendet und fehlende Wörter oder Formen ergänzt. Bei den Daten für die spätere Evaluation wurde darauf verzichtet, Formen oder Wörter in *Taggswort* zu übernehmen, um die Aussagekraft der Tests nicht negativ zu beeinflussen.

Teil		Dialekt	komplettes Korpus	Entwicklungskorpus	Testkorpus
Blick		verschiedene	**11256**	**1395**	**1564**
Blogs		verschiedene	**34834**	**4626**	**4477**
Schobinger		Zürichdeutsch	**12858**	**1575**	**1632**
Swatch		verschiedene	**34038**	**4204**	**4503**
	a2	Baseldeutsch	3167	490	415
	a3	Berndeutsch	4325	532	596
	a16	Zürichdeutsch	5055	618	640
Wiki		verschiedene	**22136**	**2527**	**2712**
	a1	Baseldeutsch	4388	409	521
	a2	Berndeutsch	4466	498	574
	a4	Zürichdeutsch	4478	542	447
alle			**115122**	**14327**	**14888**

Tabelle 4: Überblick über das NOAH-Korpus. Anzahl der Tokens in den verschiedenen Bestandteilen des Korpus und in den für die vorliegende Arbeit erstellten Subkorpora.

4.2 Automaten und Transduktoren

Endliche Automaten sind ein mathematisches Modell, um reguläre Sprachen, eine Untermenge der formalen Sprachen, zu erkennen. Eine formale Sprache $\Sigma*$ über das Alphabet Σ besteht aus den Wörtern, die aus der Verkettung der Symbole von Σ gebildet werden können (Didakowski 2005, S. 37 – 39). Für das leere Wort, d. h. leere Zeichenketten, wird üblicherweise das Zeichen Epsilon ϵ verwendet.

Endliche Transduktoren dagegen beschreiben reguläre Relationen. Diese können über das kartesische Produkt aus regulären Sprachen gebildet werden (Didakowski 2005, S. 39 – 40). Das vorliegende System zur Morphologieanalyse bildet eine solche Relation ab. Die eine Sprache ist dabei die Menge der standarddeutschen Lemmata mit der Analyse und die andere Sprache die schweizerdeutschen Dialektwörter. Die Alphabete der beiden Sprachen unterscheiden sich darin, dass die eine Sprache nur die Zeichen des Standarddeutschen und die Tags zur Annotation beinhaltet, während die andere Buchstaben beinhalten kann, die im Standarddeutschen nicht vorkommen, dafür aber keine Tags.

Die weiteren Teile dieses Unterkapitels werden weiter auf Automaten und Transduktoren eingehen.

4.2.1 Endliche Automaten

Ein endlicher Automat wird durch ein Tupel $(\Sigma, Q, q_0, F, \delta)$ definiert, das folgende Bedingungen erfüllt (siehe Didakowski 2005, S. 48):

1. Σ ist eine endliche Menge, das Eingabealphabet.
2. Q ist eine endliche Menge, die Menge der Zustände.
3. $q_0 \in Q$ ist der Startzustand.
4. $F \subseteq Q$ ist die Menge der Endzustände.
5. $\delta \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times Q$ ist eine endliche Menge, die Übergangsrelation.

Ein endlicher Automat wie in Abbildung 2 kann als Akzeptor für die Wörter einer Sprache verwendet werden. Das Eingabealphabet bezeichnet dabei die Zeichen an den Kanten und es würde im Fall eines Akzeptors für die deutsche Sprache mit dem deutschen Alphabet decken. Die Zustände $q_0 - q_6$ sind die Elemente der Menge der Zustände, der Zustand q_0 der Startzustand und die Zustände mit doppelter Umrandung sind die Endzustände beziehungsweise akzeptierenden Zustände (Menge F). Alle Pfade, die von q_0 zu einem Endzustand führen, werden akzeptiert. Die gerichteten Pfade zwischen je zwei Zuständen sind durch δ definiert. Sie können jedes Zeichen aus Σ oder die leere Zeichenkette ϵ tragen.

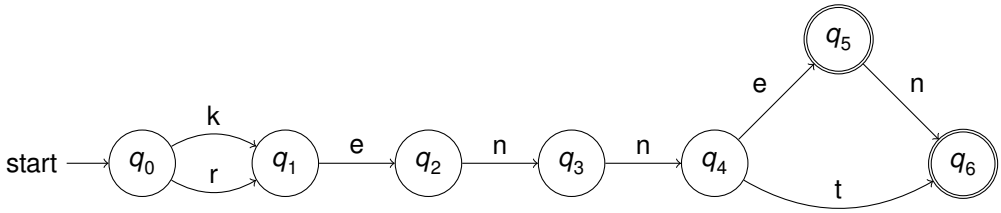

Abbildung 2: Endlicher Automat für die Sprache {kennen, kennt, kenne, rennen, rennt, renne}.

Das Beispiel in Abbildung 2 stellt einen deterministischen Automaten ohne *closure* dar, das heisst, bei jedem Zustand kann bei der Eingabe des nächsten Symbols maximal ein Zustand erreicht werden.

4.2.2 Endliche Transduktoren

Endliche Transduktoren unterscheiden sich von endlichen Automaten dadurch, dass sie zwei Zeichenmengen umfassen. Entlang der Pfeile befinden sich dementsprechend auch immer Zeichenpaare.

Ein endlicher Transduktor wird durch ein Tupel $(\Sigma, \Delta, Q, q_0, F, \delta)$ definiert, das folgende Bedingungen erfüllt (siehe Didakowski 2005, S. 49 – 50):

1. Σ ist das Eingabe- und Δ das Ausgabealphabet. Beide sind endliche Mengen.
2. Q ist eine endliche Menge, die Menge der Zustände.
3. $q_0 \in Q$ ist der Startzustand.
4. $F \subseteq Q$ ist die Menge der Endzustände.
5. $\delta \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times Q$ ist eine endliche Menge, die Übergangsrelation.

Endliche Transduktoren werden in der computergestützten Morphologie gerne für die Analyse oder Generierung verwendet. Durch Interpretation des Eingabealphabets als Ausgabealphabet und des Ausgabealphabets als Eingabealphabet kann ein Analysesystem zur Generierung verwendet werden und umgekehrt.

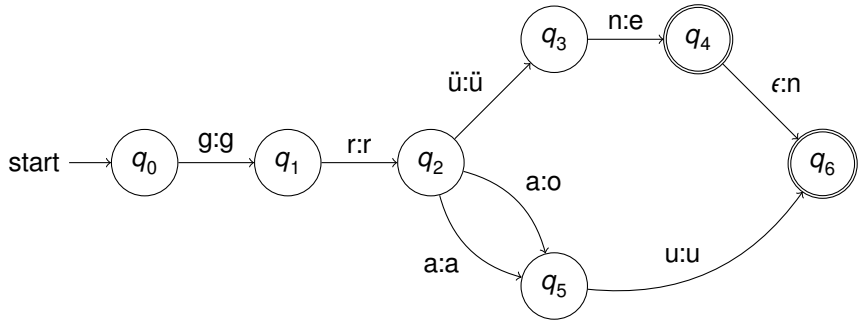

Abbildung 3: Endlicher Transduktor für die Relation {⟨grau, grau⟩, ⟨grau, grou⟩, ⟨grün, grüen⟩, ⟨grün, grüe⟩}.

Abbildung 3 stellt einen Transduktor dar, der schweizerdeutsche oder standarddeutsche Wörter von der einen auf die jeweils andere Seite überführt. Entlang der Kanten stehen jeweils Paarungen aus $\Sigma \cup \{\epsilon\}$, hier für Standarddeutsch, und $\Delta \cup \{\epsilon\}$, hier für Schweizerdeutsch.

4.2.3 Gewichtete endliche Transduktoren

Um dem Problem von vielen Ausgaben zu einer Eingabe zu begegnen, können Automaten und Transduktoren gewichtet werden. In einem System, das für ein Wort verschiedene Aussprachen ausgibt, kann mit Hilfe von Gewichten eine Rangfolge der unterschiedlichen Aussprachen gemacht werden oder den Aussprachen unterschiedliche Wahrscheinlichkeiten zugewiesen werden (siehe Mohri 2004, S. 551).

Die Gewichte dazu werden üblicherweise von Hand gesetzt oder aus grossen Datenmengen extrahiert.

Die Möglichkeit der Gewichtung erlaubt es in diesem Projekt, mit der Übergenerierung durch die vielen Dialektformen umzugehen. Aus Sicht der Analyse bedeutet Übergenerierung, dass für ein Wort eine Analyse gegeben werden kann, die in Wirklichkeit nie in Verbindung mit diesem Wort steht. Dieses Problem können Gewichte lösen, indem sie die gewünschten Analysen bevorzugen und die unerwünschten ans Ende der Rangliste schieben.

Die Einführung von Gewichten bei endlichen Transduktoren bedingt die Einführung von Gewichten entlang der Kanten und von Funktionen zur Berechnung. Gewichtete endliche Transduktoren (über S) werden durch ein Tupel $(\Sigma, \Delta, Q, q_0, F, \sigma, \lambda, \rho)$ definiert, wobei $S = (W, \oplus, \otimes, \overline{0}, \overline{1})$ ein Semiring ist und folgende Bedingungen erfüllt sind (siehe Didakowski 2005, S. 51):

1. Σ ist das Eingabe- und Δ das Ausgabealphabet. Beide sind endliche Mengen.
2. Q ist eine endliche Menge, die Menge der Zustände.
3. $q_0 \in Q$ ist der Startzustand.
4. $F \subseteq Q$ ist die Menge der Endzustände.
5. $\sigma \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times W \times Q$ ist eine endliche Menge von Übergängen, die Übergangsrelation.
6. λ ist eine Funktion von q_0 nach W, die Initialgewichtsfunktion.
7. ρ ist eine Funktion von F nach W, die Endgewichtsfunktion.

Ein Semiring wird durch ein Tupel $(W, \oplus, \otimes, \overline{0}, \overline{1})$ definiert, für das gilt (siehe Mohri 2004):

1. $(W, \oplus, \overline{0})$ ist ein kommutativer Monoid mit Identitätselement $\overline{0}$.
2. $(W, \otimes, \overline{1})$ ist ein Monoid mit Identitätselement $\overline{1}$.
3. $\otimes$ distribuiert über $\oplus$.
4. $\overline{0}$ ist negatives Element für $\otimes$: für alle $a \in W, a \otimes \overline{0} = \overline{0} \otimes a = \overline{0}$.

Ein weit verbreiteter, intuitiv verständlicher Semiring ist der probabilistische Semiring, der üblicherweise für Wahrscheinlichkeitsrechnungen verwendet wird. Entsprechend dürfen die Gewichte in der Trägermenge W nur positiv sein. Die Belegung von $\otimes$ mit $\times$ definiert, dass die Gewichte entlang eines Pfades mulitpliziert

werden. Dabei ist 1 das Identitätselement $\overline{1}$. Bei mehreren äquivalenten Pfaden werden die Gewichte der verschiedenen Pfade addiert ($\oplus$ ist $+$), um ein einzelnes Resultat zu bekommen. Identitätselement $\overline{0}$ dabei ist 0.

Ein weiterer Semiring, der auch im zu dieser Arbeit entwickelten verwendet wird, ist der tropische Semiring. Dieser wird für Optionalität verwendet und die Trägermenge W enthält die Gewichte, beliebige reelle Zahlen sein können. Positive Werte führen zu einer Bestrafung, negative Werte zu einer Bevorzugung. Die abstrakte Multiplikation $\otimes$ ist hier durch $+$ belegt, was bedeutet, dass die Gewichte entlang eines Pfades addiert werden. Das Identitätselement $\overline{1}$ dazu ist 0. Die abstrakte Addition $\oplus$ ist die im tropischen Semiring Minimumsfunktion, das heisst, dass bei äquivalenten Pfaden derjenige mit dem geringsten Gewicht bevorzugt wird. Die Belegung von $\overline{0}$ mit $+\infty$ bedeutet auch, dass ein Pfad mit $+\infty$ an einer Kante verunmöglicht ist.

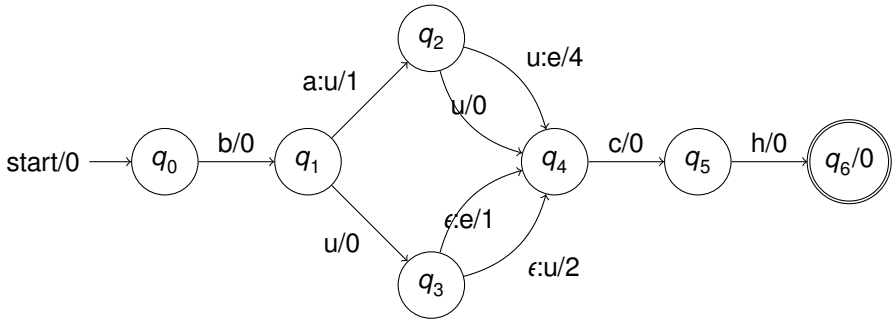

Abbildung 4: Gewichteter endlicher Transduktor für die Wortpaare {⟨buch, buuch⟩, ⟨buch, buech⟩, ⟨bauch, buuch⟩, ⟨bauch, buech⟩}. Gleiche Zeichen auf beiden Seiten werden hier durch einen Buchstaben repräsentiert.

Abbildung 4 zeigt einen gewichteten endlichen Transduktor mit einem tropischen Semiring. Das Initialgewicht ist hier das neutrale Element 0. Ein Endgewicht ist in den meisten Anwendungen das neutrale Element 0, wie es hier beim Zustand q_6 angegeben ist. In diesem Beispiel bestimmen also die Übergangsgewichte allein, welche Pfade die günstigsten sind. Dazu werden die Gewichte entlang der Pfade addiert. Jedem Element der Sprache wird dabei ein Gewicht zugewiesen:

- ⟨buch, buuch⟩ hat das Gewicht 2.
- ⟨buch, buech⟩ hat das Gewicht 1.
- ⟨bauch, buuch⟩ hat das Gewicht 1.
- ⟨bauch, buech⟩ hat das Gewicht 5.

Folgt man auf der Oberseite den Pfaden für *buch*, dann ist *buech* auf der Unterseite des besten Pfades. Mit *bauch* auf der Oberseite ist *buuch* auf der Unterseite am günstigsten.

Folgt man den Pfaden mit *buuch* auf der Unterseite, dann ist *bauch* auf der Oberseite des besten Pfades. Mit *buech* auf der Unterseite ist *buch* auf der Oberseite am günstigsten.

4.3 Finite-State-Werkzeuge

Finite-State-Werkzeuge haben in der computergestützten Morphologie Tradition, da sie einerseits die Analyse breiter Datenmengen erlauben und dabei wenig Speicher brauchen, andererseits weil mit ihnen Morphologieanalyse- und -generierungssysteme von Hand erstellt werden können. Dass für solche Systeme keine Trainingsdaten nötig sind, macht sie zu einem nützlichen Werkzeug für Sprachen mit wenig Ressourcen.

4.3.1 Ungewichtetes Werkzeug XFST

Eines der bekanntesten Systeme für die Finite-State-Methoden ist das *Xerox Finite-State Tool* (XFST) von Beesley und Karttunen (2003). Mittels der Operationen Ersetzung und Komposition, die das Interface `xfst` (Beesley und Karttunen 2003, S. 81 – 202) bietet, lassen sich viele linguistische Phänomene einfach nachbilden.

Als weiteres wichtiges Mittel erlaubt XFST *flag diacritics*, durch die Abhängigkeiten über lange Distanzen effizient umgesetzt werden können. Sie werden typischerweise zur der Begrenzung der Wörter bei einem übergenerierenden System eingesetzt, indem nur noch Wörter akzeptiert werden, deren *flag diacritics* die Form erlauben. Dies kann mit unifizierenden Merkmalen oder auch mit der Erfordernis gewisser Werte für Merkmale weiter vorne im Pfad eines Wortes gewährleistet werden.

Im Programm `xfst` ist eine Schnittstelle für Morphologien im Format von `lexc` (Beesley und Karttunen 2003, S. 203 – 278) eingebaut, worin üblicherweise die morphotaktischen Elemente einer Sprache abgebildet werden. Alternativ dazu ist das Lexikonkompilationswerkzeug `lexc` auch unabhängig verwendbar.

4.3.2 Gewichtetes Werkzeug HFST

Helsinki Finite-State Technology (HFST) von Lindén et al. (2009) ist ein System mit gewichteten Transduktoren, das sich an XFST orientiert. HFST beinhaltet ein Tool `hfst-xfst`, welches das Verhalten von XFST nachbildet und möglichst grosse Kompatibilität anstrebt. Entsprechend ist mit `hfst-lexc` eine Schnittstelle für `lexc`-Dateien mit Gewichten zugefügt. Dadurch ist das Werkzeug für erfahrene Computermorphologen einfach zu verwenden. Die Veröffentlichung als Open Source ist für die Forschung von Vorteil.

Als gewichtetes Finite-State-System eignet sich HFST nicht nur für die Morphologieanalyse, sondern auch für Rechtschreibprüfung oder Erkennung von Namen, wenn eine Rangordnung der Vorschläge erwünscht ist. Für die Gewichtung nutzt das Werkzeug HFST tropische Semiringe und die Gewichte für die Übergänge sind manuell zu setzen.

4.3.3 Komposition und Ersetzung

Die Operationen der Ersetzung und der Komposition von XFST beziehungsweise HFST werden im Rahmen des Systems *Taggswort* in grossem Umfang angewendet. Aus diesem Grund muss hier darauf eingegangen werden. Abbildung 5 enthält ein minimales Beispiel zur Interaktion zwischen Ersetzung mit dem Operator „`->`“ und Komposition mit dem Operator „`.o.`“.

```
define Lexicon [ {fuß}|{guss} ];
define Replace [ {ß} -> {ss} ];
define Compose [ Lexicon .o. Replace ];
```

Abbildung 5: Ersetzung und Komposition in XFST und HFST. `Lexicon` definiert einen einfachen Transduktor mit zwei Pfaden. `Replace` definiert eine Ersetzung und `Compose` wendet diese Ersetzung auf `Lexicon` an.

Die Variable `Lexicon` enthält einen einfachen Transduktor mit der Sprache {⟨fuß, fuß⟩, ⟨guss, guss⟩}, auf die Ersetzungen angewendet werden sollen.

