UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Thèse 2012

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

Generating Swiss German sentences from Standard German: a multi-dialectal approach

Scherrer, Yves

How to cite

SCHERRER, Yves. Generating Swiss German sentences from Standard German: a multi-dialectal approach. Doctoral Thesis, 2012. doi: 10.13097/archive-ouverte/unige:26361

This publication URL: https://archive-ouverte.unige.ch/unige:26361
Publication DOI: 10.13097/archive-ouverte/unige:26361

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES LETTRES

Doctoral Thesis

Generating Swiss German sentences from Standard German: a multi-dialectal approach

Yves Scherrer

September 28, 2012

Supervisor: Prof. Éric Wehrli

Acknowledgements

First of all, I would like to thank my parents for providing me with native speaker compentence in Swiss German dialects, without which this thesis likely would have headed in an entirely different way. While they did not actively choose Swiss German as my first language, they supported me throughout my childhood in all language-related matters, including writing pseudo-Dutch, adding graphemes to German spelling to improve the phoneme-grapheme mapping, and my first contacts with Romance languages in the Grisons and Tessin. I am very grateful to them, and to my sister Sylvie, for their continuing support throughout my studies.

Most of the work on this thesis has been carried out within the Department of linguistics at the University of Geneva, and I would like to thank all its members for their support, for interesting discussions, and for lunch company. Especially, I would like to thank my supervisor, Eric Wehrli, who accepted hiring me as an assistant, and who gladly accepted my thesis proposal even if it was only marginally related to *Fips*. I am also indebted to Paola Merlo, who supervised my DEA thesis on a different aspect of Swiss German language processing and thereby introduced me to scientific research methodology.

Special thanks are also due to my office mates for putting up with me throughout the years I spent in Geneva: Violeta Seretan, Jean-Philippe Goldman, Mar Ndiaye, and Kamel Nebhi. I am most grateful to Sharid Loáiciga for her affection far beyond the office walls and for her support during the busiest periods of my thesis writing. I also enjoyed the company of great colleagues in other offices of the seventh floor: Alexis Kauffmann, Tanja Samardzic, Lorenza Russo, Asheesh Gulati, Luka Nerima, Christopher Laenzlinger, Gabi Soare, Lena Baunaz, Antonio Leoni de León, Gabriele Musillo, Lonneke van der Plas, Bruno Cartoni, Anamaria Bentea, Richard Zimmermann, Eva Capitao, and all the others.

In 2009-2010, I had the opportunity to work on my thesis at the Center for Computational Learning Systems at Columbia University, New York. I am deeply indebted to Owen Rambow, who accepted to receive me during this year, and whose discussions and comments were very helpful for the advancement of my research. I also thank all the senior researchers, graduate students, and post-docs whose company I enjoyed during this year: Nizar Habash, Yassine Benajiba, Marine Carpuat, Yuval Marton, Daniel Bauer, Ahmed El Kholy, Mohamed Altantawy, Apoorv Agarwal, Vinod Kumar P G, Boyi Xie, Weiwei Guo, Raphael Pelossof, Josh Gordon, and all the others. This year abroad was made possible through a grant from the Swiss National Science Foundation (Grant No. PBGEP1-125929). I am grateful for this funding.

My thanks also go to the people that have supported me in earlier stages of my academic career at the University of Neuchâtel. Some of them have preceded or followed me by moving their academic center of interest to Geneva: Stéphane Borel, Sandra Schwab and Gisela Zingg.

I also appreciated the kind assistance of various researchers at the University of Zurich, who readily granted me access to different types of data and answered many questions about these. My gratitude is due to Elvira Glaser, Martin Volk, Anne Göhring, Agnes Kolmer, Gabi Bart, Noah Bubenhofer, Michael Piotrowski, Beni Ruef, and Alexandra Bünzli.

The Department of Geography at the University of Geneva allowed me to sit in their cartography class in 2008-2009 and granted me a site licence for the ArcGIS software. Likewise, I thank the Groningen people (Peter Kleiweg and Cagri Coltekin) for making available their great dialectometry software tools. I am also indebted to Hans Goebl, who provided me with an alternative dialectometric interpretation of German-speaking Switzerland. Unfortunately, the results of this recent collaboration materialized too late to be incorporated in this thesis.

Parts of the work reported in this thesis have been presented at various conferences and workshops. In this sense, this thesis has benefitted from the comments of anonymous reviewers and from interesting discussions with conference participants. I wish to thank all these people for their contribution.

The last persons to influence this thesis will likely be the members of the Committee. I thank Elvira Glaser, Paola Merlo, John Nerbonne, and Owen Rambow for accepting this task, and Jacques Moeschler for accepting to act as the President of the Jury.

Abstract

This thesis proposes to combine methods and data from two rather distant fields of language science – dialectology and human language technology – into a system that automatically transforms Standard German words and sentences into multiple Swiss German dialects. The resulting system is designed to benefit from the etymological proximity of these language varieties and from data made available by dialectological research. We qualify this scenario as **cross-lingual**, **multi-dialectal** sentence generation: cross-lingual in the sense that it describes the relation between two vertically opposed language varieties, Standard German and Swiss German, and multi-dialectal in the sense that the Swiss German side of the model is intended to cover the whole range of internal dialectal variation. While it would be most intuitive to generate dialect output in its naturally occurring spoken form, we focus here on written output.

Our work draws on previous research in two different and complementary ways. On the one hand, we use, develop and extend methods that have been proposed to describe dialects (theoretically or computationally), their variation patterns and their relation with a standard variety. In particular, our methodology recalls earlier theoretical work in generative dialectology, which views dialect change as processes formalized as rules, and posits a reference system from which the dialectal variants are derived. On the other hand, we rely on studies of the dialectological and socio-linguistic particularities of language use in German-speaking Switzerland. We draw on these resources, and in particular on the Swiss German dialect atlases *Sprachatlas der deutschen Schweiz* and *Syntaktischer Atlas der deutschen Schweiz*, to compile a **comparative multi-dialect grammar** of Swiss German as an empirical foundation of our system.

The key concept put forward in this thesis is the one of **georeferenced transfer rule**. Indeed, the relation between Standard German and Swiss German is modelled with the help of transfer rules operating on several levels of linguistic analysis. In order to account for the dialectal variation,

each rule (corresponding to one linguistic phenomenon, or **variable**) produces several dialectal **variants** with complementary areal extents. The inclusion of areal information in a linguistic rule is called georeferencing. Areal information is represented in the form of probability maps derived from the abovementioned dialectological atlases.

A large part of the thesis is dedicated to the detailed study of various aspects of the implementation of these georeferenced rules, such as:

- the different levels of linguistic analysis covered by the rules,
- the selection of linguistic variables for each analysis level,
- the creation of probability maps for the georeferencing,
- the technical choices for the implementation,
- the evaluation of the resulting system and its performance.

These aspects are discussed briefly in the following paragraphs.

In practice, cross-lingual sentence generation is an instance of machine translation. Machine translation systems usually operate on lexical and syntactic levels, by looking up word correspondences in a dictionary and arranging the target words in a syntactically correct way. Some systems perform a morpho-syntactic analysis of the source text and generate the target text according to the extracted features. A key assumption of our work, following earlier research on closely related languages, is that the lexical component (the bilingual dictionary) can be partially replaced by phonetic or graphemic transformations. These account for the etymologically related word pairs in which only a few letters have to be changed, and confine the use of explicit dictionary entries to exceptions and etymologically unrelated word pairs. In summary, the proposed system will operate on four levels of linguistic analysis: phonetics and phonology (for the stems of regularly derived words), lexicon (for the stems of irregularly derived words), morphology (to generate inflected word forms), and syntax. In terms of implementation, we distinguish the so-called word-level rules covering the former three levels from the syntax rules.

The selection of linguistic variables is complicated by the georeferencing requirement: we only consider phenomena that either have been treated in one of the dialect atlases or lack diatopical

(i.e., geographical) variation. Moreover, we discard phenomena that do not satisfy our criteria of frequency of use, diachronic stability, and feasability of implementation.

The selected atlas maps are scanned (where necessary) and digitally processed with a Geographical Information System. In the course of the discussion of the relevant cartographic techniques, we introduce the distinction between point maps and surface maps and argue that surface maps are more adequate for our use of dialect data than the point maps originally found in the atlases. Consequently, we present different interpolation methods that transform the point maps into probabilistic surface maps.

The word-level transformations are dealt with in two alternative implementations. In the first implementation, the transfer rules are conceived as regular expressions and stored in a database. In the second implementation, we use finite-state transducers to represent the rules; in the latter, georeferencing is done with a special theoretical device called **flag diacritics**. The syntactic transformations are implemented with scripts operating on syntactically annotated text. All rules are built by hand in order to facilitate the georeferencing with existing maps. Several web applications are provided for the casual use of the dialect word generation system.

The evaluation of a multi-dialectal system must operate on two levels, a linguistic one and a geographical one: it must not only be ensured that the produced output forms are linguistically correct, but also that their predicted area of occurrence is dialectologically adequate. Hence, we create a multi-dialectal corpus, based on *Wikipedia* articles from several Swiss German dialects, in order to assess different aspects of the translation system. First, we evaluate the coverage of the word-level transfer rules on the basis of word lists, and then use the same word-level rules to translate entire sentences. Furthermore, we illustrate how the geographical information contained in the word-level rules can be used to identify the dialectal origin of a text. Finally, we evaluate the accuracy of the syntactic transfer rules, and show that the integration of the latter positively impacts the performance of the translation system. While the results obtained in these experiments are encouraging, they can be further improved, for instance by validating the rule coverage empirically.

Finally, we present several dialectometrical analyses of Swiss German data. Dialectometry has emerged as a field of linguistics that investigates the application of statistical and mathematical methods in dialect research. As a consequence of our map digitization efforts, we are able to

present original results for the Swiss German dialect landscape. By using different data sets and different methods of analysis, we provide at the same time an overview of the state of the art in dialectometry and of the dialectological subdivision of German-speaking Switzerland. Our experiments give interesting insights into the Swiss German dialect landscape and suggest that this line of research is worthwhile pursuing.

Résumé en français

Cette thèse se propose de combiner des méthodes et des données de deux domaines plutôt éloignés des sciences du langage – la dialectologie et le traitement automatique du langage – dans un système qui transforme des mots et des phrases en allemand standard en une multitude de dialectes suisse allemands. De par sa conception, ce système bénéficie de la proximité étymologique de ces variétés linguistiques et des données mises à disposition par la recherche en dialectologie. Nous qualifions ce scénario de génération de phrases comme **interlingual et multidialectal** : interlingual puisqu'il décrit la relation entre deux variétés linguistiques en opposition verticale (l'allemand standard et le suisse allemand), et multidialectal puisque le côté suisse allemand du modèle est destiné à couvrir l'étendue entière de la variation dialectale interne. Bien qu'il serait plus naturel de générer des énoncés dialectaux dans leur forme orale, nous nous focalisons ici sur l'écrit.