Unter der Variable `Replace` wird ein Transduktor definiert, der beliebige Zeichenketten als Eingabe akzeptiert und dieselbe auch wieder ausgibt. Eine Ausnahme stellen die Zeichenketten dar, die ein ⟨ß⟩ enthalten. In diesen Fällen entsprechen ⟨ß⟩ in der Eingabe einem ⟨ss⟩ in der Ausgabe. Der Transduktor `Replace` enthält also Wörter wie ⟨heißen, heissen⟩, ⟨abc, abc⟩ oder ⟨hß6, hss6⟩. Neben der einfachen Ersetzung erlauben die Werkzeuge XFST und HFST auch eine optionale Ersetzung

und Einschränkungen auf einen bestimmten silbischen Kontext.

Die Komposition in `Compose` erlaubt es anschliessend, die Ersetzung in `Replace` auf das Lexikon in `Lexicon` anzuwenden. Bei der Komposition wird die Ausgabe des ersten Transduktoren als Eingabe an den zweiten Transduktor weitergegeben. ⟨fuß, fuß⟩ aus dem ersten Transduktor wird gepaart mit ⟨fuß, fuss⟩ aus dem zweiten Transduktor. Der Transduktor `Compose` enthält folglich das Paar ⟨fuß, fuss⟩. Ebenfalls wird ⟨guss, guss⟩ mit ⟨guss, guss⟩ gepaart zum neuen Paar ⟨guss, guss⟩. Weitere Wörter erkennt der neue Transduktor nicht, da `Lexicon` über keine weiteren Pfade verfügt. Zusätzliche mögliche Pfade in `Replace` werden vom Transduktor `Compose` nicht akzeptiert.

4.4 Formengenerierung

Wie es bei Finite-State-Morphologien üblich ist, ist auch dieses System wie ein Generierungswerkzeug für Wortformen aufgebaut. Während für standardisierte Orthographien ein Werkzeug für Analyse und Generierung verwendet werden kann, ist dieses System nur für eine Analyse ausgelegt – also in Gegenrichtung zur Generierung. Da bei der Analyse von Dialekten eine breite Abdeckung verschiedener Formen und Schreibweisen gewünscht ist, würde dies bei der Generierung zu einem uneinheitlichen Schriftbild führen, das schwierig zu lesen wäre.

Das Morphologieanalysesystem für Schweizerdeutsch beinhaltet als Kern eine abstrahierte Schreibung (siehe Tabellen 24 und 25 im Anhang), die über die verschiedenen regionalen Varietäten ein gemeinsames Phonemsystem abbilden. Zusätzliche Angaben historisch bedingten Unterschieden oder zur standarddeutschen Schreibung sollen erleichtern, die konkreten Formen davon zu bilden. Temporäre Markierungen zwischen Morphemgrenzen erlauben es, Ersetzungen gezielt auf die Stämme anzuwenden.

Da die Gruppe der häufig verwendeten Wörter klein ist, lassen sich die enthaltenen Wörter mit ihren Unregelmässigkeiten als Vollformen leicht auflisten, sodass sie nicht aufwendig bearbeitet werden müssen. Die Mehrheit der Lemmata, die jeweils eher selten auftreten, lässt sich aus Morphologieanalysesystemen für Standarddeutsch übernehmen. Diese Stämme lassen sich mit Hilfe von Ersetzungsregeln in schweizerdeutsche Stämme umformen und anschliessend analog zu den anderen Wörtern in dialektspezifische Lautformen überführen.

Abbildung 6 zeigt den schematischen Aufbau des Programms. Die folgenden Kapitel gehen auf die einzelnen Teile dieser Darstellung ein.

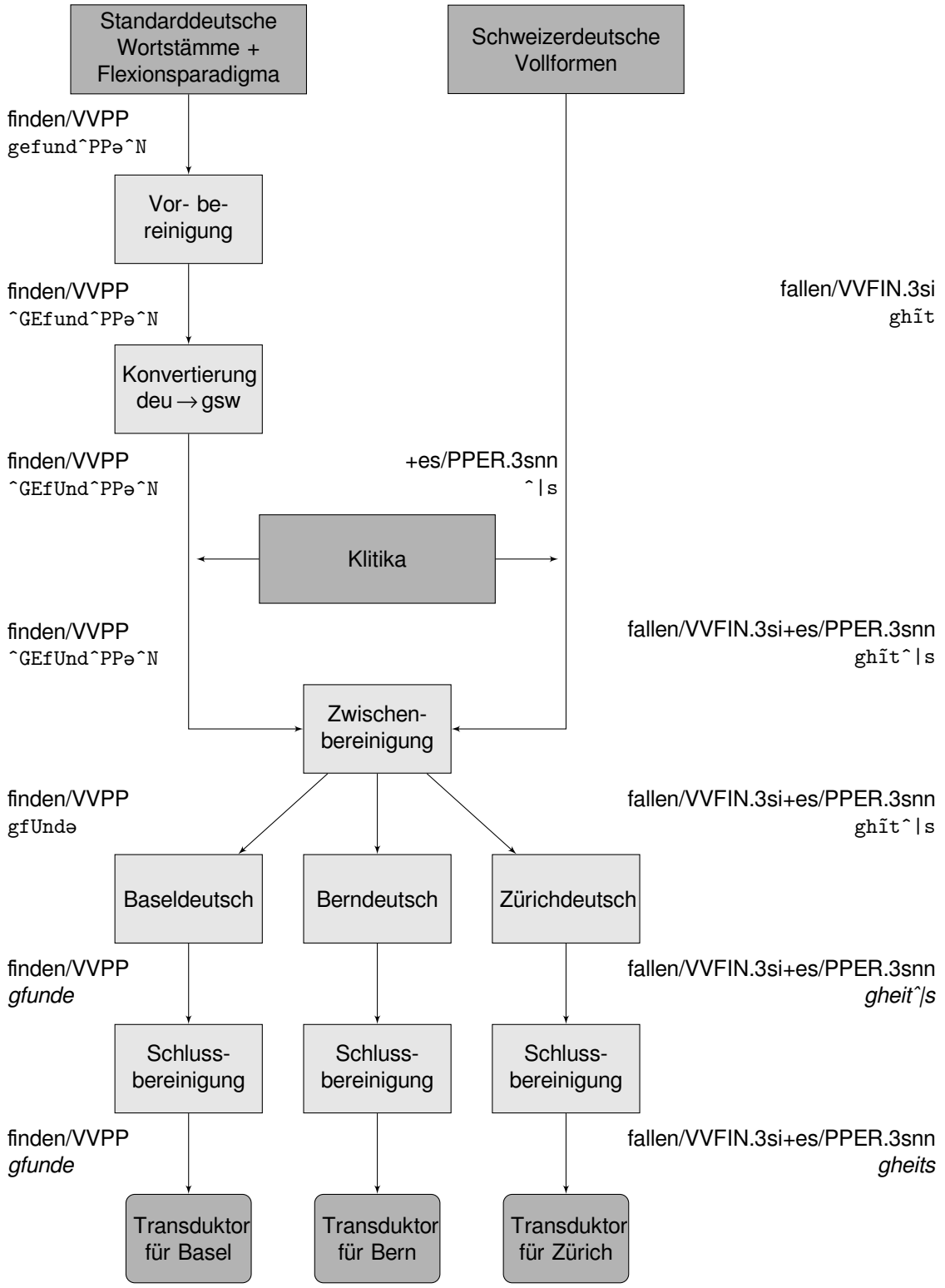

Abbildung 6: Übersicht über das Hauptskript `collection.xfst`. Links die Konversion eines standarddeutschen Stammes ins Baseldeutsche und rechts die Überführung eines schweizerdeutschen Stammes ins Zürichdeutsche. Wörterlisten in dunkelgrau und Ersetzungsregeln in hellgrau. Erklärungen zur abstrakten Schreibweise in Tabellen 24 und 25.

Der Umweg über den Zwischenschritt der abstrakten Schreibung (mit Schreibmaschinensatz ausgezeichnet) erleichtert es, das System an die Lautung anderer Dialekte anzupassen. Für die Generierung der konkreten Schreibungen eines bestimmten Dialekts können dialektspezifische Ersetzungsregeln verwendet werden. Diese sind einfach auszutauschen, wodurch dann die möglichen Formen und deren Gewichte einem anderen Dialekt entsprechen. Für die Wörter aus standarddeutschen Systemen und die vollgelisteten Wörter kann für jeden Dialekt derselbe Regelsatz verwendet werden.

4.4.1 Standarddeutsche Wortstämme

Für die offenen Wortklassen wurden die Stämme aus *Morphisto* (Zielinski et al. 2009) übernommen, die bereits nach Flexionsklassen geordnet sind und Allomorphe unregelmässiger Stämme mitenthalten. *Morphisto* beinhaltet rund 18 200 Stämme, die miteinander kombiniert werden können. Für das vorliegende System konnten die Substantivstämme (7833 Stämme), die Stämme für Namen (1052), Verbstämme (4300), Adjektivstämme (3178) und Adverbstämme (781) übernommen werden (Zahlen von Zielinski et al. 2009, S. 66). Zusammen mit den Stämmen der geschlossenen Wortklassen sollten die 30 000 häufigsten Wörter abgedeckt sein.

Die Auswahl der Stämme kann einfach erweitert werden. Bedingungen dafür sind lediglich, dass für die jeweiligen Stämme die Worthäufigkeitsklasse zum Lemma und die entsprechende Flexionsklasse bekannt sind, um die Wahl der richtigen Endungen sicherzustellen.

Abhängig von den Flexionsklassen werden zu den Stämmen die Flexionsendungen angefügt. Mit Hilfe der Flexionsklassen lassen sich dabei die Möglichkeiten für die Endungen für die schweizerdeutschen Formen eingrenzen. Beispielsweise werden schwachen Substantiven im Standarddeutschen schwache Endungen im Schweizerdeutschen zugewiesen.

Als Beispiel soll hier das Partizip des Verbs ‚finden‘ erklärt werden. In der Liste der Stämme stehen Paare wie ⟨find, `gefund`⟩. Diese Stämme werden mit ihren zugehörigen Endungen kombiniert. Da die Wahl der Endungen in flektierenden Sprachen wie Deutsch stark von den Stämmen abhängt, wurde auf eine Trennung von Stamm und Endung verzichtet. Aus ⟨find, `gefund`⟩ ergibt sich dann ein Paar ⟨finden/VVPP, `gefundˆPPəˆN`⟩.

4.4.2 Vorbereinigung

Bevor die standarddeutschen Wortstämme in schweizerdeutsche konvertiert werden, müssen noch einige Ersetzungen gemacht werden, die für alle Stämme eine gleiche Ausgangslage herstellen. Dazu gehören Löschregeln, welche für die Abtrennung standarddeutscher Endungen nötig sind, oder optionale Regeln, die für die Realisierung diverser äquivalenter Schreibungen notwendig sind.

Ein Beispiel für Zwischenschritte in Abbildung 6 ist die Ersetzung von `ge` durch das Mehrzeichensymbol `^GE`. Mehrzeichensymbole sind Zeichenketten, deren Zeichen alle an derselben Kante des Automaten oder Transduktoren liegen und so wie ein Zeichen behandelt werden. Dies wird üblicherweise für Tags und Hilfszeichen gemacht, die hier mit `^` beginnen. Da die unregelmässigen Partizipien bereits mit dem Präfix ⟨ge⟩ in den Daten von *Morphisto* enthalten sind, muss dieses ersetzt werden, um alle Verben später gleich behandeln zu können. Bei den anderen Verben ist in *Morphisto* ein Attribut vorhanden, welches direktes Setzen von `^GE` erlaubt.

4.4.3 Konvertierung ins Schweizerdeutsche

Aus sprachgeschichtlichen Gründen sind bei der Konvertierung in schweizerdeutsche Phoneme vor allem die Vokallaute betroffen, die vom Standarddeutschen – d. h. Neuhochdeutschen – in eine Form überführt werden müssen, welche dem Mittelhochdeutschen näher steht (vgl. Tabelle 5) und in einem besonderen Format dargestellt wird (siehe Tabellen 24 und 25 im Anhang). Eine der Entwicklungen, die dabei berücksichtigt werden muss, ist die Diphthongierung im Neuhochdeutschen. Dadurch wurden die hohen Langvokale /iː/ ⟨î⟩, /uː/ ⟨û⟩ und /yː/ ⟨iu⟩ zu /a͡ɪ/ ⟨ei⟩, /a͡ʊ/ ⟨au⟩ und /ɔ͡ʏ/ ⟨äu⟩ bzw. ⟨eu⟩ diphthongiert (vgl. Christen et al. 2012, S. 307). Mittelhochdeutsches *zît* und *hûs* wurde zu neuhochdeutschem *Zeit* und *Haus*. Das Schweizerdeutsche hat diese Entwicklung hingegen nicht mitgemacht und es heisst in den meisten Dialekten fortfahrend *Ziit* und *Huus*. Bei der Konvertierung ins Schweizerdeutsche muss diese Entwicklung also rückgängig gemacht und aus ⟨ei⟩ wieder ⟨ii⟩ werden.

Da die mittelhochdeutschen Diphthonge ⟨ei⟩, ⟨ou⟩ und ⟨eu⟩ als ⟨ei⟩, ⟨au⟩ und ⟨eu⟩ bzw. ⟨äu⟩ im Neuhochdeutschen weiterexistieren und somit mit den ehemaligen hohen Langvokalen zusammengefallen sind, müssen die neuhochdeutschen Diphthonge folglich bei der Konvertierung aufgeteilt werden. Für diese Entscheidung gibt es vom Standarddeutschen her keine Merkmale, die eindeutig auf die schweizerdeutsche Form hinweisen und homophone Wörter im Neuhochdeutschen können

dadurch auch im Schweizerdeutschen schlecht unterschieden werden.

Die zweite entscheidende Entwicklung, welche die Vokale betrifft, ist die neuhochdeutsche Monophthongierung (Christen et al. 2012, S. 310). Dabei wurden die mittelhochdeutschen Diphthonge ⟨ie⟩, ⟨uo⟩ und ⟨üe⟩ zu den hohen Langvokalen /iː/ (weiterhin geschrieben als ⟨ie⟩), /uː/ und /yː/. Mittelhochdeutsches *guot* und *vuoz* wurde dabei zu neuhochdeutschem *gut* und *Fuß*. Auch hier ist das Schweizerdeutsche mit *guet* und *Fuess* dem mittelhochdeutschen Lautstand näher.

Da Langvokale das Resultat der Monophthongierung im Neuhochdeutschen sind, können die Ersetzungsregeln von Hinweisen für die Vokallänge abhängig gemacht werden. Vor ⟨h⟩ und mindestens zwei Konsonanten und vor ⟨ß⟩[1] wird im Schweizerdeutschen für hohe Vokale zwingend ein Diphthong (`uə`, `ie` oder `üe`) gesetzt, da in geschlossenen Silben nur die Monophthongierung als Grund für die Vokallänge stehen kann.

In den offenen Silben, die im Standarddeutschen in der Regel lang sind, kann die Vokallänge einerseits durch die Monophthongierung und andererseits durch die Vokaldehnung in offener Silbe begründet werden. Vor einzelnen Konsonanten ist eine Längung ebenfalls möglich[2]. Auch hier könnten also Homonyme im Standarddeutschen zu Ambiguitäten führen. Zusätzlich können Lehnwörter in beiden Varietäten die gleichen Vokale aufweisen, wie es bei /kʊltuːr/ ‚Kultur' der Fall ist.

Bei den übrigen Vokalen in geschlossenen Silben kann hingegen allgemein von Kurzvokalen ausgegangen werden, die keine grossen Schwierigkeiten bereiten. Beispiele dafür sind *rund* ‚rund' in Tabelle 5 oder *gfunde* ‚gefunden' in Abbildung 6. Die verschiedenen silbischen Kontexte werden in der Implementierung nacheinander verarbeitet, wobei zuerst die spezifischen Fälle umgesetzt werden und zuletzt die einfachsten ohne Beachtung des Kontextes. Besondere Ausnahmen wie *Mueter* ‚Mutter' von mittelhochdeutsch *muoter* werden als unregelmässige Formen aufgelistet und müssen nicht durch aufwendige Regeln abgedeckt werden.

Die Kriterien zur Identifizierung der Diphthonge wie /uə̯/ können auch bei ⟨a⟩ und ⟨ä⟩ verwendet werden. So sollte die Verdumpfung zu [ɔː] wie bei *Johr* ‚Jahr' (vgl. Tabelle 1) nur beim mittelhochdeutschen Langvokal *â* möglich sein und auch beim alten Langvokal *æ* gibt es andere Lautungen als für den Kurzvokal *ä*.

Bei den Konsonanten dagegen steht das Standarddeutsche dem älteren Sprach-

[1] Ich verwende in den standarddeutschen Lemmata das scharfe S, obwohl dessen Gebrauch in der Schweiz unüblich ist. Kompatibilität zu anderen Systemen und die besser ableitbare Aussprache des Vokals davor begründen diese Entscheidung.

[2] In vielen Dialekten ist ein gelängter Vokal von anderer Qualität als ein überlieferter Langvokal. /tsʊːɡ̊/ ‚Zug' vs. /huːs/ ‚Haus'

deu	→	gsw	Beispiel			Kommentar	Mittelhochdeutsch
ei	→	ī	Zeit	–	Ziit		zît
ei	→	aĭ	heiß	–	heiss		heiz
u	→	uə	Fuß	–	Fuess	Langvokal in geschlossener Silbe	vuoz
u	→	U	Zug	–	Zug	gedehnter Vokal	zuc
u	→	uə	gut	–	guet	Monophthongierung	guot
u	→	ū	Kultur	–	Kultur	Lehnwort	-
u	→	U	rund	–	rund	Kurzvokal	runt

Tabelle 5: Ersetzungsregeln für Vokale vom Standarddeutschen zum Schweizerdeutschen. ⟨ei⟩ ist grundsätzlich ambig, während bei ⟨u⟩ die Ambiguität durch den silbischen Kontext begrenzt werden kann.

stand näher und die Ersetzungen können in die selbe Richtung wie die historische Entwicklung gemacht werden (vgl. Tabelle 6). Die beiden wichtigsten Ersetzungen sind die Verschiebung von ⟨k⟩ zu ⟨ch⟩ /x/ und der Ausfall von /n/. Die Verschiebung von ⟨k⟩ zu ⟨ch⟩ geschieht vor allem am Wortanfang, ist aber auch im Wort möglich. /n/ kann sowohl am Wortende oder vor Kompositionsgrenzen, aber auch in Wortstämmen ausfallen. Der Buchstabe ⟨n⟩ wird dabei jedoch nur durch einen Stellvertreter ersetzt, der auch bei den vollgelisteten Lemmata in Verwendung ist und separat behandelt wird.

deu	→	gsw	Beispiel			Kommentar	Mittelhochdeutsch
k	→	ch	Kopf	–	Chopf		kopf
k	→	k	Kanton	–	Kanton	jüngere Lehnwörter	-
st	→	št	fest	–	fescht		veste
n	→	ˆN	Stein	–	Stei/Stein	nach Vokal	stein

Tabelle 6: Ersetzungsregeln für Konsonanten vom Standarddeutschen zum Schweizerdeutschen. Standarddeutschem ⟨k⟩ können sowohl `ch` als auch `k` entsprechen. ⟨st⟩ und ⟨n⟩ können dagegen grundsätzlich ersetzt und später behandelt werden.