Notre travail se base sur des travaux de recherche antérieurs de deux manières différentes et complémentaires. D'un côté, nous utilisons, développons et étendons des méthodes qui ont été proposées pour la description (théorique ou computationnelle) des dialectes, de leurs types de variation, et de leur relation avec une variété standardisée. En particulier, notre méthodologie rappelle celle de la dialectologie générative, qui voit les changements dialectaux comme des processus formalisés à l'aide de règles, et qui postule un système de référence à partir duquel les variantes dialectales sont dérivées. De l'autre côté, nous nous appuyons sur des études des particularités dialectologiques et sociolinguistiques de l'utilisation langagière en Suisse alémanique. Nous nous basons sur ces ressources, et notamment sur les atlas dialectologiques *Sprachatlas der deutschen Schweiz* et *Syntaktischer Atlas der deutschen Schweiz*, afin de compiler une **grammaire comparative multidialectale** du suisse allemand, qui servira de base empirique à notre système.

Le concept-clé mis en avant dans cette thèse est celui de **règle de transfert géoréférencée**. En effet, la relation entre l'allemand standard et le suisse allemand est modélisée à l'aide de règles de transfert qui opèrent à différents niveaux d'analyse linguistique. Afin de prendre en compte la variation dialectale, chaque règle (correspondant à un phénomène, ou une **variable**, linguistique) produit plusieurs **variantes** dialectales avec des étendues spatiales complémentaires. L'intégration d'informations géographiques dans une règle linguistique est appelée géoréférencement. L'information géographique est représentée sous forme de cartes probabilistes, qui sont dérivées des atlas dialectologiques susmentionnés.

Une grande partie de la thèse est dédiée à l'étude détaillée de divers aspects de l'implémentation de ces règles géoréférencées, comme par exemple :

- les différents niveaux d'analyse linguistique couvertes par les règles,
- la sélection des variables linguistiques pour chacun de ces niveaux,
- la création de cartes probabilistes pour le géoréférencement,
- les choix techniques de l'implémentation,
- l'évaluation du système résultant et ses performances.

Ces aspects sont discutés brièvement dans les paragraphes suivants.

Dans la pratique, la génération interlinguale de phrases est un cas particulier de traduction automatique. Généralement, les systèmes de traduction automatique opèrent aux niveaux lexical et syntaxique : ils recherchent des correspondances de mots dans un dictionnaire bilingue et arrangent les mots cibles dans un ordre syntaxiquement correct. Quelques systèmes effectuent une analyse morpho-syntaxique du texte source et génèrent le texte cible en fonction des traits extraits du premier. Une hypothèse clé de notre travail – à la suite de travaux antérieurs sur des langues étymologiquement proches – est que la composante lexicale (le dictionnaire bilingue) peut être partiellement remplacé par des transformations phonétiques ou graphémiques. Celles-ci tiennent compte des paires de mots étymologiquement reliés, dans lesquelles seulement quelques lettres doivent être remplacées. Ainsi, elles et restreignent l'utilisation d'un dictionnaire explicite aux exceptions et aux paires de mots étymologiquement non reliés. En résumé, le système proposé opérera sur quatre niveaux d'analyse linguistique : phonétique et phonologie (pour les radicaux des mots à dérivation régulière), lexique (pour les radicaux des

mots à dérivation irrégulière), morphologie (pour générer des formes fléchies), et syntaxe. A des fins d'implémentation, nous distinguons les règles à l'échelle du mot, couvrant les trois premiers niveaux, des règles syntaxiques.

La sélection des variables est compliquée par l'exigence du géoréférencement : nous considérons seulement les phénomènes qui ont été traités dans un atlas dialectologique et ceux qui ne présentent pas de variation diatopique (géographique). De plus, nous rejetons les phénomènes qui ne satisfont pas à nos critères de fréquence d'utilisation, de stabilité diachronique, et d'implantabilité.

Les cartes sélectionnées sont numérisées (si nécessaire) et transformées à l'aide d'un système d'informations géographiques. Dans le cadre de la discussion des techniques cartographiques pertinentes, nous introduisons la distinction entre cartes de points et cartes de surfaces, et nous soutenons que les cartes de surfaces sont plus adéquats pour notre utilisation de données dialectales que les cartes de points originellement présents dans les atlas. Par conséquent, nous présentons différentes méthodes d'interpolation qui transforment les cartes de points en cartes de surfaces probabilistes.

Nous proposons deux implantations alternatives des transformations à l'échelle du mot. Dans la première, les règles de transfert sont conçues comme des expressions régulières et stockées dans une base de données. Dans la seconde, les règles sont représentées par des transducteurs à états finis, et le géoréférencement est effectué à l'aide du concept de **flag diacritics**. Les transformations syntaxiques sont implantées à l'aide de scripts qui nécessitent du texte syntaxiquement annoté. Toutes les règles sont créées manuellement, afin de faciliter le géoréférencement avec des cartes existantes. Plusieurs applications web sont fournies pour une utilisation occasionnelle du système de génération de mots dialectaux.

L'évaluation d'un système multidialectal doit opérer sur deux niveaux, un niveau linguistique et un niveau géographique : il ne faut pas seulement s'assurer que les formes produites sont linguistiquement correctes, mais aussi que leur aire d'apparition prédite soit conforme à la réalité dialectologique. Ainsi, nous créons un corpus multidialectal, basé sur des articles de *Wikipedia* pour différents dialectes suisse allemands, afin d'évaluer différents aspects du système de traduction. D'abord, nous évaluons la couverture des règles à l'échelle du mot sur la base de listes de mots. Ensuite, nous utilisons les mêmes règles pour traduire des phrases entières. De

plus, nous illustrons comment l'information géographique contenue dans les règles à l'échelle du mot peut être utilisée pour identifier l'origine dialectale d'un texte. Enfin, nous évaluons l'exactitude des règles de transfert syntaxiques et montrons que l'intégration de ces dernières a un impact positif sur les performances globales du système de traduction. Bien que les résultats obtenus dans ces expériments soient encourageants, ils peuvent être améliorés, par exemple à travers la validation empirique de la couverture des règles.

Enfin, nous présentons plusieurs analyses dialectométriques de données suisse allemandes. La dialectométrie s'est positionnée comme un domaine de la linguistique qui étudie l'application de méthodes statistiques et mathématiques dans la recherche dialectologique. Suite à nos efforts de numérisation de cartes, nous pouvons présenter des résultats originaux pour l'aire dialectale suisse allemande. En utilisant des jeux de données différents et des méthodes d'analyse différentes, nous fournissons une vue d'ensemble à la fois de l'état de l'art en dialectométrie et la division dialectologique de la Suisse alémanique. Nos expériments donnent un éclairage intéressant sur le paysage dialectologique suisse allemand et suggèrent que cet axe de recherche mérite d'être poursuivi à l'avenir.

Zusammenfassung auf deutsch

In dieser Dissertation wird vorgeschlagen, die Methoden und Daten von zwei weit entfernten Forschungsgebieten der Sprachwissenschaft – der Dialektologie und der Maschinellen Sprachverarbeitung – in einem System zu vereinen, das standarddeutsche Wörter und Sätze automatisch in verschiedene schweizerdeutsche Dialekte überträgt. Das resultierende System ist darauf ausgelegt, die etymologische Nähe dieser Sprachvarietäten und die durch dialektologische Forschung verfügbar gemachten Daten zu nutzen. Wir bezeichnen dieses Szenario als **sprachübergreifende, multidialektale** Sprachgenerierung: sprachübergreifend, da es die Beziehung zwischen zwei sich auf der vertikalen Ebene unterscheidenden Sprachvarietäten (Standarddeutsch und Schweizerdeutsch) beschreibt, und multidialektal, da die schweizerdeutsche Seite des Modells die gesamte interne dialektale Variation umfassen soll. Obwohl es am natürlichsten wäre, Dialektäusserungen in gesprochener Form zu generieren, beschränken wir uns hier auf die schriftliche Ausgabe.

Diese Arbeit stützt sich auf bisherige wissenschaftliche Arbeiten in zwei unterschiedlichen, komplementären Hinsichten. Einerseits benutzen, entwickeln und erweitern wir Methoden, die für die (theoretische oder maschinelle) Beschreibung von Dialekten, deren Variationsmuster und deren Verhältnis zu einer Standardvarietät vorgeschlagen worden sind. Namentlich gleicht unsere Methodologie derjenigen der generativen Dialektologie, die Dialektwandel als durch Regeln formalisierte Prozesse ansieht, mit welchen die verschiedenen dialektalen Varianten von einem Referenzsystem abgeleitet werden. Andererseits berufen wir uns auf Studien der dialektologischen und soziolinguistischen Eigenschaften des schweizerdeutschen Sprachgebrauchs. Wir benutzen diese Ressourcen, und insbesondere die Dialektatlanten *Sprachatlas der deutschen Schweiz* und *Syntaktischer Atlas der deutschen Schweiz*, um eine **vergleichende multidialektale Grammatik** des Schweizerdeutschen zusammenzutragen, die uns dann als empirische

Grundlage des Übersetzungssystems dient.

Das zentrale Konzept, das in dieser Dissertation vorgeschlagen wird, ist das der **georeferenzierten Transferregel**. In der Tat wird die Beziehung zwischen Standarddeutsch und Schweizerdeutsch mit Hilfe von Transferregeln modelliert, die auf verschiedenen linguistischen Ebenen arbeiten. Die dialektale Variation wird berücksichtigt, indem jede Regel (einem sprachlichen Phänomen, oder einer sprachlichen Variable, entsprechend) mehrere dialektale Varianten mit komplementären räumlichen Ausdehnungen erzeugt. Die Integration von räumlichen Informationen in eine sprachliche Regel wird Georeferenzierung genannt. Räumliche Informationen werden als Wahrscheinlichkeitskarten dargestellt, die aus den obengenannten Dialektatlanten extrahiert werden.

Ein grosser Teil dieser Dissertation widmet sich der detaillierten Analyse von verschiedenen Aspekten der Implementation dieser georeferenzierten Regeln, wie zum Beispiel:

- die durch die Regeln abgedeckten Analyseebenen,
- die Auswahl der sprachlichen Variablen für jede Analyseebene,
- die Erstellung von Wahrscheinlichkeitskarten für die Georeferenzierung,
- die technischen Details der Implementation,
- die Evaluation des resultierenden Systems und dessen Leistung.