4.4.4 Vollformenlexika

Falls Lemmata der Klassen Adjektiv, Adverb, Substantiv oder Verb zwischen Standarddeutsch und Schweizerdeutsch deutlich abweichen, sind sie direkt in der abstrakten Schreibung (siehe Tabellen 24 und 25 im Anhang) erfasst. Beispiele wie ⟨Montag/NN.sm, `mān̄tig`⟩ zeigen, dass diese Vorgehensweise ein Aufblähen der Er-

setzungsregeln verhindert. Dieses Verfahren bietet sich auch für Wörter wie ⟨fallen/VVINF, ghĩə⟩ an, die nur auf Schweizerdeutsch existieren. Ebenso sind die übrigen Wortarten (Adpositionen, Artikel, Interjektionen, Konjunktionen, Pronomina, Partikeln und Zahlen) bereits als schweizerdeutsche Wortformen erfasst.

Den Extremfall der Volllistung sieht man beispielsweise bei den Artikeln oder Pronomina, wo alle Formen nacheinander aufgelistet sind und nur noch die Konvertierung in die spezifischen Dialekte erfolgen muss. Abbildung 7 zeigt einen Auszug aus dem Vollformenlexikon für die Artikel mit Paaren wie ⟨eine/ART.irsn, əs⟩.

```
! nominative/accusative
!! masc 'ein'/'einen'
eine/ART.irsm:əˆN       Final   "weight: 4.6" ;
eine/ART.irsm:nəˆN      Final   "weight: 4.6" ;
!! fem 'eine'
eine/ART.irsf:əˆN       Final   "weight: 4.8" ;
eine/ART.irsf:nəˆN      Final   "weight: 4.8" ;
!! neut 'ein'
eine/ART.irsn:əs        Final   "weight: 5.0" ;
eine/ART.irsn:nəs       Final   "weight: 5.0" ;
eine/ART.irsn:əˆN       Final   "weight: 5.0" ;
```

Abbildung 7: Auszug aus `articles.lexc`. Definition der Nominativ-/Akkusativformen des unbestimmten Artikels.

Vor allem bei Endungen, die einem bestimmten Muster folgen, gibt es auch hier eine Aufteilung von Stamm und Endung. Diese wird aber nur bei regelmässigen Mustern angewendet.

4.4.5 Klitika

Personal- und Reziprokpronomina, sowie das Indefinitpronomen ‚man' können als Klitika an finite Verbformen wie bei *hani* ‚habe ich', an Imperative wie bei *macheds* ‚macht es', an Konjunktionen wie bei *öbs* ‚ob es', an Pronomina wie bei *mers* ‚man es' und an einige Präpositionen wie in *bimer* ‚bei mir' angefügt werden. Ähnlich können auch Artikel an Konjunktionen klitisiert werden, beispielsweise *ondem* ‚und dem'. Jeweils separate Dateien pro Kasus für die Pronomina ermöglichen eine lexikographische Übersichtlichkeit bei allfälligen Erweiterungen.

Merkmalsübereinstimmung

Flag diacritics geben Informationen zum Subjekt des Satzes weiter und verhindern ungrammatische Kombinationen aus Verben und Pronomina. Die Beispiele

4.1 und 4.2 zeigen, wie die *flag diacritics* bei der Analyse des Tokens *machi* helfen. Das Personalpronomen ‚ich' kann an das Verb enklitisiert werden, wenn dieses in Person und Zahl übereinstimmt, was hier mit `@U.form.ich@` sichergestellt ist, wie in Beispiel 4.1. In Beispiel 4.2 tragen Verb und Personalpronomen einen unterschiedlichen Wert bei der Person und diese Variante wird als Analyse blockiert.

(4.1) mach -i
`@U.form.ich@` `@U.form.ich@`
machen/VVFIN.1si ich/PPER.1s*n
‚mache ich'

(4.2) * mach -i
`@U.form.du@` `@U.form.ich@`
machen/VVIMP.s ich/PPER.1s*n
‚mach ich!'

Da das *flag diacritic* `form` dem Subjekt des Satzes folgt, lässt es sich auch bei der Unterscheidung zwischen klitisiertem Personalpronomen und klitisiertem Reflexivpronomen verwenden, wie in den Beispielen 4.3 bis 4.6 zur Analyse des Tokens *machmer*. In Beispiel 4.3 stimmt das Pronomen in Zahl und Person mit dem Verb überein und ist somit als Reflexivpronomen zu bestimmen. Das konkurrierende Personalpronomen hingegen unifiziert nicht mit dem Verb und verhindert die Analyse wie in 4.4.[3] Demgegenüber darf ein Pronomen nicht als Reflexivpronomen annotiert werden, wenn es nicht unifiziert (wie in Beispiel 4.5), sondern es muss dann auf das Personalpronomen ausgewichen werden (wie in Beispiel 4.6).

(4.3) mach -mer
`@U.form.ich@` `@U.form.ich@`
machen/VVFIN.1si sich/PRF.1s*d
‚[ich] mache mir'

(4.4) * mach -mer
`@U.form.ich@` `@D.form.ich@`
machen/VVFIN.1si ich/PPER.1s*d
‚[ich] mache mir'

(4.5) * mach -mer
`@U.form.du@` `@U.form.ich@`
machen/VVIMP.s sich/PRF.1s*d

[3] Das `U` steht dabei für Unifikation mit den gleichen Werten. `form` ist der Name der Variable und `ich` oder `du` der Wert, mit dem sie belegt ist. Ein `D` gibt an, dass der zugehörige Wert für diese Variable nicht gesetzt sein darf.

,mach mir!‘

(4.6)

mach	-mer
`@U.form.du@`	`@D.form.ich@`
machen/VVIMP.s	ich/PPER.1s*d

,mach mir!‘

Ein zweites *flag diacritic* regelt das Zusammenspiel der Verben mit den Verbpartikeln. Dazu gehört die Verhinderung von Verbpartikeln bei Imperativen (wie im Standarddeutschen *mach vor!* statt *vormach!*) oder die Verhinderung der Klitisierung bei Anwesenheit einer Verbpartikel, was morphosyntaktisch unmöglich ist. Die Satzklammer verhindert nämlich die gleichzeitige Belegung der Stelle vor und nach dem Verb.

Bei den Konjunktionen und Präpositionen sind Klitika nur dann erlaubt, wenn entsprechende *flag diacritics* für die erlaubten Kasus gesetzt sind. Damit wird eine Übergenerierung verhindert. Für die Pronomina stellt das *flag diacritic* auch die Kongruenz bezüglich Kasus sicher. Da Artikel nur an Wörter ohne Kasus klitisiert werden können, reicht hier ein *flag diacritic* als Bedingung für eine Klitisierung. Die Kombination aus Präposition und Artikel (APPRART) gilt dagegen als eigene Wortart und wird auch ohne Klitisierung erledigt.

4.4.6 Zwischenbereinigung

Vor der Umwandlung in die Dialektformen werden die Mehrzeichensymbole `^GE`, `^PP` und `^N` eliminiert. Das Symbol `^N` kann dabei gelöscht werden oder durch `n` ersetzt werden, wie es bei einigen Schreibern vor Wörtern, die mit Vokal beginnen, geschrieben wird.

Die Symbole `^GE` und `^PP` können dagegen als Zirkumfix betrachtet werden, welches das richtige Präfix bei den Partizipien ermöglicht. Falls beide Präfixe zusammen vorkommen, wird `^GE` in Abhängigkeit mit dem nächsten Konsonanten durch das Präfix *ge-* in seinen verschiedenen Realisierungen im Schweizerdeutschen ersetzt. Sonst werden beide Symbole gelöscht.

4.4.7 Überführung in dialektspezifische Lautformen

Ausgehend von der abstrakten Schreibweise können verschiedene dialektspezifische Lautformen generiert werden. Die Anpassung durch weitere Dialekte sollte einfach zu bewerkstelligen sein, um das System für andere Dialekte als diejenigen von

Basel, Bern oder Zürich anpassen zu können.

Wie bei der Konversion vom Standarddeutschen ins Schweizerdeutsche betreffen die meisten Regeln für die dialektspezifischen Formen die Vokale. Mit einer Ausgangslage, die dem Mittelhochdeutschen nahe steht, können die meisten Entwicklungen mit einfachen Ersetzungsregeln nachgebildet werden. Wie in Tabelle 7 zu erkennen ist, geht es oft um Vorverlagerung von /uː/ oder um Entrundung von Umlauten. An der Verteilung der Phoneme ändert sich dabei jedoch wenig.

Phonem	Beispiel	Zürich	Wallis	Basel	Mittelhochdeutsch
ū	Haus	Huus	Hüüs	Huus	hûs
ǖ	Häuser	Hüüser	Hiischer	Hiiser	hiuser
ā	Jahr	Jaar	Jaar	Joor	jâr
ī	frei	frei	frii	frei	vrî

Tabelle 7: Dialektspezifische Lautformen. Bei ‚Haus' lässt sich die Vorverlagerung des Vokals im Wallis erkennen. ‚Häuser' zeigt die Entrundung des Vokals wie im Wallis und in Basel üblich. Der Vokal in ‚Jahr' wird in Basel verdumpft. ‚Frei' zeigt ausser in den höchstalemannischen Dialekten (z. B. Walliserdeutsch) die Hiatusdiphthongierung.

Aus der Verteilung der Vokale in Tabelle 7 lassen sich dialektspezifische Regeln zum Beispiel fürs Baseldeutsche (Tabelle 8) ableiten. Durch die feine Unterteilung der Phoneme in der abstrakten Darstellung ist bei den Ersetzungsregeln der Kontext nur bei dialektspezifischen Lautwandelphänomenen notwendig.

gsw	→	Basel	Beispiel			Kommentar	Standarddeutsch
ū	→	uu	hūs	–	huus		Haus
ǖ	→	ii	hǖsər	–	hiiser	Entrundung	Häuser
ā	→	oo	jār	–	joor	Verdumpfung	Jahr
ī	→	ei	frī	–	frei	Hiatusdiphthongierung	frei

Tabelle 8: Ersetzungsregeln für Vokale des Baseldeutschen. Die dialekttypischen Lautungen bzw. Schreibungen können hier berücksichtigt werden.

Abweichend vom mittelhochdeutschen Lautstand sind für die Hiatusdiphthongierung (siehe Christen et al. 2012, S. 308) von mhd. *vrî* zu *frei* selbständige Phoneme enthalten, die in den Dialekten ohne dieses Phänomen mit den Langvokalen zusammengeführt werden können.

Bei den Konsonanten sind die ziemlich regelmässigen Phänomene der Vokalisierung von /l(ː)/ (von mhd. *welt* zu *Wäut* [wæut] ‚Welt') um Bern und der Degemina-

tion von /lː/, /mː/ und /nː/ (von mhd. *stimmen* zu *stime* [ʃtimə] ‚stimmen') im Norden des Sprachgebiets zu erwähnen. Da diese einen neueren Sprachstand darstellen, können ihre lautgesetzlichen Entwicklungen direkt als Ersetzungsregeln in die dialektspezifischen Module übernommen werden.

4.4.8 Schlussbereinigung

Im letzten Bereinigungsschritt vor dem Ablegen in das binäre Transduktorenformat werden noch Klitisierungsgrenzen wie ^| gelöscht oder gewisse Laute, die alle Dialekte betreffen, behandelt. Ebenfalls unter diesen Punkt fallen die Schreibungen von /ʃ/ (⟨sch⟩, ⟨sh⟩ oder ⟨s⟩), die von den verschiedenen Schreibern unterschiedlich verwendet werden, sowie die Gross- und Kleinschreibung für Substantive und für die übrigen Wortarten am Satzanfang. Zusätzlich gehören auch Regeln dazu, welche Varianten bei der Schreibung von Umlauten (⟨ö⟩ oder ⟨oe⟩) und Apostrophen (gerade, gebogen) berücksichtigen und die Ersetzung durch äquivalente Schreibungen zulassen.

4.5 Gewichte

Mit den umfangreichen Ersetzungen bei der Anpassung an die schweizerdeutschen Lautformen steigt auch die Zahl möglicher Analysen pro Wort an. Gewichte sind eine Möglichkeit, sich wieder einen Überblick darüber zu verschaffen, indem sie eine Ordnung der Analysen ermöglichen. Ziel ist es dabei, möglichst oft die richtige Analyse an erster Stelle zu haben und ungewollte oder unwahrscheinliche Analysen durch Strafpunkte an das Ende der Rangliste zu verschieben.

Als Grundlage für die Gewichtung in *Taggswort* dienen Worthäufigkeitsklassen für Grundformen. Mit ihnen lassen sich die am häufigsten verwendeten Wörter bevorzugen, womit man in möglichst vielen Fällen die richtige Grundform unter den ersten Möglichkeiten hat. Das Schema wird aber auch auf die Gewichte der grammatischen und der lautlichen Formen übertragen. Die grammatischen Kategorien einer Wortart treten in einem Text unterschiedlich oft auf und entsprechend können damit Gewichte berechnet werden, welche die selteneren Kombinationen zugunsten der häufigsten bestrafen. Indem die Gewichte für die Formen zur Worthäufigkeitsklasse addiert werden, können die häufigsten Lemmata, sowie ihre häufigsten Formen, bei der Priorisierung berücksichtigt werden.

Um die Lautentsprechungen miteinfliessen zu lassen, werden Lautentsprechungen,

die in einem Kontext selten vorkommen, bestraft und solche Analysen in der Rangfolge nach unten verschoben. Auch für diese Gewichte, die zu den vorherigen addiert werden, bilden Häufigkeitsverteilungen die Grundlage.

4.5.1 Worthäufigkeitsklassen

Basis der Worthäufigkeitsklassen der Lemmata ist das Verhältnis der Häufigkeit eines bestimmten Lemmas zur Häufigkeit des frequentesten Lemmas. Dieses ist im Standarddeutschen der bestimmte Artikel. Die anschliessende Logarithmisierung sorgt für ein leserlicheres Zahlenformat und gibt im Zusammenhang mit der Rundung auf die nächste Ganzzahl die Worthäufigkeitsklasse. Die Häufigkeitsklasse HK eines Wortes wird also durch folgende Formel ermittelt (siehe IDS 2012, Benutzerdokumentation zu DeReWo S. 7):

$$HK(\text{Wort}) := \lfloor log_2 \left(\frac{f(\text{häufigstesWort})}{f(\text{Wort})} \right) + 0,5 \rfloor$$

Das bedeutet, dass das häufigste Wort etwa $2^{HK(wort)}$-mal so oft auftritt wie das Wort selbst. Dieser Vergleich ist natürlich mit jedem Wortpaar möglich, was dann interessant wird, wenn man zwei Lemmata als alternative Analysen für eine Form vergleichen möchte.

Mangels genügend grosser Korpora für Schweizerdeutsch werden hier die Worthäufigkeitsklassen des Standarddeutschen aufs Schweizerdeutsche transferiert. Selbst wenn die unterschiedlich häufige Verwendung gewisser Wörter ein Dialektmerkmal darstellen kann, sollten doch beide Varietäten einander genug ähneln, um für eine Ordnung der möglichen Analyse nützlich zu sein.

Diese Worthäufigkeitsklassen sind bereits in den aus *Morphisto* wiederverwendeten Daten[4] enthalten und müssen lediglich bei den manuell eingefügten Lemmata aus der Wortgrundformenliste *DeReWo* (IDS 2012) ergänzt werden.

4.5.2 Gewichtung der Formen

Damit nicht nur die gebräuchlichsten Wörter, sondern auch deren gebräuchlichsten Formen bei der Analyse bevorzugt werden, erhalten die Formen eine Gewichtung

[4] Online verfügbar auf `https://github.com/GreatStuff660/morphisto/blob/master/src/basestems.xml` und `https://github.com/GreatStuff660/morphisto/blob/master/src/adverbien.xml` (aufgerufen am 23. Februar 2016)

anhand ihrer grammatischen Kategorien. Dazu wurden die Häufigkeitsverteilungen im Entwicklungskorpus verwendet. Ein Transfer von Daten aus dem Standarddeutschen ist durch die Unterschiede in der Grammatik nur schwer möglich und wurde deshalb als Option verworfen. Beispielsweise würde das Fehlen des Genitivs im Schweizerdeutschen eine starke Verschiebung in der Häufigkeit der anderen Kasus bewirken.

Die verschiedenen Formen sind anhand der Verteilung ihrer grammatischen Kategorien in Abhängigkeit der Wortart gewichtet. Die Gewichte w der einzelnen Formen berechnen sich aus dem Anteil, wie oft die entsprechende Wortart in dieser Form vorkommt. Aufbauend auf dem binären Logarithmus des Kehrwertes ist diese Zahl kompatibel mit der Gewichtung nach Häufigkeitsklassen und kann zu jener addiert werden. Die Formel für die Häufigkeitsklassen wird entsprechend abgeändert:

$$w(\text{Kategorien}|\text{Wortart}) := log_2 \left(\frac{f(\text{Kategorien}|\text{Wortart})}{f(\text{Wortart})} \right)$$

Tabelle 9 zeigt, wie die Formen des bestimmten Artikels ‚die' und des unbestimmten Artikels ‚eine' verteilt sind. Da der definite und der indefinite Artikel unterschiedlich viele Formen haben, wurden sie wie unterschiedliche Wortarten behandelt.

	definit		indefinit	
Morphologisches Tag	Vorkommen	Gewichtung	Vorkommen	Gewichtung
drsm / irsm	177	2,3	70	1,6
drsf / irsf	215	2,1	61	1,8
drsn / irsn	102	3,1	55	2,0
drp*	103	3,1	-	-
ddsm / idsm	39	4,5	5	5,4
ddsf / idsf	161	2,5	15	3,8
ddsn / idsn	19	5,6	7	4,9
ddp*	77	3,5	-	-
Total	893		213	

Tabelle 9: Gewichtung der Artikelformen. Die verschiedenen Formen der Artikel sind entsprechend der Verteilung ihrer grammatischen Kategorien innerhalb des Lemmas gewichtet. Ein Gewicht 2,3 bedeutet, dass die ensprechende Form einmal in $2^{2,3}$ vorkommt. Diese Gewichte werden anschliessend zum Gewicht (also der Häufigkeitsklasse) des Lemmas addiert.