Diese Aspekte werden in den folgenden Abschnitten kurz angesprochen.

Die sprachübergreifende Satzgenerierung kann als eine Instanz der maschinellen Übersetzung angesehen werden. Üblicherweise arbeiten maschinelle Übersetzungssysteme auf der lexikalischen und syntaktischen Ebene, indem Wörter in einem zweisprachigen Wörterbuch nachgeschlagen werden und die Wörter der Zielsprache syntaktisch korrekt angeordnet werden. Manche Systeme führen eine morpho-syntaktische Analyse des Quelltextes durch und generieren den Zieltext aufgrund der von ersterem extrahierten Merkmale. Im Anschluss an bisherigen computerlinguistische Studien zu nahe verwandten Sprachen argumentieren wir in unserer Arbeit, dass die lexikalische Komponente (das zweisprachige Wörterbuch) teilweise durch phonetische und graphemische Transformationen ersetzt werden kann. Diese behandeln die etymologisch verwandten Wortpaare, in denen nur ein paar wenige Buchstaben verändert

werden müssen, und beschränken die Benutzung von expliziten Wörterbucheinträgen auf Ausnahmen und etymologisch nicht verwandte Wortpaare. Das vorgeschlagene System arbeitet auf vier Analyseebenen: Phonetik und Phonologie (für die Stämme regelmässig abgeleiteter Wörter), Lexikon (für die Stämme unregelmässig abgeleiteter Wörter), Morphologie (für die Generierung flektierter Wortformen), und Syntax. In Bezug auf die Implementierung unterscheiden wir die sogenannten Regeln auf Wortebene, die die ersten drei Ebenen umfassen, von den Syntaxregeln.

Die Auswahl der sprachlichen Variablen wird durch die Erfordernisse der Georeferenzierung erschwert: wir berücksichtigen nur Phänomene, die entweder in einem der Dialektatlanten behandelt wurden oder keinerlei diatopische (d.h. geografische) Variation aufweisen. Darüber hinaus scheiden wir diejenigen Phänomene aus, die unsere Kriterien von Nutzungshäufigkeit, diachronischer Stabilität und Implementierbarkeit nicht erfüllen.

Für die Transformationen auf Wortebene schlagen wir zwei alternative Implementationen vor. In der einen werden die Transferregeln als reguläre Ausdrücke betrachtet und in einer Datenbank gespeichert. In der anderen werden die Regeln mittels endlicher Transduktoren formalisiert; in diesem Ansatz wird die Georeferenzierung mit Hilfe von sogenannten **flag diacritics** bewerkstelligt. Die Syntaxtransformationen werden mit Skripts implementiert, die syntaktisch annotierten Text benötigen. Alle Regeln werden von Hand erstellt, was die Integration existierender Karten zur Georeferenzierung vereinfacht. Mehrere Web-Applikationen werden für die gelegentliche Benutzung des Wortgenerierungssystems angeboten.

Die Evaluierung eines multidialektalen Systems muss auf zwei Ebenen geschehen, der linguistischen und der geografischen: es muss nicht nur gewährleistet werden, dass die Ausgabe des Systems linguistisch korrekt ist, sondern auch, dass deren Verbreitungsgebiet dialektologisch adäquat vorhergesagt wird. Daher stellen wir ein multidialektales Korpus auf der Basis von *Wikipedia*-Artikeln von verschiedenen schweizerdeutschen Dialekten zusammen, um die verschiedenen Aspekte des Übersetzungssystems zu beurteilen. Zuerst evaluieren wir die Abdeckung der Transferregeln auf Wortebene aufgrund von Wortlisten und wenden dann dieselben Regeln an, um ganze Sätze zu übertragen. Darüber hinaus zeigen wir, wie die in den Regeln enthaltenen geografischen Informationen benutzt werden können, um die dialektale Herkunft eines Textes zu bestimmen. Schliesslich evaluieren wir die Genauigkeit der syntaktischen Transferre-

geln und zeigen, dass sich deren Integration positiv auf die Leistung des Übersetzungssystems auswirkt. Obschon die in diesen Experimenten erreichten Resultate viel versprechend sind, können sie noch weiter verbessert werden, beispielsweise durch die empirische Validierung der Regelabdeckung.

Zum Schluss präsentieren wir mehrere dialektometrische Analysen mit Deutschschweizer Daten. Die Dialektometrie hat sich zu dem Bereich der Sprachwissenschaft entwickelt, der die Anwendung statistischer und mathematischer Methoden in der Dialektforschung untersucht. Als Ergebnis unserer Bemühungen zur Kartendigitalisierung können wir originelle Studien für die schweizerdeutsche Dialektlandschaft vorstellen. Durch die Verwendung verschiedener Datensätze und verschiedener Analysemethoden können wir eine Übersicht sowohl des heutigen Standes der Technik als auch der dialektologischen Unterteilung der deutschen Schweiz bieten. Unsere Experimente gewähren interessante Einblicke in die schweizerdeutsche Dialektlandschaft und regen zu weiterer Forschung in diesem Bereich an.

Contents

Acknowledgements **3**

Abstract **5**

Résumé en français **9**

Zusammenfassung auf deutsch **13**

1. Introduction **27**

1.1. Goals 27

1.2. Swiss German dialects 29

1.2.1. Dialects, varieties and languages: some definitions 29

1.2.2. Variables, variants and isoglosses 32

1.2.3. Languages in Switzerland 33

1.2.4. German dialects 35

1.2.5. The dialectological division of Swiss German dialects 39

1.2.6. Diglossia 41

1.2.7. Impacts on cross-lingual, multi-dialectal sentence generation 42

1.2.8. Some words about language change and borrowing 47

1.3. Outline 48

2. Related Work **51**

2.1. Introduction 51

2.2. A short history of dialectology 52

2.2.1. The constitution of the discipline 53

2.2.2. Big enterprises 54

2.2.3. New theories and new technologies . . . 57
2.2.4. Summary . . . 61
2.3. A brief history of Human Language Technologies . . . 61
2.4. Central pieces of related research . . . 68
2.5. Closely related languages and cognate words . . . 72
2.6. Dialectometry . . . 74
2.7. Computational dialectology . . . 75
2.8. Spoken dialect . . . 77
2.9. Computational processing of historical texts . . . 78
2.10. Natural Language Processing of Standard German . . . 78
2.11. Conclusion . . . 82

3. Data sources **85**
3.1. Overview . . . 85
3.2. Spelling . . . 86
3.2.1. Standard German spelling . . . 87
3.2.2. The Dieth spelling rules . . . 88
3.2.3. Conclusion . . . 89
3.3. *Sprachatlas der deutschen Schweiz* . . . 89
3.3.1. History of the project . . . 89
3.3.2. Contents . . . 94
3.3.3. Kleiner Sprachatlas . . . 95
3.3.4. Using SDS data for our work . . . 97
3.3.5. Summary . . . 105
3.4. *Syntaktischer Atlas der deutschen Schweiz* . . . 105
3.4.1. Methodology . . . 105
3.4.2. Contents . . . 108
3.4.3. Using SADS data in our work . . . 109

4. Maps **111**
4.1. Introduction . . . 111
4.2. Geographical Information Systems . . . 111
4.3. Digitizing SDS maps . . . 113

4.4. Importing SADS data . . . 119
4.5. Dialectal variation: discrete or continuous? . . . 121
4.6. Tessellation and interpolation . . . 123
4.6.1. Tessellation . . . 124
4.6.2. Interpolation methods . . . 126
4.6.3. Bandwidth estimation . . . 131
4.6.4. Value normalization . . . 133
4.6.5. Map algebra . . . 135
4.7. Web access . . . 135
4.8. Conclusion . . . 136

5. Georeferenced transfer rules **139**
5.1. Introduction . . . 139
5.2. Database implementation of word-level transfer rules . . . 140
5.2.1. General system architecture . . . 140
5.2.2. Morpho-syntactic analysis and compound splitting . . . 143
Fips . . . 143
TreeTagger . . . 144
Compound splitting with Morphisto . . . 145
Summary . . . 146
5.2.3. Lexical rules . . . 147
5.2.4. Inflection classes . . . 149
5.2.5. Regular expressions . . . 149
5.2.6. Phonetic transformations . . . 152
5.2.7. Different application contexts of phonetic transformations . . . 153
5.2.8. Phonetic or graphemic rules? . . . 154
5.2.9. Lexical restrictions of phonetic transformations . . . 156
5.2.10. Affix generation . . . 157
5.2.11. Suppletive morphology: short verbs . . . 158
5.2.12. Rule combination . . . 159
5.2.13. A full example . . . 160
5.2.14. Limitations of the database approach . . . 165

5.3. Finite-state implementation of word-level transfer rules . . . 166
5.3.1. Finite-state methods . . . 166
Finite-state automata . . . 167
Finite-state transducers . . . 169
Composing finite-state machines . . . 170
Weighted transducers . . . 171
5.3.2. Georeferencing with flag diacritics . . . 172
5.3.3. The *xfst* toolkit . . . 174
5.3.4. General architecture . . . 175
Derivational affixes . . . 176
Pre-annotation of phoneme classes . . . 176
Simplified application of phonetic rules . . . 178
Pruning at the end of the derivation process . . . 179
5.3.5. Determining lexical restrictions from Middle High German data . . . 179
5.3.6. Conclusion . . . 181
5.4. Syntactic transfer rules . . . 183
5.4.1. The CoNLL-X annotation format . . . 183
5.4.2. Rule implementation . . . 185
5.4.3. An example . . . 188
5.4.4. Converting *Fips* output to the *CoNLL-X* format . . . 189
5.4.5. Conclusion . . . 190
5.5. Data structures . . . 191
5.6. Summary . . . 193

6. Experiments **197**
6.1. Introduction . . . 197
6.2. The *Wikipedia* corpus . . . 197
6.3. Coverage of the word-level transfer rules . . . 200
6.3.1. Creation of the morphological dictionary . . . 200
6.3.2. Evaluation . . . 202
6.3.3. Comparison of the word-level implementations . . . 203
6.3.4. Error analysis . . . 205