Bei Inhaltswörtern muss im Gegensatz zu den Artikelwörtern von einer Unabhängigkeit zwischen dem Lemma und den Häufigkeiten dessen Realisierungen ausgegangen werden. Die Gründe dafür liegen im Aufbau des Programms, in dem die Endungen separat von den Stämmen vorliegen und in der Verfügbarkeit sprachlicher Ressourcen. Tabelle 10 zeigt, wie die Gewichte für Singular und Plural bei maskulinen Substantiven aussehen. Diese Gewichte lassen sich bei den Substantiven, welche diesen Kategorien folgen, zum Gewicht des Lemmas – das ja der Häufigkeitsklasse entspricht – addieren. Falls bei einem entsprechenden Wort Singular und Plural formengleich sind, gilt die Singularform als wahrscheinlicher, unabhängig davon, was beim Wort selbst für eine Verteilung vorherrscht.

Morphologisches Tag	Vorkommen	Gewichtung	‚Stein' (HK 11)	‚Wald' (HK 10)
sm	497	0,5	*Stei* 11,5	*Wald* 10,5
pm	214	1,7	*Stei* 12,7	*Wälder* 11,7
Total	711			

Tabelle 10: Gewichtung der Substantivformen mit Beispielen. Die Verteilung von Singular und Plural bei maskulinen Substantiven bestimmt die Gewichtung des entsprechenden Numerus.

Während bei den formenreichen Artikeln ein grosses Vorkommen vorliegt und bei den Substantiven die Zahl der Formen sehr klein ist, braucht es beispielsweise bei den Verben eine Möglichkeit, die Fülle an Formen zu behandeln. Mit der Annahme von Unabhängigkeit zwischen gewissen Kategorien kann man diese separat auszählen und anschliessend wieder miteinander kombinieren, wobei dann ungesehene Kombinationen grösstenteils vermieden werden können. Bei den Verben wurde darum für den Modus (inkl. Konjunktive, finite und infinite Kategorien) auf der einen Seite und für die Kombinationen aus Person und Zahl auf der anderen Seite Gewichte berechnet, die anschliessend addiert werden können. Phänomene wie Ersatzinfinitive anstelle von Partizipien und die verschieden gebräuchlichen Konjunktivformen waren Grund für eine separate Berechnung der Gewichte für den Modus der Voll-, Hilfs- und Modalverben.

4.5.3 Lautentsprechungen

Auch bei den Lautersetzungen sollen ungewollte Analysen durch Gewichte bestraft werden. Ziel davon ist es, falsche Freunde als Vorschläge zu vermeiden.

Die meisten Ersetzungsregeln für die Überführung zu schweizerdeutschen Phonemen greifen nur in einem bestimmten Umfeld. Mit Hilfe der Grundformenliste

DeReWo als Datensatz und den entsprechenden schweizerdeutschen Lautungen konnten den verschiedenen Ersetzungsalternativen in einem Kontext Häufigkeiten zugewiesen werden. Aus diesen Zahlen wurden dann, analog zur Gewichtung der Formen, Gewichte für die Ersetzungen berechnet.

```
# U
# long vowels: usually before ß and before hCC
define VRuleU1 [ {uß} -> {uəss}     ];
define VRuleU2 [ {u}  -> {uə}       || _ {h} Consonant Consonant ];

# long or short vowels: in front of max. 1 Cons
define VRuleU3 [ {u} (->) {ū}::3.5  || _ ({h}) (Consonant|{ch}) [.#.|Vowel] ];
define VRuleU4 [ {u} (->) {uə}::2.4 || _ ({h}) (Consonant|{ch}) [.#.|Vowel] ];
define VRuleU5 [ {u}  ->  {U}::0.5  || _ ({h}) (Consonant|{ch}) [.#.|Vowel] ];

# short vowel as default
define VRuleUdef [ {u} -> {U}       || _ \[ə] ];

# All rules for 'u'
define VRuleU [ VRuleU1 .o. VRuleU2 .o. VRuleU3 .o.
                VRuleU4 .o. VRuleU5 .o. VRuleUdef ];
```

Abbildung 8: Ersetzung von standarddeutschem ⟨u⟩ (Vereinfachtes Code-Beispiel). `VRuleU1` und `VRuleU2` lassen keine Wahl zu und tragen darum keine Gewichte. Bei `VRuleU3`–`VRuleU5` dagegen gibt es nicht genug Informationen und verschiedenen Möglichkeiten werden anhand ihrer Wahrscheinlichkeit gewichtet.

Als Beispiel soll hier auf die Ersetzung des standarddeutschen ⟨u⟩ (siehe auch Tabelle 5) eingegangen werden. Als allererstes wird ⟨u⟩ in den Diphthongen ⟨au⟩, ⟨äu⟩ und ⟨eu⟩ behandelt und von den weiteren Ersetzungen nicht mehr beeinflusst. Das weitere Verfahren mit ⟨u⟩, welches in Abbildung 8 dargestellt ist, soll nun erklärt werden. Wie in Kapitel 4.4.7 erklärt wurde, kann ⟨u⟩ vor ⟨ß⟩ nur einem Diphthong entsprechen – das gleiche gilt auch vor ⟨h⟩ und zwei weiteren Konsonanten. Da diese Regeln also zwingend sind, kann eine Gewichtung ausbleiben und die Ersetzung von ‚Fuß‘ nach `fuəss` wird nicht bestraft.

Vor maximal einem Konsonanten (allfälliges ⟨h⟩ zur Markierung der Länge ausgeschlossen) besteht diese Eindeutigkeit nicht. Die Möglichkeiten eines Langvokals (wie in ⟨Kultur/NN.sf, `kUltūr`⟩), eines Diphthongs (wie ⟨gut/ADJD.p, `guət`⟩) oder eines gelängten Vokals (wie in ⟨Zug/NN.sm, `zUg`⟩) mussten ausgezählt werden, um darauf eine Gewichtung aufzubauen. Dass bei `ū` das höchste Gewicht steht, spiegelt den Umstand wider, dass dieser Laut selten einem ⟨u⟩ in diesem Umfeld entspricht.

In allen anderen Kontexten muss davon ausgegangen werden, dass standarddeutsches ⟨u⟩ kurz ist und im Schweizerdeutschen `U` wie in ⟨rund/ADJD.p, `rUnd`⟩ aufweist. Deshalb kann auf eine Gewichtung verzichtet werden. Der Ausschluss von

`ə` danach stellt sicher, dass die Diphthonge nicht nochmals behandelt werden.

4.5.4 Gewichtung der Dialektformen

Bei der Konvertierung von der abstrakten Darstellung des Schweizerdeutschen in die verschiedenen Dialekte sind die Gewichte von geringerer Bedeutung. Da die abstrakte Darstellung bei den meisten Dialekten die Phoneme abbildet, braucht es nur oberflächliche Änderungen, die wenig Konfliktpotenzial bergen. Lediglich bei einzelnen dialektspezifischen Anpassungen der Phoneme, wie zum Beispiel der Fortisierung im Zürichdeutschen (von `dUnkəl` zu *tunkel* ‚dunkel') ist von einem Nutzen der Gewichtung auszugehen.

Bei der Mehrheit der Ersetzungen handelt es sich um verschiedene Schreibungen für die gleiche Lautung. Damit ist die Gefahr von Verwechslungen mit anderen Wörtern gering. Ob im Zürichdeutschen für `ī` nun ⟨ii⟩ oder ⟨y⟩ wie beispielsweise in *Ziit*/*Zyt* ‚Zeit' gewählt wird, hängt vor allem von den Gewohnheiten des Schreibers oder der Schreiberin ab. Schreibungen wie ⟨yy⟩, die im Zürichdeutschen eher eine Randerscheinung sind, brauchen keine Gewichte. Dies lässt sich durch die Betrachtung aus der Gegenrichtung leicht begründen: Da hinter geschriebenem ⟨yy⟩ nur das Phonem `ī` stehen kann, wären bei einem allfälligen Vorkommen alle Analysen gleich stark bestraft und die Rangfolge unverändert.

Die Variante ⟨ie⟩ für `ī` ist dagegen pauschal mit einem hohen Wert bestraft. Grund dafür ist, dass in den meisten Dialekten ⟨ie⟩ für /iə̯/ steht. Wie gross die Verwechslungsgefahr ist und ob das Gewicht nicht zu hoch ist, konnte im Umfang dieser Arbeit nicht berechnet werden. Benötigt wäre für eine solche Aufgabe ein Korpus, das mit Phonemen annotiert ist.

4.6 Verwendung

Die mit `collection.xfst` (siehe Abbildung 6) erstellten binären Transduktordateien lassen sich nach ihrer Erstellung unabhängig vom restlichen System verwenden.

Verschiedene Programme um HFST[5] sind zur Arbeit mit den binär gespeicherten Transduktoren ausgelegt. Zur einfachen Analyse ist das Skript `hfst-lookup` mit der binären Datei als Argument zu verwenden. Für eine Integration in Program-

[5]Erhältlich via `https://github.com/hfst` (aufgerufen am 3. April 2016)

miersprache wie Python oder Java gibt es die Pakete `hfst-optimized-lookup-python` beziehungsweise `hfst-optimized-lookup-java`.

Kombinationen aus verschiedenen dialektspezifischen Transduktoren können mit dem Skript `hfst-disjunct` erstellt werden. Für weitere Operationen gibt es entsprechende Programme.

5 Evaluation

Das Ziel einer Finite-State-Morphologie ist einerseits eine möglichst breite Abdeckung der Formen in den Korpora, für deren Analyse das System entwickelt wurde. Andererseits sollen die vorgeschlagenen Analysen für ein Wort möglichst korrekt sein.

Als Ursachen für ein suboptimales System nennen Beesley und Karttunen (2003, S. 313 – 319) die *sins of omission*, die zu einer fehlenden Analyse führen (falsch negativ), und die *sins of commission*, die eine falsche Analyse zur Folge haben (falsch positiv). Fehlende Analysen ergeben sich dadurch, dass Wortstämme dem System nicht bekannt sind oder dass sie inkorrekt verarbeitet werden. Mit der fehlenden korrekten Analyse senkt dies die Ausbeute. Als Untergruppe dazu gibt es die *sins of partial omission*, bei der nicht alle erwarteten Analysen gegeben werden können.

Die falschen Negativen zeigen, dass zwar Analysen gemacht werden, aber viele davon inkorrekt sind. Das kann bedeuten, dass einzelne Wörter irrtümlicherweise als andere Wörter lemmatisiert werden oder dass übergeneriert wird. Durch die umfangreichen Ersetzungsregeln im vorliegenden System existiert eine Übergenerierung von Formen, deren Auswirkungen aber durch eine gewichtete Rangordnung abgefangen werden kann. Probleme ergeben sich nur, wenn falsche Analysen rangmässig vor den richtigen kommen, das heisst, wenn sie präferiert sind. Solche Fehler senken als falsche Positive die Präzision des Systems oder vermindern die Rankingqualität.

Während Tokens ohne Analyse sofort auffallen, muss für das Auffinden einzelner fehlender und inkorrekter Analysen die Ausgabe des Systems mit einem Goldstandard abgeglichen werden.

Als Goldstandard dient hier ein eigens vom Verfasser annotierter Teil des NOAH-Korpus mit 14 888 Tokens. Da das Testkorpus repräsentativ für das gesamte NOAH-Korpus ist, kann von den Resultaten auf den Testdaten über das ganze Korpus sowie über dessen einzelne Teile generalisiert werden.

5.1 Abdeckung

Dass *sins of omission* andere Auswirkungen haben als *sins of partial omission* und *sins of commission* wurde dadurch berücksichtigt, dass für die Abdeckung nicht nur der prozentuale Anteil der Tokens, für die eine Analyse gemacht werden kann, berechnet wurde. Die wichtigste Frage ist nämlich, wie viele Tokens korrekt analysiert werden können. Diese Zahl wird im folgenden Kapitel jeweils unter *Tokens mit korrekter Analyse* aufgeführt.

Für den Vergleich mit anderen Systemen werden auch die Fragen gestellt, wie viele Tokens und wie viele Types analysiert werden können. Die Antworten dazu stehen in den Tabellen jeweils unter *Tokens mit Analyse* und *Types mit Analyse*.

5.1.1 Analyse auf dem kompletten Testkorpus

Für wie viele der rund 15 000 Tokens im Testkorpus konnte die korrekte Analyse gefunden werden? Einschliesslich aller Wortarten sind dies 79%. Schliesst man dabei die im Morphologieanalysesystem nicht behandelten Wortarten wie Eigennamen, fremdsprachiges Material und nicht-sprachliche Elemente aus, steigt dieser Anteil auf 86%. Tabelle 11 zeigt, dass besonders bei der Adaption fürs Berndeutsche der Anteil hoch ist und fast an die Kombination der Systeme fürs Basel-, Bern- und Zürichdeutsche herankommt.

Transduktor		Tokens mit korrekter Analyse	Tokens mit Analyse	Types mit Analyse
Basel	alle Wortarten	0,786	0,850	0,686
	ohne FM, NE, XY	0,856	0,898	0,742
Bern	alle Wortarten	**0,789**	0,852	0,691
	ohne FM, NE, XY	**0,859**	0,902	0,747
Zürich	alle Wortarten	0,785	0,847	0,681
	ohne FM, NE, XY	0,855	0,897	0,738
Basel+Bern	alle Wortarten	**0,790**	0,854	0,695
+Zürich	ohne FM, NE, XY	**0,860**	0,902	0,750

Tabelle 11: Abdeckung auf dem Testkorpus (relativer Anteil) nach dialektspezifischen Analysewerkzeug. Resultate für alle im Korpus auftretenden Tokens und Resultate ohne die Wortarten FM, NE und XY, die nicht behandelt wurden.

Die andere Frage dieser Evaluation ist, wie viele Tokens eine Analyse bekamen. Der

Anteil der Tokens mit mindestens einer Analyse liegt mit zirka 85% bzw. 90% für die behandelten Wortarten noch höher. Der Leistungsunterschied entsteht durch Tokens, für die zwar eine Analyse gemacht wurde, die korrekte aber Analyse nicht vorliegt. Erwartungsgemäss sinkt der Anteil der Wörter mit mindestens einer, aber keiner korrekten Analyse mit dem Ausschluss von Eigennamen und Wortklassen, die nicht zur behandelten Sprache gehören.

Zum Vergleich ist es interessant zu untersuchen, wie viele der Tokens im Testkorpus einem gängigen Tagger für Standarddeutsch bekannt sind. Für diesen Test wurden die Lemmata aus dem Goldstandard als unabhängige Wörter an den *TreeTagger* (Schmid 1995) übergeben und überprüft, ob dem Lemma aus dem Goldstandard ein Lemma durch den *TreeTagger* zugewiesen werden konnte. Mit dem *TreeTagger*-Modell für Standarddeutsch[1] erhalten 90% bzw. 94% Tokens eine Analyse. Diese Zahl ist zwar höher als mit dem vorliegenden System, doch der *TreeTagger* musste die Zuordnung zwischen schweizerdeutschen Wörtern und standarddeutschen Lemmata und die Verarbeitung der Wortformen nicht leisten.

Analyse nach Wortarten

Ein Überblick zur Abdeckung bei den einzelnen Wortarten soll nun zeigen, wo das System am meisten an Abdeckung einbüsst. Tabelle 12 zeigt den Anteil der korrekt erkannten Tokens und den Anteil, den diese Tokens im Korpus einnehmen.

Wortart	Tokens mit korrekter Analyse	Tokens mit Analyse	Types mit Analyse	Anzahl Tokens	Anzahl Types
ADJA	*0,662*	0,752	0,705	715	546
ADJD	0,723	0,814	0,771	328	253
ADV	0,901	0,949	0,872	970	335
APPO	*0,500*	1,000	1,000	4	3
APPR	**0,959**	0,982	0,898	909	108
APPRART	**0,970**	0,987	0,948	535	116
APZR	**1,000**	1,000	1,000	1	1
ART	**0,971**	0,996	0,952	1086	62
CARD	0,807	0,861	0,770	259	135
FM	*0,000*	0,590	0,528	183	127
ITJ	*0,194*	0,355	0,273	31	22

weiter auf der nächsten Seite

[1] Online verfügbar auf `www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` (aufgerufen am 5. Februar 2016)

Wortart	Tokens mit korrekter Analyse	Tokens mit Analyse	Types mit Analyse	Anzahl Tokens	Anzahl Types
KOKOM	**0,986**	1,000	1,000	69	7
KON	**0,992**	0,994	0,920	513	25
KOUI	0,938	1,000	1,000	16	5
KOUS	**0,970**	0,992	0,984	131	61
NE	*0,129*	0,392	0,363	1021	615
NN	*0,583*	0,668	0,600	2395	1752
PAV	0,870	0,935	0,900	46	30
PDAT	0,907	0,973	0,913	75	23
PDS	**0,963**	0,988	0,917	81	12
PIAT	0,873	0,937	0,891	79	46
PIDAT	**1,000**	1,000	1,000	11	4
PIS	0,944	0,986	0,963	142	54
PPER	**0,980**	0,995	0,982	395	54
PPOSAT	0,936	0,994	0,985	155	68
PRELS	**0,957**	0,979	0,917	94	12
PRF	0,948	0,987	0,909	77	11
PTKA	0,923	1,000	1,000	13	5
PTKAM	**1,000**	1,000	1,000	1	1
PTKANT	0,889	0,889	0,778	18	9
PTKINF	0,917	1,000	1,000	12	4
PTKNEG	**1,000**	1,000	1,000	78	12
PTKVZ	0,881	0,985	0,971	67	35
PTKZU	0,940	1,000	1,000	50	7
PWAV	0,889	0,963	0,944	27	18
PWS	**1,000**	1,000	1,000	29	11
TRUNC	*0,000*	0,083	0,083	12	12
VAFIN	**0,980**	0,996	0,978	816	90
VAINF	0,864	1,000	1,000	22	11
VAPP	0,943	0,968	0,786	158	14
VMFIN	0,943	0,989	0,978	87	46
VMINF	**0,967**	1,000	1,000	30	12
VVFIN	0,881	0,943	0,926	404	310
VVIMP	*0,667*	0,889	0,875	9	8
VVINF	0,848	0,899	0,880	197	166
VVIZU	*0,500*	0,625	0,571	8	7
VVPP	0,851	0,891	0,856	484	361
XY	*0,000*	0,061	0,130	165	46

weiter auf der nächsten Seite

Wortart	Tokens mit korrekter Analyse	Tokens mit Analyse	Types mit Analyse	Anzahl Tokens	Anzahl Types
$(	**0,986**	0,986	0,750	347	12
$,	**1,000**	1,000	1,000	568	1
$.	**0,999**	0,999	0,875	965	8

Tabelle 12: Abdeckung nach Wortarten im Testkorpus. Entsprechend dem Testkorpus mit verschiedenen Dialekten wurde eine Kombination der Systeme für Basel-, Bern- und Zürichdeutsch verwendet.