6.4. Translation of entire sentences . . . 206
6.4.1. Measures of linguistic distance and similarity . . . 208
6.4.2. Measuring language distance and translation difficulty . . . 209
6.4.3. Measuring translation performance and localization precision . . . 211
6.4.4. Discussion . . . 212
6.4.5. Conclusion . . . 215
6.5. Shibboleth-based dialect identification . . . 216
6.5.1. Introduction and related work . . . 217
6.5.2. The model . . . 219
Word lookup and dialect identification . . . 219
Computing the probability map for a word . . . 220
Computing the probability map for a segment . . . 222
6.5.3. Experiments and results . . . 222
Data . . . 222
Baseline: N-gram model . . . 223
Our model . . . 225
Variations . . . 229
6.5.4. Discussion . . . 231
6.6. Correctness of the syntactic transfer rules . . . 231
6.6.1. Corpus frequencies . . . 232
6.6.2. Testing the syntax rules with the TIGER treebank . . . 232
6.6.3. Analysis of transformation errors . . . 235
6.6.4. Testing the syntax rules with *Fips* annotations . . . 237
6.6.5. Analysis of *Fips* parsing errors . . . 239
6.6.6. Summary . . . 242
6.7. Putting the pieces together . . . 242
6.8. Conclusion . . . 245

7. Dialectometry **247**
7.1. Introduction . . . 247
7.2. Methods . . . 248
7.2.1. Distance matrices . . . 249
7.2.2. Spatial autocorrelation . . . 251

7.2.3. Clustering . . . 253
7.2.4. Multidimensional scaling . . . 255
7.3. Experiments . . . 258
7.3.1. Inferring dialect landscapes with SDS data . . . 258
7.3.2. The influence of linguistic levels on the inferred dialect landscapes . . . 263
7.3.3. Inferring dialect landscapes with SADS data . . . 265
7.3.4. The influence of the inquiry technique on the inferred dialect landscapes 266
7.3.5. Localizing dialect texts . . . 268
The Archimob corpus . . . 269
Computing the linguistic similarity of two comparable texts . . . 271
Cognate word pairs and identical word pairs . . . 272
Normalization by text length . . . 272
Evaluation and visualisation . . . 274
7.4. Conclusion . . . 279

8. Conclusion **281**
8.1. What we have achieved . . . 281
8.2. What we could have done but haven't . . . 283
8.3. What remains to be done . . . 286
8.4. Epilogue . . . 287

A. A comparative multi-dialect grammar of Swiss German **289**
A.1. Introduction . . . 289
A.1.1. Geographical references . . . 290
A.2. Phonetics and phonology . . . 291
A.2.1. Vowels . . . 292
Standard German diphthongs . . . 292
Standard German long monophthongs . . . 296
Closing . . . 297
Opening . . . 297
Unrounding . . . 298
Shortening . . . 298
Lengthening . . . 299

Schwa syncope and apocope . . . 300
Schwa quality . . . 301
Semi-vowel insertion in hiatus . . . 301
General vowel quality . . . 302
Umlaut as a morpho-phonological process . . . 303
A.2.2. Consonants . . . 304
The pronunciation of *ch* and *s* . . . 304
The outcomes of Standard German *k* . . . 305
Aspiration of initial *p* and *t* . . . 306
Lenition and Fortification . . . 307
Intervocalic consonant gemination . . . 309
l-vocalization . . . 309
Nasal and glide gemination . . . 310
s and *sch* . . . 310
st and *sp* . . . 311
n-drop and *n*-insertion . . . 311
Staub's law . . . 312
nd . . . 313
chs . . . 314
rn, *rm* and *ln* . . . 314
A.3. Derivational morphology . . . 315
Noun diminutives . . . 315
Verb diminutives . . . 316
Derivation of diminutive adjectives . . . 317
Adjectives ending in *-end, -ern, -en* . . . 317
Verbal prefixes . . . 318
Nouns ending in *-ung* . . . 319
Nouns ending in *-ling* . . . 319
Nouns ending in *-end* . . . 319
Phonetically reduced derivational suffixes . . . 320
Abstract nouns . . . 320
Deverbal intensifying affixes . . . 321
Masculine agent nouns . . . 321

Feminine derivation of agent nouns . . . 321
Compound nouns and *Kompositionsfuge* . . . 322
A.4. Morphosyntax . . . 323
A.4.1. Topological fields . . . 323
A.4.2. Determiners and pronouns . . . 325
The definite article . . . 326
The indefinite article . . . 328
Prepositional dative marking . . . 329
Demonstrative determiners . . . 330
Interrogative pronouns and determiners . . . 331
Personal pronouns . . . 331
Polite form . . . 336
Reflexive and reciprocal pronouns . . . 337
Relative pronouns . . . 337
Possessive pronouns . . . 339
Indefinite pronouns and determiners . . . 340
Pronominal partitive . . . 341
Expletive *es* . . . 341
Definite article before names . . . 342
Indefinite articles before occupational titles and names of origin . . . 343
Adverbs and articles – word order and article doubling . . . 343
A.4.3. Nouns . . . 344
Inflection . . . 344
Relics of case marking . . . 348
Paraphrase of Standard German genitive case . . . 349
A.4.4. Adjectives . . . 351
Inflection . . . 351
Inflection of predicative and copredicative adjectives . . . 352
Comparison forms . . . 353
Comparative particles . . . 353
A.4.5. Prepositions . . . 354
Place-name prepositions . . . 355

A.4.6. Verbs . . . 355
The inflectional paradigm in Standard German and in Swiss German . 355
Regular verbs . . . 357
Irregular verbs . . . 360
Verb classes in Highest Alemannic dialects . . . 363
Paraphrase of the preterit and pluperfect tense . . . 363
Future tense . . . 365
Passive voice . . . 366
Periphrastic expression of conditional mood . . . 367
Periphrastic indicative forms . . . 367
Gerund . . . 368
Continuous forms . . . 368
Verb doubling . . . 369
Ellipsis of movement verbs . . . 370
Word order in modal and auxiliary verb clusters . . . 371
Word order in serial verb clusters . . . 375
A.4.7. Subordinate clauses . . . 381
Infinitival subordinate clauses . . . 381
Double complementizer in indirect questions . . . 382
Embedded main clause structures . . . 383
Long *wh*-extractions *(Satzverschränkung)* . . . 384
Relative clauses . . . 384
A.4.8. Characteristics of spoken language . . . 385
A.5. Lexical resources . . . 386
A.6. Conclusion . . . 388

B. List of digitized SDS maps **391**

Bibliography **427**

1. Introduction

1.1. Goals

This thesis finds its motivation in the inherent tensions between two rather distant fields of language science, dialectology and human language technology.