Auf die Abdeckung der korrekten Analyse auf allen Wortarten haben die Substantive den grössten Einfluss. Die nicht richtig erkannten Substantive (NN) machen 6,7% aller Tokens im Testkorpus aus. Mit 6,0% folgen die Eigennamen (NE), deren Behandlung allerdings nicht Kern dieser Arbeit war. An nächster Stelle folgen die Adjektive, deren nicht erkannte Formen 1,6% (für ADJA) und 0,6% (für ADJD) ausmachen. Die nicht erkannten Formen von Vollverben machen zusammen rund 1% aus, wobei die Hälfte davon Partizipformen sind. Die nicht behandelten Kategorien für fremdsprachiges Material (FM) und Nichtwörter (XY) machen 1,2% bzw. 1,1% des Korpus aus.

Eine Fehleranalyse sollte sich auf die Substantive und Adjektive und in geringerem Mass auch auf die Verben konzentrieren. Ad hoc gebildete Komposita wie *Dracheriter* ‚Drachenreiter' sind als Problem zu erwarten, da die Implementierung der Substantive keine Komposita zulässt und sich einzig auf die lexikalisierten Komposita wie *Bundesrepublik* abstützen kann, die oft vorkommen.

Dass die von *Morphisto* übernommenen Daten eher eine knappe Abdeckung ergeben, zeigt sich auch bei der Anwendung des *TreeTaggers* auf die nicht korrekt analysierten Formen der Substantive. Von 999 Substantiven wurden 156 vom *TreeTagger* als solche erkannt und mit Lemmata versehen. Für zusätzliche 53 wurde eine Analyse gegeben, jedoch mit der falschen Wortart. Bei den attributiven Adjektiven konnte für 37 aus 242 eine Analyse mit Lemma gegeben werden, 15 davon als Adjektiv. Unter den 91 nicht richtig erkannten prädikativen Adjektiven konnten 22 analysiert werden, davon 19 Mal mit der korrekten Wortart.

Andere Gründe für nicht erkannte Formen liegen in spezifischen Dialektformen. Mit der Konzentration auf die Dialekte zwischen Basel, Bern und Zürich sind die alpinen Dialekte im Süden der Schweiz nicht behandelt. Walliserdeutsche Formen wie *Gscheich* ‚Geschenk' konnten beispielsweise nicht analysiert werden, da die dahinter liegenden Lautwandelphänomene den im System behandelten Dialekten fremd sind. Auch ostschweizerische und bündnerdeutsche Formen wie *schnellor* ‚schneller' oder *Szena* ‚Szene' blieben deshalb unerkannt.

Die beiden Untergruppen der Vollverben mit der schlechtesten Abdeckung VVIMP und VVIZU treten im Korpus selten auf. Durch Verbpartikeln erweitertes *sein*, das damit zum Vollverb wird (etwa *zämezsii* ‚zusammenzusein'), ist sehr selten. Die Inspektion der unerkannten Infinitive zeigt, dass davon knapp ein Drittel Partikelverben wie *schtillschtah* ‚stillstehen' und *fithalte* ‚fithalten' sind. Wenig darüber liegen die Partizipien, die wie die Infinitive die Partikel stets an sich binden.

Ebenfalls eine relativ schlechte Quote weisen Postpositionen, Interjektionen und Kardinalzahlen auf. Ein Blick auf die unerkannten Formen zeigt allerdings, dass bei den Postpositionen bloss ein Type, der zweimal auftritt, nicht korrekt analysiert werden konnte. Bei den Kardinalzahlen fallen Dezimalzahlen auf und die Erweiterung einiger Zahlen durch ein ⟨i⟩ wie *12i* ‚12'/‚12 Uhr'/‚12 Jahre'. Für die Interjektionen ist eine breite Variation der Schreibungen auszumachen, die besondere Aufmerksamkeit benötigt.

Erstglieder bei elliptischen Komposita (TRUNC) sind dagegen ein Punkt, der bei einer zukünftigen Behandlung der Kompositabildung berücksichtigt werden kann. Die tiefe Erkennungsquote bei diesen seltenen Wortarten fällt aber im Gegensatz zu den Substantiven nicht ins Gewicht.

5.1.2 Analyse nach Textgattungen

Zwischen den verschiedenen Teilen des Testkorpus zeigen sich deutliche Unterschiede bezüglich der Abdeckung. Wie Tabelle 13 zeigt, sind die Zahlen im Teil *Blogs* am besten, während sie im Teil *Swatch* am tiefsten sind.

Während der Vergleich der Anzahl Types mit der Anzahl Tokens für die *Blog*-Daten auf einen beschränkten Wortschatz hindeutet, weisen die Daten von *Swatch* bei vergleichbarer Textlänge ein breiteres Vokabular auf.

Ähnliches gilt in den Teilen *Blick* und *Schobinger*, die ebenfalls eine ähnliche Textlänge aufweisen. Die Texte von *Schobinger* weisen ein weniger breites Vokabular auf als die von *Blick*, wobei aber der Unterschied in der Abdeckung zwischen diesen beiden Texten kleiner ist.

Beim Vergleich mit *TreeTagger* über die Tokens ohne FM, NE und XY schneidet das System bei den *Blog*-Texten besser ab. Mit 94,4% der Tokens mit Analyse liegt es über den 93,0% von *TreeTagger*. Bei den Texten von *Schobinger* liegt das System ein wenig unter den 95,1% des *TreeTaggers*. Schlechter ist es bei den *Blick*-Daten (*TreeTagger*: 95,0%), bei den *Wiki*-Daten (*TreeTagger*: 94,3%) und bei den *Swatch*-Daten (*TreeTagger*: 94,4%).

Teil		Tokens mit korr. Analyse	Tokens mit Analyse	Types mit Analyse	Anzahl Tokens	Anzahl Types
Blick	alle Wortarten	0,820	0,866	0,786	1564	837
	ohne FM, NE, XY	0,876	0,909	0,831	1445	751
Blogs	alle Wortarten	**0,850**	0,909	0,805	4477	1590
	ohne FM, NE, XY	**0,908**	0,944	0,851	4179	1425
Schobinger	alle Wortarten	0,841	0,879	0,793	1632	691
	ohne FM, NE, XY	0,892	0,923	0,832	1540	647
Swatch	alle Wortarten	*0,717*	0,793	0,644	4503	1998
	ohne FM, NE, XY	*0,817*	0,857	0,704	3863	1662
Wiki	alle Wortarten	0,765	0,844	0,747	2712	1297
	ohne FM, NE, XY	0,819	0,886	0,792	2492	1153

Tabelle 13: Abdeckung nach Textgattung. Zahlen für alle im Testkorpus auftretenden Tokens und Zahlen ohne die Wortarten FM, NE, XY, die nicht behandelt wurden. Berechnung mit demselben Transduktor wie in Tabelle 12.

Spezialfall Substantive

Da bei der fehlenden korrekten Analyse von Substantiven die meisten Prozentpunkte verloren gegangen sind (*Blick* 8%, *Blogs* 3.9%, *Schobinger* 6%, *Swatch* 10% und *Wiki* 9%), soll diese Wortart genauer angeschaut werden. Als offene Wortart sind Substantive durch die Breite des Vokabulars und die Textgattung besonders beeinflusst. Dadurch ergibt sich für die Substantive je nach Textgattung eine unterschiedliche Abdeckung, was in Tabelle 14 ersichtlich ist. Auffällig ist dabei die vergleichsweise hohe Zahl bei den *Blog*-Daten und die tiefe Zahl bei den *Swatch*-Daten.

Teil	Tokens mit korrekter Analyse	Tokens mit Analyse	Types mit Analyse	Anzahl Tokens	Anzahl Types
Blick	0,573	0,679	0,657	274	245
Blogs	**0,701**	0,793	0,751	551	406
Schobinger	0,553	0,678	0,640	208	172
Swatch	*0,540*	0,594	0,528	867	665
Wiki	0,546	0,649	0,634	495	393
alle	0,583	0,668	0,600	2395	1752

Tabelle 14: Abdeckung der Substantive nach Textgattung.

Der Vergleich mit *TreeTagger* zeigt, dass die Substantivlemmata bei den *Swatch*-Texten allgemein ein Problem sind (*TreeTagger* 81,9%). Besser war *TreeTagger* bei den Daten von *Blick* (84,3%), bei den Daten von *Wikipedia* (86,5%) und den Daten von *Schobinger* (88,0%). Die Daten aus den *Blogs* waren für beide Systeme die einfachsten (*TreeTagger*: 88,6%).

	Teil	Substantive (Vorkommen bzw. Länge)
häufigste	Blick	*Dialäkt* ‚Dialekt' (5); *Stuck* ‚Stück' (3); *Lüüt* ‚Leute' (3)
	Blog	*Bus* ‚Bus' (8); *Schuel* ‚Schule' (7); *Hotel* ‚Hotel' (7)
	Schobinger	*Tokter* ‚Doktor' (5); *Wääg* ‚Weg' (4); *Krimi* ‚Krimi' (4)
	Swatch	*Modäu* ‚Modell' (9); *Marke* ‚Marke' (8); *Uhr* ‚Uhr' (7)
	Wiki	*Göttin* ‚Göttin' (7); *Schtaat* ‚Staat' (6); *Name* ‚Name' (6)
längste	Blick	*Chindheitserinnerige* ‚Kindheitserinnerungen' (20); *Bahnhof-olte-dialäkt* ‚Bahnhof-Olten-Dialekt' (20); *Uci-weltranglischte* ‚UCI-Weltrangliste' (19)
	Blog	*Ohgottichwechsledennmaldstrassesiite-art* ‚Oh-Gott-ich-wechsle-dann-mal-die-Straßenseite-Art' (40); *Peanutbuttersandwiches* ‚Peanutbuttersandwiches' (22); *Kokosnussverchäufer* ‚Kokosnussverkäufer' (19)
	Schobinger	*Mèèndig-sälbschtmord* ‚Montagsselbstmord' (20); *Korperazioonsbürger* ‚Korporationsbürger' (19); *Hüüraatsvermittleri* ‚Heiratsvermittlerin' (19)
	Swatch	*Wohltätigkeits-Galavrastaltig* ‚Wohltätigkeitsgalaveranstaltung' (29); *Grand-Feu-Emailzifferblatt* ‚Grand-Feu-Emailzifferblatt' (27); *IP-Sync-Grandmaster-Lösige* ‚IP-Sync-Grandmaster-Lösungen' (26)
	Wiki	*Rächtschryb-tradition* ‚Rechtschreibtradition' (21); *Fasnachts-grubbierige* ‚Fasnachtsgruppierungen' (21); *Überlandstrassenbahn* ‚Überlandstraßenbahn' (20)

Tabelle 15: Typische Substantive nach Korpusteil. In den *Swatch*-Daten sind die längsten Wörter verhältnismässig länger als in den anderen Teilen.

Um einen Einblick in das Vokabular zu geben, zeigt Tabelle 15 die häufigsten und die längsten Substantive in den Testdaten, gruppiert nach Textgattung. Die Substantive in den *Blick*-Daten stammen aus Artikeln zu Dialektologie (im Bezug auf die Dialektausgabe) oder aus Nachrichten. Auch alltägliche Sprache kommt vor. Mit *Chindheitserinnerige* findet sich ein eher alltägliches Wort als eines der längsten Wörter im Text. Die Daten aus den *Blogs* nehmen einen Mittelweg ein, zwischen gewöhnlichen Substantiven als häufigste und einem Kompositum, das einen zusammengezogenen Satz enthält, als längstes. Ähnlich stehen auch die *Schobinger*-Daten zwischen üblichen Wörtern und spontan gebildeten Komposita. Auffällig sind jedoch die Substantive bei *Swatch*, wo sich selbst die häufigsten Substantive der Uhrmacherei zuordnen lassen, was die Ausrichtung des Unternehmens widerspiegelt. Die längsten Substantive sind hier mit einer Ausnahme alle länger als

bei den anderen Daten. Bei den *Wiki*-Daten gibt die Wahl der Substantive einen Hinweis auf die Artikel, aus denen die Sprachdaten stammen.

Die Betrachtung der Substantivlängen gibt Hinweise darauf, wo die Probleme liegen. Wie im Standarddeutschen können im Schweizerdeutschen theoretisch beliebig lange Komposita gebildet werden und je länger ein Wort ist, umso eher ist es ein Kompositum. Dies ist deshalb wichtig, weil sich das System bei den Komposita vollständig auf die Abdeckung der Lemmata von *Morphisto* verlässt und Komposita deshalb nur erkannt werden können, wenn sie in den Daten von *Morphisto* vorliegen. Die obigen Beobachtungen bestätigend weisen die Texte von *Swatch* die längsten Substantive auf (9,2 Zeichen ± 4,3). Am kürzesten (6,4 Zeichen ± 3,2) sind sie in den Texten aus den *Blogs*, wo ihre Abdeckung am besten ist. Dazwischen liegen die Daten von *Blick* (7,2 Zeichen ± 3,4), von *Schobinger* (7,5 Zeichen ± 3,5) und von *Wikipedia* (7,9 Zeichen ± 3,7). Von einer verbesserten Behandlung der Komposita können vor allem die *Swatch*-Texte profitieren.

5.1.3 Analyse nach Dialekten

Um den Nutzen der Aufteilung des Systems für unterschiedliche Dialekte zu überprüfen, sind die dialektspezifischen Systeme auf den dialektannotierten Teilen des Testkorpus getestet worden. Zusätzlich wurde auch die Abdeckung der Kombination dieser Systeme berechnet (Tabelle 16).

		Transduktor			
Dialekt	Teil	Baseldeutsch	Berndeutsch	Zürichdeutsch	Bas.+Ber.+Zü.
Basel	Swatch a2	**0,823**	0,820	0,820	0,823
	Wiki a1	**0,821**	0,810	0,810	0,821
Bern	Swatch a3	0,804	**0,851**	0,804	0,851
	Wiki a2	0,770	**0,791**	0,769	0,791
Zürich	Swatch a16	0,884	0,884	0,884	0,884
	Wiki a4	0,804	0,804	0,804	0,804

Tabelle 16: Korrekte Analyse nach Dialekt für die verschiedenen Systeme. Anteil der Tokens aller Wortarten ausser FM, NE und XY, für welche die korrekte Analyse möglich ist.

Während für die bern- und baseldeutschen Texte mit den entsprechenden Systemen eine bessere Analyse erreicht werden konnte, machte dies bei den zürichdeutschen Texten keinen Unterschied. Obwohl die Kombination der Systeme logischerweise die beste Abdeckung ermöglicht, fällt doch auf, dass diese nicht höher ist als die

Abdeckung des zugehörigen dialektspezifischen Systems.

Um die Ursachen der unterschiedlichen Abdeckung zu ergründen, sollen hier die Wörter, welche von den Dialektunterschieden betroffen sind, betrachtet werden. Tabelle 17 zeigt jene Wörter, die von den Analysesystemen für die anderen Dialekte nicht erkannt wurden. Dass ein Wort im System für den eigenen Dialekt nicht, in einem anderen System jedoch korrekt erkannt wurde, ist nicht eingetreten. Das bedeutet, dass die Aufteilung des Morphologieanalysesystems in Dialekte in diesem Test keine negativen Auswirkungen auf die Abdeckung hat.

Dialekt	Teil	Mit anderen Systemen unerkannte Tokens
Basel	Swatch a2	*Briedere* ‚Brüder‘
	Wiki a1	*gheert* ‚gehört‘; *grindet* ‚gegründet‘; *Hytte* ‚heute‘; *Mieh* ‚Mühe‘; *ver* ‚für‘
Bern	Swatch a3	*au* ‚alle‘; *auem* ‚allem‘; *aus* ‚als‘; *bsungers* ‚besonders‘; *drunger* ‚darunter‘; *gstangä* ‚gestanden‘; *härgsteut* ‚hergestellt‘; *häufä* ‚helfen‘; *jewius* ‚jeweils‘; *Mau* ‚Mal‘; *Mittupunkt* ‚Mittelpunkt‘; *Modäu* ‚Modell‘; *Schlüssu* ‚Schlüssel‘; *Sümbou* ‚Symbol‘; *Ungerem* ‚unter dem‘; *uurautä* ‚uralten‘; *viusiitegi* ‚vielseitige‘
	Wiki a2	*aus* ‚als‘; *auso* ‚also‘; *Hiuf* ‚Hilfe‘; *Mau* ‚Mal‘; *Schuud* ‚Schuld‘; *Usbiudig* ‚Ausbildung‘; *viu* ‚viel‘; *Vokau* ‚Vokal‘; *zaut* ‚gezahlt‘; *Zuefau* ‚Zufall‘ *ou* ‚auch‘ (von Baseldeutsch erkannt)
Zürich	Swatch a16	-
	Wiki a4	-

Tabelle 17: Unterschiede in der Abdeckung nach Dialekt. Wörter, die nur durch die Systeme für die anderen Dialekte nicht erkannt wurden.

Die baseldeutschen Wörter, welche durch die anderen Systeme nicht erkannt wurden, weisen alle eine Entrundung der vorderen gerundeten Vokale auf. Dieses Phänomen des Baseldeutschen existiert in den Dialekten von Bern und Zürich nicht.

Die berndeutschen Wörter, deren Analyse durch die anderen Systeme unmöglich war, enthalten entweder eine Vokalisierung von /l/ zu /ʊ/ oder eine Ersetzung von /nd/ durch /ŋ/. Diese beiden Phänomene sind charakteristisch für das Berndeutsche. Ein Spezialfall ist *ou* ‚auch‘, das zwar fürs Baseldeutsche, aber nicht fürs Zürichdeutsche erkannt wurde. Diese Realisierung des Diphthongs ist für beide Dialekte unüblich. Während dies beim Baseldeutschen durch eine höhere Gewichtung gelöst wurde, wurde diese Realisierung beim Zürichdeutschen ganz unterdrückt.

Dass die zürichdeutschen Texte von allen drei Systemen gleich gut abgedeckt werden konnten, zeigt neben der Tabelle oben auch der Umstand, dass keines der Wörter dieser Texte durch die anderen Systeme schlechter verarbeitet werden konnte.

Mögliche Unterschiede begrenzen sich hier lediglich auf die Gewichtung.

5.2 Gewichte

Dieser Abschnitt dreht sich um die Frage, ob die Gewichtung hilft, in möglichst vielen Fällen die richtige Analyse zu bevorzugen. Dazu muss die Rangfolge, die durch die Gewichtung gegeben ist, betrachtet werden.

Für die Beantwortung dieser Frage bieten sich Metriken aus dem *Information Retrieval* an, welche die Rangordnung von Antworten berücksichtigen können. Für diese Arbeit am besten geeignet ist der *Mean Reciprocal Rank* (MRR). „Der **MMR** weist dort jeder Frage einen Wert zu, der gleich dem Kehrwert des Ranges der ersten korrekten Antwort der N besten Kandidaten ist [...]“ (Neumann 2010, S. 587). Ein verbreiteter Wert für N ist 10, was auch hier zur Evaluation verwendet wird.

Der *Reciprocal Rank* (RR) wird für ein Wort mit folgender Formel (vereinfacht) berechnet (Büttcher et al. 2010, S. 409):

$$RR = \frac{1}{\min(k|Res[k] \in Rel)}$$

In der Formel steht k für den Rang eines Resultats, $Res[k]$ entsprechend für das Resultat mit diesem Rang. Beachtet werden nur relevante – in diesem konkreten Fall korrekte – Resultate aus der Menge Rel. Durch die Funktion $\min()$ werden alle Ränge ausser dem kleinsten ignoriert, wodurch folgt, dass nur das erste relevante Resultat gesucht wird. Ergänzend zur Formel gilt, dass wenn die richtige Analyse unter den N ersten Vorschlägen nicht vorkommt, der Wert Null als RR vergeben wird. Anschliessend wird der Durchschnitt aller RR berechnet, um den MRR zu erhalten.

Anhand eines Beispiels soll hier der MRR verdeutlicht werden. Tabelle 18 zeigt einen Satz und die jeweiligen Analysevorschläge für jedes Token. Der RR entspricht jeweils dem Kehrwert des Ranges.

Das Token *New-Gent-Modäu* ‚New-Gent-Modell‘, für das keine Analyse gemacht werden kann, zeigt, wie die Abdeckung in den MRR einfliesst. Denn selbst wenn die Gewichtung immer die richtige Analyse bevorzugen würde, kann die Obergrenze von 90% nicht übertroffen werden. Um den Einfluss der Abdeckung aus dem MRR zu eliminieren, wird in diesem Kapitel deshalb jeweils zusätzlich der MRR ausschliesslich auf den Tokens mit korrekter Analyse angegeben.

Token	Analysen	Rang	RR
D’	**die/ART.drsf** , die/ART.drp*	1	1/1
Kollektion	**Kollektion/NN.sf**	1	1/1
besteit	**bestehen/VVFIN.3si** , bestehen/VVFIN.3pi , ...	1	1/1
us	aus/APZR , aus/PTKVZ , **aus/APPR.d** , ...	3	1/3
zäh	**zehn/CARD** , zäh/ADJD.p , Zehe/NN.sf , ...	1	1/1
New-Gent-Modäu		-	0
i	**in/APPR.d** , ich/PPER.1s*n , in/APPR.a	1	1/1
lüchtendä	leuchtend/ADJA.pdsfw , **leuchtend/ADJA.pdp*s** , ...	2	1/2
Farbä	Farbe/NN.sf , **Farbe/NN.pf**	2	1/2
.	**./$.**	1	1/1
		MRR:	0,733

Tabelle 18: Beispiel für MRR. Die korrekte Analyse ist jeweils fett gedruckt.

Als Untergrenze wurde jeweils der MRR berechnet und zwar auf einer zufällig erstellten Rangfolge der Analysen. Auch hier wurden zusätzlich Zahlen ohne Einfluss der Abdeckung berechnet.

Mit dem MRR verwandt ist die *Mean Average Precision* (MAP), die mit mehreren korrekten Antworten in einer Rangordnung umgehen kann. Da im Testkorpus aber immer nur eine korrekte Analyse steht, unterscheiden sich die beiden Metriken für diese Anwendung nicht (vgl. Büttcher et al. 2010, S. 408–409).

5.2.1 Analyse auf dem kompletten Testkorpus

Die Evaluation der Gewichte über das ganze Korpus hinweg zeigt mit einem MRR von bis 73% bzw. 85% ohne Einfluss der Abdeckung positive Resultate. Tabelle 19 stellt den MRR für die einzelnen Dialekttransduktoren dar.

Wie im letzten Kapitel wurden auch hier einmal für alle Tokens und einmal für die Tokens der behandelten Wortarten (d. h. ohne FM, NE und XY) die Zahlen berechnet. Trotz der schlechtesten Abdeckung konnte das System für Zürichdeutsch den besten MRR erreichen. Entsprechend steigt der MRR für Zürichdeutsch noch beim Ausschluss der unerkannten Wörter weiter. Grund dafür könnte unter Umständen sein, dass viele der Testdaten in diesem Dialekt verfasst sind. Ein tieferer MRR bei einer höheren Abdeckung weist aber auch auf eine Übergenerierung bei den anderen Dialekten hin.

Transduktor		Tokens mit korr. Analyse	MRR f. alle Tokens System	MRR f. alle Tokens zufällig	MRR f. erk. Tokens System	MRR f. erk. Tokens zufällig
Basel	alle Wortarten	0,786	*0,658*	0,421	*0,837*	0,535
	ohne FM, NE, XY	0,856	*0,715*	0,454	*0,835*	0,530
Bern	alle Wortarten	0,789	0,663	0,432	0,840	0,547
	ohne FM, NE, XY	0,859	0,720	0,467	0,838	0,543
Zürich	alle Wortarten	0,785	**0,668**	0,437	**0,850**	0,556
	ohne FM, NE, XY	0,855	**0,726**	0,475	**0,849**	0,555
Basel+Bern	alle Wortarten	0,790	0,666	0,420	0,843	0,531
+Zürich	ohne FM, NE, XY	0,860	0,724	0,456	0,842	0,530

Tabelle 19: MRR auf dem Testkorpus nach dialektspezifischem Analysewerkzeug. Resultate für alle im Korpus auftretenden Tokens und Resultate ohne die Wortarten FM, NE, XY, die nicht behandelt wurden. Die Berechnung nur auf den erkannten Wörtern beseitigt den Einfluss der Abdeckung.

Analyse nach Wortarten

Wie bei der Abdeckung soll hier für die Gewichtung eine Untersuchung nach Wortarten zeigen, wo noch Verbesserungsmöglichkeiten vorhanden sind. Tabelle 20 zeigt den MRR durch die Rangfolgen des Systems pro Wortart sowie den MRR durch eine zufällige Reihenfolge und die Abdeckung der Analyse zum Vergleich.

Wortart	Tokens mit korrekter Analyse	MRR f. alle Tokens System	MRR f. alle Tokens zufällig	MRR f. erk. Tokens System	MRR f. erk. Tokens zufällig	Anzahl Tokens
ADJA	0,662	0,290	0,215	*0,439*	0,324	715
ADJD	0,723	0,610	0,225	0,844	0,312	328
ADV	0,901	0,810	0,438	0,899	0,486	970
APPO	0,500	0,375	0,286	0,750	0,571	4
APPR	0,959	0,728	0,360	0,759	0,375	909
APPRART	0,970	0,768	0,410	0,792	0,423	535
APZR	1,000	1,000	0,333	**1,000**	0,333	1
ART	0,971	0,700	0,375	0,721	0,386	1086
CARD	0,807	0,801	0,715	**0,993**	0,886	259
FM	0,000	0,000	0,000	-	-	183
ITJ	0,194	*0,077*	0,129	*0,400*	0,667	31

weiter auf der nächsten Seite

Wortart	Tokens mit korrekter Analyse	MRR f. alle Tokens System	MRR f. alle Tokens zufällig	MRR f. erk. Tokens System	MRR f. erk. Tokens zufällig	Anzahl Tokens
KOKOM	0,986	0,650	0,309	0,660	0,313	69
KON	0,992	0,901	0,470	0,908	0,474	513
KOUI	0,938	0,625	0,448	0,667	0,478	16
KOUS	0,970	0,750	0,276	0,773	0,285	131
NE	0,129	0,124	0,101	**0,957**	0,783	1021
NN	0,583	0,531	0,347	0,911	0,595	2395
PAV	0,870	0,841	0,528	0,967	0,607	46
PDAT	0,907	0,508	0,165	0,560	0,182	75
PDS	0,963	*0,296*	0,305	*0,308*	0,317	81
PIAT	0,873	0,558	0,258	0,639	0,296	79
PIDAT	1,000	*0,470*	0,571	*0,470*	0,571	11
PIS	0,944	0,645	0,309	0,684	0,327	142
PPER	0,980	0,756	0,206	0,772	0,211	395
PPOSAT	0,936	0,758	0,179	0,810	0,191	155
PRELS	0,957	0,932	0,533	**0,974**	0,557	94
PRF	0,948	0,570	0,141	0,601	0,149	77
PTKA	0,923	*0,385*	0,423	*0,417*	0,459	13
PTKAM	1,000	0,333	0,250	*0,333*	0,250	1
PTKANT	0,889	0,889	0,589	1,000	0,663	18
PTKINF	0,917	0,844	0,329	0,920	0,358	12
PTKNEG	1,000	1,000	0,321	1,000	0,321	78
PTKVZ	0,881	0,607	0,395	0,689	0,449	67
PTKZU	0,940	*0,327*	0,552	*0,348*	0,587	50
PWAV	0,889	0,539	0,492	0,607	0,554	27
PWS	1,000	0,955	0,261	**0,955**	0,261	29
TRUNC	0,000	0,000	0,000	-	-	12
VAFIN	0,980	0,913	0,441	0,931	0,450	816
VAINF	0,864	0,136	0,126	*0,158*	0,146	22
VAPP	0,943	0,943	0,548	1,000	0,581	158
VMFIN	0,943	0,763	0,329	0,810	0,349	87
VMINF	0,967	0,967	0,160	1,000	0,166	30
VVFIN	0,881	0,624	0,307	0,708	0,349	404
VVIMP	0,667	*0,152*	0,296	*0,228*	0,444	9
VVINF	0,848	0,702	0,195	0,828	0,230	197
VVIZU	0,500	0,500	0,171	**1,000**	0,342	8
VVPP	0,851	0,837	0,549	**0,983**	0,645	484
XY	0,000	0,000	0,000	-	-	165

weiter auf der nächsten Seite

Wortart	Tokens mit korrekter Analyse	MRR f. alle Tokens System	MRR f. alle Tokens zufällig	MRR f. erk. Tokens System	MRR f. erk. Tokens zufällig	Anzahl Tokens
$(	0,986	0,986	0,986	**1,000**	1,000	347
$,	1,000	1,000	1,000	**1,000**	1,000	568
$.	0,999	0,999	0,999	**1,000**	1,000	965

Tabelle 20: MRR nach Wortarten im Testkorpus. Berechnung mit dem selben System wie in Tabelle 12.

Unterschreitung der Untergrenze

Bei den Wortarten ITJ, PDS, PIDAT, PTKA, PTKZU und VVIMP ist die Rangfolge durch die Gewichtung schlechter, als eine zufällige Ordnung erwarten lässt. Die Untersuchung der Gewichte bei den Interjektionen scheint bei der geringen Abdeckung wenig sinnvoll. Auf die anderen Wortarten soll hier aber eingegangen werden.

Bei den substituierenden Demonstrativpronomina (PDS) konnte in einem separaten Test zur Lemmatisierung ein MRR von 47,5% beobachtet werden. Ein Test nach Lemmatisierung und Wortart (ohne morphologische Merkmale) war dagegen nicht besser als der Test auf der vollen Analyse. Das bessere Abschneiden bei der Auslassung der Wortart kann auf die damit nicht mehr erfolgte Unterscheidung zwischen PDS und PDAT zurückgeführt werden. Dass der MRR trotzdem noch weit hinter der guten Abdeckung zurückbleibt, liegt wohl am Umstand, dass die formenähnlichen Artikel andere Lemmata tragen. So kann *dä* sowohl für substituierende oder attribuierende Demonstrativpronomina stehen, beide /dæː/ bzw. /dɛː/ ‚dieser' (Lemma: ‚diese'), aber ebenso für /də/ ‚der' (Lemma: ‚die').

Bei den attribuierenden Indefinitpronomina mit Determiner (PIDAT) gibt die Betrachtung der Merkmale Lemma und Wortart Aufschluss. Betrachtet man einzig die Lemmatisierung, dann wird ein MRR = 1 erreicht, was heisst, dass immer der erste Vorschlag das korrekte Lemma beinhaltet. Keine Verbesserung im Vergleich zur vollen Analyse gibt der Vergleich der Wortart. Zusammen bedeutet das, dass die Lemmata von PIDAT durch gleiche Lemmata einer anderen Wortart konkurriert werden. Es handelt sich dabei um substituierende oder attribuierende Formen der Indefinitpronomina. Dies bedeutet zudem, dass die syntaktisch begründete Unterscheidung der Indefinitpronomina für eine tokenbasierte Analyse nicht geeignet ist. So folgt beispielsweise ‚alle' dem gleichen Flexionsschema, egal wie es verwendet wird. Mit einem Tagger ist dieses Problem aber einfach zu beheben, da die Mehrdeutigkeit durch ein Sprachmodell einfach aufgelöst werden kann.

Auch bei den Partikeln vor Adverbien und Adjektiven und der Infinitivpartikel *zu* bringt die blosse Betrachtung des Lemmas eine Verbesserung, während die Betrachtung der Wortart selbst keinen Anstieg des MRR mit sich bringt. Bei PTKZU ist eine Verwechslungsmöglichkeit mit der Präposition *zu* zu erwarten und die alleinige Betrachtung des Lemmas führt zu einem MRR = 0,96. Weniger hoch ist der Nutzen bei PTKA, weil die Partikel *am* bei Superlativen nicht gegen die Partikel *am* im Progressiv, sondern auch gegen Formen der Präposition *an* mit Artikel, die als *an* lemmatisiert wird, zu kämpfen hat. Für die PTKA *zu* kann Ähnliches wie PTKZU gelten.

Nicht viel anderes kann bei den Imperativen gesagt werden. Mit dem Vergleich nur auf dem Lemma basierend ist der MRR bei diesen Wörtern ebenfalls besser. Konkurrierende Formen mit gleichem Lemma finden sich hier vor allem unter den finiten Verbformen.

Tiefer MRR

Bei den attributiven Adjektiven und den Infinitiven der Hilfsverben hinkt der MRR ebenfalls deutlich der Abdeckung hinterher, obwohl diese Wörter bereits von der Gewichtung profitieren.[2] Da die Abdeckung hauptsächlich bei den Adjektiven miteinfliesst, sollen hier diejenigen Wortarten betrachtet werden, deren MRR für die erkannten Tokens unter 0,5 liegt.

Die Infinitive der Hilfsverben (VAFIN) schneiden bei einer Betrachtung auf die Lemmatisierung beschränkt mit 65,5% (ohne Bereinigung) deutlich besser ab, während ein Test auf der Wortart alleine schlecht abschneidet. Beim Vergleich der Lemmata ist die Unterscheidung zwischen dem Verb *sein* und dem Possessivpronomen *sein* aufgehoben. Dies kann den MRR bereits verbessern. Der immer noch vorhandene Unterschied zur Abdeckung liegt wohl unter anderem darin, dass *si* ‚sein' sich lautlich auch mit den Personalpronomina *si* ‚sie' überschneidet. Die Gewichtung der Analysen für diese Tokens reicht also nicht für eine Disambiguierung aus. Auch in diesem Fall kann ein Sprachmodell zur Lösung dieses Problem verwendet werden.

Das schlechte Resultat bei den attributiven Adjektiven hingegen lässt sich kaum durch die Konkurrenz mit einer anderen Wortart begründen. Hier liegt die Ursache eher darin, dass das Paradigma der Adjektive viele zusammenfallende Formen aufweist (siehe Tabelle 21). Während der MRR mit der Berechnung über den

[2] Bei PTKAM liegt der MMR auch hinter der Abdeckung, doch mit nur einem Token hat ein Vergleich mit zufälligen Rangordnungen keine Aussagekraft.

Kehrwert schon den zweiten Vorschlag stark bestraft, liegen die Gewichte, also die Grundlage der Rangordnung, relativ nahe beieinander. Abbildung 9 zeigt die Ausgabe des Analysesystems mit den Gewichten.

		Singular			Plural
Flexion	Kasus	m	f	n	*
stark	Nom./Akk.	schööne	schööni	schööns	schööni
	Dativ	schöönem	schööne(r)	schöönem	schööne
schwach	Nom./Akk.	schöön(i)	schöön(i)	schöön(i/e)	schööne
	Dativ	schööne	schööne	schööne	schööne

Tabelle 21: Formen des Adjektivs ‚schön' im Positiv. Viele Endungen erscheinen mehrmals in der Tabelle. Die Steigerungsformen *schöner* und *schönst* folgen dem gleichen Muster, sind aber meistens von den Formen des Positivs abgrenzbar.

```
hfst[1]: up schööne
schön/ADJA.pdsfw        13.59961
schön/ADJA.pdp*s        13.89941
schön/ADJA.pdp*w        14.00000
schön/ADJA.pdsmw        14.00000
schön/ADJA.pdsnw        14.00000
schön/ADJA.prsms        14.00000
schön/ADJA.prp*w        14.09961
schön/ADJA.prsnw        14.29980
schön/ADJA.pdsfs        16.39941
```

```
hfst[1]: up schööni
schön/ADJA.prp*s        13.59961
schön/ADJA.prsfw        13.89941
schön/ADJA.prsfs        14.09961
schön/ADJA.prsnw        14.29980
schön/ADJA.prsmw        15.00000
```

Abbildung 9: Ausgabe bei der Analyse von Adjektiven mit dem System für Berndeutsch.

Die oben geäusserten Vermutungen werden durch den Vergleich des MRR auf Lemma-, Wortarten- und Lemma-Wortarten-Ebene gestützt. Bei der alleinigen Betrachtung des Lemmas übertrifft der MRR die Abdeckung korrekter Analysen und auch der MRR auf den Wortarten steht nur knapp unter der Abdeckung. Wirklich aussagekräftig ist aber der MRR unter der blossen Auslassung der morphologischen Kategorien. Mit dieser kommt der MRR auf 59,3% und damit in die Nähe der Abdeckung (66,2%). Dies bestätigt die Vermutung, dass sich hier vor allem die lautgleichen Formen mit unterschiedlicher Analyse konkurrieren und ohne Wortumfeld nicht unterschieden werden können. Für solche Fälle muss der Einsatz von Sequenztagging geprüft werden, um diese syntaktisch bedingten Mehrdeutigkeiten aufzulösen.

5.2.2 Analyse nach Dialekten

Weniger klar ist der Nutzen der Gewichtung bei den nach Dialekt getrennten Systemen. Tabelle 22 vergleicht den MRR bei den dialektspezifischen Texten pro System. Während bei den zürichdeutschen Texten der MRR am höchsten ausfällt, wenn für sie das System für Zürichdeutsch angewendet wird, ist bei den baseldeutschen oder berndeutschen Texten das dialektspezifische System nur geringfügig besser als die Kombination der drei Systeme. Eher überraschend ist dagegen, dass das System für Zürichdeutsch auf den baseldeutschen Texten zwar die schlechtere Abdeckung erreicht (siehe Kapitel 5.1.3), die Gewichte relativ zur Abdeckung aber höher sind.

		Transduktor			
Dialekt	Teil	Baseldeutsch	Berndeutsch	Zürichdeutsch	Bas.+Ber.+Zü.
Basel	Swatch a2	0,666 (0,810)	0,663 (0,808)	**0,668** (**0,815**)	0,666 (0,810)
	Wiki a1	**0,678** (0,826)	0,676 (0,834)	0,677 (**0,836**)	0,676 (0,824)
Bern	Swatch a3	0,692 (0,861)	**0,733** (**0,862**)	0,692 (0,861)	0,733 (0,861)
	Wiki a2	0,638 (0,828)	**0,658** (0,832)	0,640 (**0,833**)	0,657 (0,831)
Zürich	Swatch a16	0,745 (0,843)	0,749 (0,847)	**0,753** (**0,852**)	0,747 (0,845)
	Wiki a4	0,707 (0,879)	0,712 (0,885)	**0,721** (**0,897**)	0,714 (0,888)

Tabelle 22: MRR in den dialektannotierten Korpusteilen. In Klammern der MRR für die Tokens mit korrekter Analyse. FM, NE und XY sind bei der Berechnung nicht berücksichtigt.

Auf einen Nutzen der Aufteilung in dialektspezifische Systeme weist also höchstens der Test auf den zürichdeutschen Texten hin. Dies ist so zu werten, dass beim System für Zürichdeutsch weniger Formen aus anderen Dialekten oder Übergenerierungen vorkommen, welche die gewünschte Analyse konkurrieren. Die tiefere Abdeckung auf dialektfremden Texten unterstützt diese Vermutung. Möglicherweise ist auch die Bestrafung unerwünschter Analysen primär im baseldeutschen System noch mangelhaft.

Die gewählten Dialekte um Basel, Bern und Zürich sind möglicherweise zu nahe verwandt, um den Nutzen der Aufteilung zu zeigen. Dafür, dass die Formen des einen Dialekts diejenigen eines anderen dieser Dialekte konkurrieren, wurden keine Hinweise gefunden. Da aber das hier entwickelte Morphologieanalysesystem im Hinblick auf andere Dialekte offen gelassen wurde, kann keine generelle Aussage zum Nutzen dialektspezifischer Module gezogen werden. Diese Frage kann also erst vollständig beantwortet werden, wenn genug Entwicklungs- und Testdaten in den anderen Dialekten des Schweizerdeutschen vorliegen.

6 Fazit

Mit einer korrekten Analyse von über 90% der Tokens der behandelten Wortarten in ausgewählten Texten kann die Entwicklung des Morphologieanalysesystems für Schweizerdeutsch *Taggswort* als Erfolg bewertet werden. Auf Texten mit einem breiteren Vokabular macht der Anteil der Tokens der behandelten Wortarten immer noch 86% aus, wobei hier neben den berücksichtigten Dialekten noch weitere vertreten sind. Mit einem *Mean Reciprocal Rank* von 72% (im Vergleich zu 86% als Obergrenze dazu) trägt die Gewichtung zweifellos positiv bei, um die korrekten Analysen zu präferieren.

Neben dem Morphologieanalysesystem ist im Rahmen dieser Arbeit auch ein Korpus von rund 29 000 Tokens mit feiner morphologischer Annotation entstanden. Dieses stellt eine wertvolle Basis für weitere Forschungen zur schweizerdeutschen Morphologie und Syntax dar und kann mit Hilfe des entwickelten Systems ohne grossen Aufwand erweitert werden.

Im Rahmen dieser Arbeit wurde die bereits existierende Wortartenauszeichnung für Schweizerdeutsch angepasst, sodass die syntaktischen Phänomene besser beschrieben werden können. Besonders sticht dabei das vorgeschlagene Tag *PTKAM* heraus, das eine Analyse des *am*-Progressivs als Konjugationsform sauber und konsistent ermöglicht.

Ein Nachteil der Kompatibilität mit Tagsets für das Standarddeutsche ist die Feinheit der Aufteilung in die Wortarten. Die Unterscheidung zwischen attribuierendem und substituierendem Gebrauch bei den Pronomina führt jeweils beim seltener verwendeten Gebrauch zu einem schlechteren MRR. Für reine Morphologieanalyse ist die syntaktische Unterscheidung in *attribuierend* und *substituierend* nicht besonders sinnvoll und die Ambiguität kann nur im Wortumfeld aufgelöst werden.

Mit dem aufs Schweizerdeutsche zugeschnittenen Auszeichnungsschema für die morphologischen Kategorien *STTS.gsw* ist es möglich, die Formen der Wörter eindeutig zu beschreiben. Ziel dieser Aufgabe war es, die Zahl der Analysen zu reduzieren und trotzdem die Kompatibilität zu Schemata für das Standarddeutsche zu bewahren. Eine der Anpassungen ist der Verzicht auf die gemischte Flexion der

Adjektive, was durch den unterschiedlichen Gebrauch der Flexionstypen bedingt ist. Mit dem Fehlen des Präteritums entfällt zudem die Kategorie Tempus und die beiden Konjunktive sind als unabhängige Modi interpretiert worden. Bei den Kasus fällt erstens das Fehlen des Genitivs auf, zweitens wurde ein Sammlungstag für Nominativ und Akkusativ geschaffen für diejenigen Fälle, in denen sich diese nicht klar unterscheiden lassen.

Auch bei den Adjektiven stellt die feine Einteilung trotz Auslassen der gemischten Flexion ein Problem bei der Disambiguierung der Formen dar. Begründung für diese Feinheit ist nicht nur die Kompatibilität, sondern auch die lange linguistische Tradition und vor allem die Verteilung der Formen über das Paradigma. Auch dieses Problem muss durch eine Nachbearbeitung mit Hilfe eines syntaktischen Sprachmodells gelöst werden.

Die Klitika wurden mit einer einfachen Verkettung mit den Wörtern, an die sie angefügt werden können, integriert. Mit *flag diacritics* können dazu die Wörter, die sie tragen können, ausgewählt oder sogar ungrammatische Kombinationen innerhalb der Tokens verhindert werden. Seitens der linguistischen Beschreibung sind die Klitika wie unabhängige Wörter interpretierbar, die mit einem Pluszeichen verbunden sind. Auf der Seite der Oberflächenform hingegen sind Wörter mit Klitika als ganze Tokens aufzufassen und die Morphemgrenzen sind nicht markiert.

Für die Ambiguitäten, welche durch den fehlenden orthographischen Standard entstehen, sind die Gewichte eine funktionierende Lösung. Die Fälle, in denen die Gewichte eine suboptimale Analyse ausgeben, lassen sich durch formengleiche Lemmata erklären. Beispielsweise konkurrieren sich Analysen bei den Partikeln und Pronomina.

Die Übernahme von Stämmen aus Ressourcen für das Standarddeutsche garantiert bereits eine gewisse Abdeckung. Da die übernommenen Wortarten offen sind, sind sie durch die verschiedenen Textgattungen besonders beeinflusst. Hierfür können aber Ressourcen aus weiteren Systemen integriert werden.

Durch die manuelle Erstellung eines Vollformenkerns mit schweizerdeutschen Stämmen konnte eine sehr gute Abdeckung erreicht werden ohne komplexe Ersetzungsregeln anwenden zu müssen. Besonders geeignet ist dieses Vorgehen bei den geschlossenen Wortklassen, die damit fast vollständig abgedeckt werden konnten. Auch für häufige und sehr unregelmässige oder für Schweizerdeutsch spezifische Wörter der offenen Wortklassen eignet sich diese Art der lexikographischen Erfassung.

Weniger Nutzen als erwartet ist bislang in der separaten Behandlung für die dia-

lektspezifischen Lautformen zu erkennen. Das System für das Zürichdeutsche konnte auf den entsprechenden Texten die beste Gewichtung ohne Verlust bei der Abdeckung erreichen. Beim System für das Baseldeutsche konnte dies aber nicht erreicht werden. Zukünftige Experimente sollten sich also um die Verbesserung des Systems für Baseldeutsch konzentrieren. Erweiterungen für Dialekte, vor allem für die alpinen, deren Vokalsystem sich von den Mittellandsdialekten unterscheiden, könnten aber einen deutlicheren Nutzen zeigen.

Weitere Erweiterungen betreffen das Vokabular, entweder für Wörter, für welche die Ersetzungsregeln nicht greifen, oder für die aus *Morphisto* übernommenen Stämme. Bei letzteren verdient die Kompositabildung besondere Beachtung, wie auch die von den bereits bekannten Stämmen ausgehende Derivation. Für beide Punkte lässt sich der Aufbau von standarddeutschen Systemen adaptieren. Die bisherige Abdeckung, die auf einfache Weise erreicht werden konnte, sollte mit entsprechend geringem Aufwand noch ein beträchtliches Stück angehoben werden können.

Neben der Erweiterung des Systems selbst ist auch die Weiterverarbeitung der vorgeschlagenen Analysen für die Wörter ein Thema für künftige Arbeiten. Man könnte einen morphologischen Tagger bauen, der die wortbasierten morphologischen Analysen im Kontext von Sätzen entsprechend ihrer Abfolgewahrscheinlichkeit klassifiziert. Eine entsprechende Nachbearbeitung unseres Morphologieanalysesystems für Rumantsch Grischun[1] mit einem CRF-Tagger konnte bereits mit wenig Trainingsdaten (4500 Tokens) ermöglicht werden. Ob und wie die Gewichtung der Analysen darin integriert werden kann, muss mit Experimenten geprüft werden.

[1] Interaktiv verfügbar auf `http://kitt.ifi.uzh.ch/kitt/rumansh/dev/` (aufgerufen am 4. April 2016)

Bibliographie

Baumgartner, Heinrich, Konrad Lobeck, Robert Schläpfer, Rudolf Hotzenköcherle, Doris Handschuh, Rudolf Trüb, Paul Zinsli und Walter Haas, Hrsg. (1962-2003). *Sprachatlas der deutschen Schweiz*. Ab Bd. 5 Herausgabe fortgef. von Robert Schläpfer, Rudolf Trüb, Paul Zinsli. Bern und Basel: Francke.

Baumgartner, Reto, Martina Bachmann, Rolf Badat, Daniel Hegglin, Susanna Tron und Melanie Widmer (2013). *Morphologieanalyse für Rumantsch Grischun*. Universität Zürich, Institut für Computerlinguistik. Zürich. URL: `http://kitt.cl.uzh.ch/kitt/rumansh/documentation.pdf`.

Beesley, Kenneth R. und Lauri Karttunen (2003). *Finite State Morphology*. Standford (Kalifornien): CSLI Publications.

Bolzern, T. (2015). *Die Technik versteht jetzt Schweizerdeutsch*. Hrsg. von 20 Minuten. (online, aufgeschaltet am 26. Nov. 2015, aufgerufen am 16. Feb. 2016). URL: `http://www.20min.ch/digital/news/story/11708663`.

Büttcher, Stefan, Charles L. A. Clarke und Gordon V. Cormack (2010). *Information Retrieval - Implementing and Evaluating Search Engines*. Cambridge (Massachusetts): MIT Press.

Christen, Helen, Elvira Glaser und Matthias Friedli, Hrsg. (2012). *Kleiner Sprachatlas der deutschen Schweiz*. 4. Aufl. (1. Aufl. 2010). Frauenfeld: Huber.

Crysmann, Berthold, Silvia Hansen-Schirra, George Smith und Dorothea Ziegler-Eiseles (2005). *TIGER Morphologie-Annotationsschema*. Universität des Saarlandes, Universität Stuttgart, Universität Potsdam. URL: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-morph.pdf`.

Didakowski, Jörg (2005). *Robustes Parsing und Disambiguierung mit gewichteten Transduktoren*. Bd. 23. Linguistics in Potsdam. Potsdam: Universitätsverlag Potsdam. URL: `http://www.dwds.de/static/website/publications/text/LIP23.pdf`.

Dieth, Eugen (1986). *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*. Hrsg. von Christian Schmid. 2. Aufl. / bearb. und hrsg. von Christian Schmid-Cadalbert (1. Aufl. 1938). Lebendige Mundart. Aarau etc.: Sauerländer.

Duden online, Hrsg. (2016). *am.* (online, aufgerufen am 15. Februar 2016). Dudenverlag, Bibliographisches Institut GmbH. Berlin. URL: http://www.duden.de/rechtschreibung/am.

Dürscheid, Christa und Karina Frick (2014). „Keyboard-to-Screen-Kommunikation gestern und heute: SMS und WhatsApp im Vergleich". In: *Networx* 64, S. 149–181. URL: http://www.mediensprache.net/networx/networx-64.pdf.

Gesmundo, Andrea und Tanja Samardžic (2012). „Lemmatisation as a Tagging Task". In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*. The Association for Computer Linguistics, S. 368–372. URL: http://www.aclweb.org/anthology/P12-2072.

Hollenstein, Nora und Noëmi Aepli (2014). „Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging". In: *COLING 2014, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Hrsg. von Marcos Zampieri, Liling Tan, Nikola Ljubešić und Jörg Tiedemann, S. 85–94. URL: http://www.aclweb.org/anthology/W/W14/W14-53.pdf.

Hulden, Mans, Iñaki Alegria, Izaskun Etxeberria und Montse Maritxalar (2011). „Learning word-level dialectal variations as phonological replacement rules using a limited parallel corpus". In: *Proceedings of EMNLP 2011. Conference on Empirical Methods in Natural Language Processing*. Hrsg. von Association for Computational Linguistics, S. 39–48. URL: http://www.aclweb.org/anthology/W11-2605.

IDS Institut für Deutsche Sprache, Programmbereich Korpuslinguistik (2012). *Korpusbasierte Wortgrundformenliste DeReWo, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation*. Mannheim. URL: http://www.ids-mannheim.de/derewo.

Klaper, David (2014). *11-712: NLP Lab Report: A Dependency Parser for Swiss German*. URL: https://github.com/DKlaper/gsw-DepParser/blob/master/Report/reportDKlaper.pdf.

Linde, Sonja (2011). *Referenzkorpus Althochdeutsch. Kurzbeschreibung*. www.sprachgeschichte.de/DDD. URL: http://www2.hu-berlin.de/sprachgeschichte/mitarbeiter/richling/Manual.pdf.

Lindén, Krister, Miikka Silfverberg und Tommi A. Pirinen (2009). „HFST Tools for Morphology - An Efficient Open-Source Package for Construction of Morphological Analyzers". In: *State of the Art in Computational Morphology. Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009, Zurich, Switzerland, September 2009. Proceedings*. Hrsg. von Cerstin Mahlow und Michael Piotrowski. Bd. 41. Communications in Computer and Information Sci-

ence. Springer, S. 28–47.

Marti, Werner (1985a). *Bärndütschi Schrybwys: ein Wegweiser zum Aufschreiben in berndeutscher Sprache: mit einer Einführung über allgemeine Probleme des Aufschreibens und einem Wörterverzeichnis nebst Beispielen*. 2., überarb. Aufl. (1. Aufl. 1972). Bern: A. Francke.

Marti, Werner (1985b). *Berndeutsch-Grammatik für die heutige Mundart zwischen Thun und Jura*. Bern: A. Francke.

Mohri, Mehryar (2004). „Weighted Finite-State Transducer Algorithms. An Overview“. In: *Formal Languages and Applications*. Hrsg. von Carlos Martín-Vide, Victor Mitrana und Gheorghe Păun. Bd. 148. Studies in Fuzziness and Soft Computing. Berlin und Heidelberg: Springer, S. 551–563.

Neumann, Günter (2010). In: *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Hrsg. von Kai-Uwe Carstersen, Christian Ebert, Cornelia Ebert, Susanne J. Jekat, Ralf Klabunde und Hagen Langer. 3. Aufl. Heidelberg: Spektrum Akademischer Verlag. Kap. Text-basiertes Informationsmanagement, S. 576–615.

Rehbein, Ines und Sören Schalowski (2013). „STTS goes Kiez - Experiments on Annotating and Tagging Urban Youth Language“. In: *JLCL* 28.1, S. 199–227. URL: http://www.jlcl.org/2013_Heft1/8Rehbein.pdf.

Rios, Anette und Richard Castro Mamani (2014). „Morphological Disambiguation and Text Normalization for Southern Quechua Varieties“. In: *COLING 2014, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Hrsg. von Marcos Zampieri, Liling Tan, Nikola Ljubešić und Jörg Tiedemann, S. 39–47. URL: http://www.aclweb.org/anthology/W/W14/W14-53.pdf.

Rumjanzewa, Marina (2013). *Die Verschriftlichung der Mundart*. Hrsg. von Neue Zürcher Zeitung. (online; aufgeschaltet am 1. Feb. 2013). URL: http://www.nzz.ch/feuilleton/die-verschriftlichung-der-mundart-1.17973385.

Scherrer, Yves (2007). „Phonetic Distance Measures for the Induction of a Translation Lexicon For Dialects“. Diplomarbeit. Universität Genf.

Scherrer, Yves (2011). „Morphology Generation for Swiss German Dialects“. In: *Systems and Frameworks for Computational Morphology - Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings*. Hrsg. von Cerstin Mahlow und Michael Piotrowski. Bd. 100. Communications in Computer and Information Science. Springer, S. 130–140. URL: https://archive-ouverte.unige.ch/unige:22778.

Scherrer, Yves und Owen Rambow (2010). „Natural Language Processing for the Swiss German Dialect Area“. In: *Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing, KONVENS 2010, September 6-8, 2010, Saarland University, Saarbrücken, Ger-*

many. Hrsg. von Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde und Angelika Storrer. universaar, Universitätsverlag des Saarlandes, S. 93–102.

Schiller, Anne, Simone Teufel, Christine Stöckert und Christine Thielen (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universitäten Tübingen und Stuttgart. Tübingen und Stuttgart. URL: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schmid, Helmut (1995). „Improvements in Part-of-Speech Tagging with an Application to German". In: *Proceedings of the EACL SIGDAT-Workshop*. (überarbeitete Version). URL: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schmid, Helmut und Florian Laws (2008). „Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging". In: *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*. Hrsg. von Donia Scott und Hans Uszkoreit, S. 777–784. URL: http://www.aclweb.org/anthology/C08-1098.

Siebenhaar, Beat und Alfred Wyler (1997). *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. 5., vollst. überarb. Aufl. (1. Aufl. 1984). Zürich: Edition „Pro Helvetia". URL: http://home.uni-leipzig.de/siebenh/pdf/Siebenhaar_Wyler_97.pdf.

Stark, Elisabeth, Simone Ueberwasser und Beni Ruef (2009–2015). *Swiss SMS Corpus*. URL: https://sms.linguistik.uzh.ch.

Suter, Rudolf (1992). *Baseldeutsch-Grammatik*. 3., überarb. Aufl. (1. Aufl. 1976). Grammatiken und Wörterbücher des Schweizerdeutschen in allgemeinverständlicher Darstellung. Basel: Christoph-Merian-Verlag.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister und Kathrin Beck (2015). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Universität Tübingen. URL: http://www.sfs.uni-tuebingen.de/fileadmin/user_upload/ascl/tuebadz-stylebook-1508.pdf.

Teufel, Simone und Christine Stöckert (1996). *ELM-DE: EAGLES Specifications for German morphosyntax: Lexicon Specification and Classification Guidelines*. URL: http://www.ilc.cnr.it/EAGLES96/pub/eagles/lexicons/elm_de.ps.gz.

Ueberwasser, Simone (2013). „Non-standard data in Swiss text messages with a special focus on dialectal forms". In: *Non-standard Data Sources in Corpus-based Research. (=TSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln 5. Hrsg: Christiane M. Bongartz und Claudia M. Riehl)*. Hrsg. von Marcos Zampieri und Sascha Diwersy. Aachen: Shaker Verlag, S. 7–24. URL: http://ueberwasser.eu/UeFiles/uni/Tagungen/

2012Koeln/ueberwasser.pdf.

Ueberwasser, Simone (2015a). *Normalization*. Hrsg. von sms4science.ch. (online, aufgerufen am 16. Februar 2016). URL: https://sms.linguistik.uzh.ch/bin/view/SMS4Science/Normalization.

Ueberwasser, Simone (2015b). *Part of speech tagging*. Hrsg. von sms4science.ch. (online, aufgerufen am 16. Februar 2016). URL: https://sms.linguistik.uzh.ch/bin/view/SMS4Science/PoS#German_40both_dialectal_and_non_45dialectal_41.

Van Pottelberge, Jeroen (2005). „Ist jedes grammatische Verfahren Ergebnis eines Grammatikalisierungsprozesses? Fragen zur Entwicklung des *am*-Progressivs". In: *Grammatikalisierung im Deutschen*. Hrsg. von Torsten von Leuschner, Tanja Mortelmans und Sarah Groodt. Berlin: De Gruyter, S. 169–192.

Weber, Albert und Bund Schwyzertütsch (1948). *Zürichdeutsche Grammatik: ein Wegweiser zur guten Mundart*. Grammatiken und Wörterbücher des Schweizerdeutschen in allgemeinverständlicher Darstellung. Zürich: Schweizer Spiegel-Verlag.

Wikimedia Commons, Hrsg. (2015). *File:Swiss German location.svg*. (online, aufgerufen am 27. Januar 2016). Wikimedia Foundation Inc. Los Angeles, California. URL: https://commons.wikimedia.org/wiki/File:Swiss_German_location.svg.

Zielinski, Andrea, Christian Simon und Tilman Wittl (2009). „Morphisto: Service-Oriented Open Source Morphology for German". In: *State of the Art in Computational Morphology - Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009, Zurich, Switzerland, September 2009. Proceedings*. Hrsg. von Cerstin Mahlow und Michael Piotrowski. Bd. 41. Communications in Computer and Information Science. Springer, S. 64–75.

A Tabellen

Tag	Morphologische Kategorien					Wortart
ADJA	Grad	Kasus	Numerus	Genus	Flexion	attributives Adjektiv
	[pcs*]	[rd]	[sp]	[mfn*]	[sw]	
	[p*]	[*]	[*]	[*]	[*]	(invariabel)
ADJD	Grad					prädikatives/adverbiales
	[pcs*]					Adjektiv
ADV						Adverb
APPO		Kasus				Postposition
APPR		[ad*]				Präposition
APPRART	Def.	Kasus	Numerus	Genus		Präposition mit Artikel
	[di]	[ad]	[sp]	[mfn*]		
APZR						rechter Teil einer Zirkumposition
ART	Def.	Kasus	Numerus	Genus		Artikel
	[di]	[rd]	[sp]	[mfn*]		
CARD						Kardinalzahl
FM						Fremdsprachiges Material
ITJ						Interjektion
KOUI						Konjunktion mit Infinitiv
KOUS						unterordnende Konjunktion
KON						nebenordnende Konjunktion
KOKOM						vergleichende Konjunktion
NN			Numerus	Genus		reguläres Substantiv
			[sp]	[mfn*]		
NN		Kasus	Numerus	Genus	Flexion	substantiviertes Adjektiv
		[rd]	[sp]	[mfn*]	[sw]	
NE			Numerus	Genus		Eigennamen
			[sp*]	[mfn*]		
PDAT		Kasus	Numerus	Genus		attr. Demonstrativpronomen
PDS		[rd]	[sp]	[mfn*]		substit. Demonstrativpronomen
PIAT		Kasus	Numerus	Genus		attr. Indefinitpronomen
PIDAT		[rd*]	[sp*]	[mfn*]		attr. Indefinitpronomen mit Determiner
PIS		([nad])				substit. Indefinitpronomen
PPER		Person	Numerus	Genus	Kasus	Personalpronomen
		[123]	[sp]	[mfn*]	[nad]	

weiter auf der nächsten Seite

Tag	Morphologische Kategorien				Wortart
PPOSAT	Kasus	Numerus	Genus		attr. Possessivpronomen
PPOSS	[rd]	[sp]	[mfn*]		substit. Possessivpronomen
PRELS					Relativpronomen
PRF	Person	Numerus	Genus	Kasus	Reflexivpronomen
	[123]	[sp]	[*]	[ad]	
PWAT	Kasus	Numerus	Genus		attr. Interrogativpronomen
PWS	[rd]	[sp]	[mfn*]		substit. Interrogativpronomen
PWAV					adverbiales Interrogativpronomen
PAV					Pronominaladverbien
PTKZU					Partikel *zu* vor Infinitiven
PTKNEG					Negationspartikel *nicht*
PTKVZ					Partikel bei Partikelverben
PTKANT					Antwortpartikel
PTKA					Partikel bei Adjektiv/Adverb
PTKAM					Partikel *am* bei Verlaufsform
PTKINF					Infinitivpartikeln *go, la, cho*
TRUNC					Kompositions-Erstglied
VAFIN	Person	Numerus	Modus		finite Hilfsverbform
VMFIN	[123]	[sp]	[ijk]		finite Modalverbform
VVFIN					finite Vollverbform
VAIMP		Numerus			Hilfsverb im Imperativ
VVIMP		[sp]			Vollverb im Imperativ
VAINF					Hilfsverb im Infinitiv
VMINF					Modalverb im Infinitiv
VVINF					Vollverb im Infinitiv
VVIZU					Verbinfinitiv mit *zu*
VAPP					Partizip Perfekt (Hilfsverb)
VMPP					Partizip Perfekt (Modalverb)
VVPP					Partizip Perfekt (Vollverb)
XY					Aussersprachliches
$,					Komma
$.					satzbeendendes Satzzeichen
$(					satzinterne Satzzeichen

Tabelle 23: Tagset *STTS.gsw* für Wortarten und morphologische Merkmale. Für jede Wortart sind die morphologischen Kategorien in der Reihenfolge angegeben, wie sie getaggt werden sollen. Für jede morphologische Kategorie sind auch die möglichen Belegungen aufgeführt.
Grad: p: Positiv; c: Komparativ; s: Superlativ
Person: 1: erste; 2: zweite; 3: dritte
Kasus: n: Nominativ; a: Akkusativ; d: Dativ; r: Nominativ/Akkusativ
Numerus: s: Singular; p: Plural
Genus: m: Maskulinum; f: Femininum; n: Neutrum
Modus: i: Indikativ; j: Konjunktiv I; k: Konjunktiv II
Flexion: s: stark; w: schwach
Definitheit: i: indefinit; d: definit

Zeichen	Aussprache	Beispiel		Mhd.	Kommentar
a	ɑ a	land	‚Land‘	land	
ā	ɑː ɔː oː	jār	‚Jahr‘	jâr	
ä	æ ɛ	nächt	‚Nächte‘	nähte	Sekundärumlaut von /a/
ǟ	æː ɛː	nǟchi	‚Nähe‘	næhe	manchmal auch [œː]
e	e	setzə	‚setzen‘	setzen	Primärumlaut von /a/
ē	eː ɛː	sē	‚See‘	sê	
ë	æ ɛ	hëlfə	‚helfen‘	hëlfen	
ə	ə	alləs	‚alles‘	alles	ə in Nebensilben
I	ɪ e	chInd	‚Kind‘	kint	
ī	iː i	zīt	‚Zeit‘	zît	
ĩ	iː eɪ̯ aɪ̯	frĩ	‚frei‘	vrî	Hiatusdiphthongierung
i	i	liəbi	‚Liebe‘	liebe	/i/ in Endungen
o	o ɔ	holz	‚Holz‘	holz	
ō	oː ɔː	rōt	‚rot‘	rôt	
ö	ø œ	hölzər	‚Hölzer‘	hölzer	
ő	øː œː	šő̄n	‚schön‘	schœne	
U	ʊ o	rUnd	‚rund‘	runt	
ū	uː u	hūs	‚Haus‘	mûre	
ũ	uː ɔʊ̯ aʊ̯	bũə	‚bauen‘	bûwen	Hiatusdiphthongierung
u	u				eventuell für /u/
Ü	ʏ ø	wÜnšə	‚wünschen‘	wünschen	
ǖ	yː y	fǖr	‚Feuer‘	viur	
ỹ	yː œɪ̯ ɔɪ̯	nỹ	‚neu‘	niuwe	Hiatusdiphthongierung
ü	y				eventuell für /y/
y	iː	šwyz	‚Schwyz‘	-	⟨y⟩ in Namen
y	y ʏ	typ	‚Typ‘	-	⟨y⟩ in Fremdwörtern
aĭ	aɪ̯ eɪ̯	haĭss	‚heiss‘	heiȝ	
aŭ	aʊ̯ ɔʊ̯	baŭm	‚Baum‘	boum	
äŭ	ɔɪ̯ œɪ̯	bäŭm	‚Bäume‘	böume	
iə	iə̯	liəbi	‚Liebe‘	liebe	
uə	uə̯	guət	‚gut‘	guot	
üə	yə̯	güəti	‚Güte‘	güete	

Tabelle 24: Vokalphoneme für die Überführung in verschiedene Dialekte. In der oberen Hälfte sind die Monophthonge, in der unteren die Diphthonge aufgeführt.

Zeichen	Aussprache	Beispiel		Mhd.	Kommentar
b	b	baŭm	‚Baum‘	boum	
ch	x kh	chInd	‚Kind‘	kint	nach hochalemannischem Lautstand
	x xː	machə	‚machen‘	machen	Lenis und Fortis
d	d t	dÜnn	‚dünn‘	dünne	wie im Standarddeutschen
	d	redə	‚reden‘	reden	
f	f	hafə	‚Hafen‘	haven	
ff	fː f	trëffə	‚treffen‘	treffen	wie im Standarddeutschen
g	g	guət	‚gut‘	guot	
gg	k	eggə	‚Ecke‘	ecke	
h	h	holz	‚Holz‘	holz	
j	j	jār	‚Jahr‘	jâr	
k	kx kh	kə	‚kein‘	kein	zu unterscheiden von /gh/
ck	kx kh	štÜck	‚Stück‘	stücke	wie im Standarddeutschen
l	l	mālə	‚malen‘	mâlen	
l	lː l	fallə	‚fallen‘	vallen	regional degeminiert
m	m	namə	‚Name‘	name	
mm	mː m	šwImmə	‚schwimmen‘	swimmen	regional degeminiert
n	n	maĭnə	‚meinen‘	meinen	
nn	nː n	chönnə	‚können‘	künnen	regional degeminiert
p	p	špIl	‚Spiel‘	spil	
pp	p	rappə	‚Rappen‘	rappe	wie im Standarddeutschen
qu	kxʊ	quëllə	‚Quelle‘	qwelle	
r	r	tiər	‚Tier‘	tier	
rr	rː r	charrə	‚Karren‘	karre	oft degeminiert
s	s	sī	‚sein‘	sîn	
ss	sː s	wassər	‚Wasser‘	waʒʒer	
š	ʃ ʃː	wäšə	‚waschen‘	waschen	
	ʃ	fešt	‚fest‘	veste	
t	t	guət	‚gut‘	guot	
tt	t	bIttə	‚bitten‘	bitten	wie im Standarddeutschen
v	f	vIl	‚viel‘	vil	wie im Standarddeutschen
	ʊ	vasə	‚Vase‘		in Fremdwörtern
w	ʊ	wassər	‚Wasser‘	waʒʒer	
x	ks	häx	‚Hexe‘	hecse	wie im Standarddeutschen
z	ts	zīt	‚Zeit‘	zît	
tz	ts	setzə	‚setzen‘	setzen	wie im Standarddeutschen

Tabelle 25: Konsonantenphoneme für die Überführung in verschiedene Dialekte.

B Teile des Programms

Bestandteile des Morphologieanalysesystems für Schweizerdeutsch *Taggswort*:

- Wörterlisten
  - `adjectives.lexc`
    Standarddeutsche Adjektivstämme nach Smor-Klassen
  - `adjectivesIrr.lexc`
    Schweizerdeutsche Stämme für unregelmässige Adjektive
  - `adpositions.lexc`
    Postpositionen, Zirkumpositionen und Präpositionen mit und ohne Artikel
  - `adverbs.lexc`
    Standarddeutsche Adverbstämme
  - `adverbsIrr.lexc`
    Schweizerdeutsche Adverbstämme
  - `articles.lexc`
    Bestimmter und unbestimmter Artikel
  - `cardinals.lexc`
    Kardinalzahlen
  - `cliticAccPron.lexc`
    Klitika von Personal- und Reflexivpronomina im Akkusativ
  - `cliticArt.lexc`
    Klitika von Artikeln
  - `cliticDatPron.lexc`
    Klitika von Personal- und Reflexivpronomina im Dativ
  - `cliticNomPron.lexc`
    Klitika von Personal- und Indefinitpronomina im Nominativ
  - `conjunctions.lexc`
    Konjunktionen
  - `interjections.lexc`
    Interjektionen

  - `interpunct.lexc`
    Satzzeichen
  - `nouns.lexc`
    Standarddeutsche Substantivstämme nach Smor-Klassen
  - `nounsIrr.lexc`
    Schweizerdeutsche Stämme für unregelmässige Substantive
  - `particles.lexc`
    Partikeln
  - `particlesVZ.lexc`
    Partikeln für Partikelverben
  - `pronouns.lexc`
    Pronomina
  - `verbs.lexc`
    Standarddeutsche Vollverben nach Smor-Klassen
  - `verbsIrr.lexc`
    Schweizerdeutsche unregelmässige Vollverben, Hilfsverben und Modalverben

- Transduktoren mit Ersetzungen
  - `cleanPre.xfst`
    Bereinigung vor der Konversion zum Schweizerdeutschen
  - `cleanMid.xfst`
    Bereinigung vor der Überführung in Dialektformen
  - `cleanPost.xfst`
    Bereinigung nach der Überführung in Dialektformen
  - `convertToGsw.xfst`
    Überführung in schweizerdeutsche abstrakte Phoneme
  - `dialectAbstract.xfst`
    Ausgangspunkt für weitere Dialekte
  - `dialectBE.xfst`
    Erzeugung der berndeutschen Lautformen
  - `dialectBS.xfst`
    Erzeugung der baseldeutschen Lautformen
  - `dialectZH.xfst`
    Erzeugung der zürichdeutschen Lautformen

- Sammlung aller Teile
  - `collection.xfst`
    Sammlung der einzelnen Wortarten aus `hfst`-Dateien und Durchfüh-

rung der Ersetzungen

– `Makefile`
Sammlung aller Befehle für die Erstellung der dialektspezifischen Systeme mit Hilfe des Buildsystems `make`

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Studiendekanat
Rämistr. 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass
die Masterarbeit von mir selbst und ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.lehre.uzh.ch/plagiate/20110314_LK_Merkblatt Plagiat.pdf).

.....

Ort und Datum Unterschrift

Lebenslauf

Persönliche Angaben

Name	Reto Flavio Baumgartner
Wohnort	Knonau ZH
E-Mail	retoflavio.baumgartner@uzh.ch
Geburtsdatum	8. April 1989

Studium

2009 – 2013	Bachelor-Studium an der Universität Zürich in Computerlinguistik und Sprachtechnologie, Skandinavistik und slavische Sprachwissenschaft
2013 – 2016	Master-Studium an der Universität Zürich in Computerlinguistik und Sprachtechnologie und Skandinavistik
Herbst 2014	Auslandsemester an der Universität Göteborg in Schweden