Dialectology deals with spoken, substandard language varieties and their geographic distribution. Dialectological work is essentially field work: data collection, usually through interviews of native informants, is one of its most important and most time-consuming tasks. The specific constraints of oral data collection – the informants should not be biased by the form and content of the interview – led dialectologists to pay close attention to the linguistic and extralinguistic context of the examined dialect communities and to establish links with ethnographic studies. However, confronted to the mass of data collected, dialectology has long been criticized for its lacking capacity of synthesis and generalization.

As variation is a crucial and defining characteristic of dialects, finding variation in the linguistic material is the very goal of dialectology. This goal contrasts with the status of variation in human language technology. One can distinguish two basic approaches to human language technology (Clark, Fox, and Lappin 2010, p. 1), but neither of them has traditionally granted much importance to variation.

The first approach, referred to by the denomination 'Natural Language Processing' (NLP), can be defined as the branch of computer science whose object of study is language. This field is characterized by the use of engineering methods to accomplish language-related tasks and is heavily oriented towards practical applications. As the demand for such applications usually is

limited to written data in standardized languages, questions of language variation tend to be either neglected or ingenuously circumvented.[1]

The other approach, Computational Linguistics (CL), can be viewed as the branch of linguistics that uses computers for modeling purposes. It aims at the large-scale validation of formal theories of language. For some 50 years, theoretical linguistics has been claiming the formal nature of language; however, this claim presupposes a series of generalizations and abstractions at the social level of the language community as well as at the individual level of the speaker-hearer. These abstractions have mainly cut off the questions of language variation and substandard varieties from theoretical linguistics. Even if variation has regained some interest in theoretical linguistics in the last decades, it has only marginally found its ways into CL.

The sketched distinction between NLP and CL is, at best, an epistemological shortcut. Most research groups combine characteristics of both approaches. Nevertheless, as neither approach pays much attention to variation in general, and geographic variation in particular, the points of contact between human language technology and dialectology are few.

Variation, be it geographic or not, is an essential characteristic of human language, as is acknowledged by sub-fields of linguistics like dialectology or sociolinguistics. Hence, we believe that computer models of language processing should also be able to account for variational phenomena. We further believe that the current methods of representing and interpreting dialect data cannot do justice to the huge efforts involved in their collection. Therefore, we see a lot of benefits in the integration of dialectological work in computational models. For the CL community, the use of non-standard, spoken data presents interesting challenges for modeling, and the NLP community may be interested by new applications and use cases arising from the new type of data. For dialectologists, the constraints of the formal language models require a more concise and synthetic representation of the data, and computational implementations of such models allow the data to be used by laymen in more intuitive ways than before.

This thesis is an attempt to overcome the inherent tensions between dialectology and human language technology. Concretely, we propose a system that generates words and sentences in

1 Text-to-speech and speech-to-text applications are of course excepted here. We will discuss these applications in a subsequent section, as they are crucial for our own work.

multiple dialectal varieties. We have chosen the dialects of German-speaking Switzerland as target varieties into which Standard German words and sentences are transformed. The choice of Swiss German dialects is essentially based on personal and practical considerations. Our system is based on a linguistic analysis of the source language, Standard German, and on a set of transfer rules that account for the differences between Standard German and Swiss German. These rules are triggered selectively with respect to the Swiss German dialect area chosen by the user. The proposed system is designed to benefit from the etymological closeness of the respective language varieties, and from data made available by dialectological research.

In the remainder of this chapter (Section 1.2), we introduce our object of study, the Swiss German dialects. We start by defining some essential dialectological concepts (Sections 1.2.1 and 1.2.2). Then, we present the (socio-)linguistic situation of German-speaking Switzerland. Swiss German dialects are in a three-fold opposition with ambient language varieties. First, they are in opposition with the other languages spoken in Switzerland, i.e. French, Italian, Romansh, and many immigrant languages. Second, they are in opposition with other dialects of German, spoken north and east of the Swiss territory. This second opposition can be extended to the internal differences among the Swiss German dialects. Third, they are in contrast with the standard variety of German. In Sections 1.2.3 to 1.2.6, we will turn to each of these types of opposition. Finally, we will investigate how these characteristics affect our project and its potential uses (Section 1.2.7).

1.2. Swiss German dialects

1.2.1. Dialects, varieties and languages: some definitions

In our attempt to define key concepts like 'dialect' or 'language', we start with the concept of '**language variety**'. A language variety is any way of speaking of a group of people that we wish to consider as a single entity (Chambers and Trudgill 1998, p. 5). By postulating a language variety, one merely decides to make abstraction of the individual variation occurring among the members of the group. The groups can be defined by various criteria. A **dialect** is a language variety defined by the geographical origin of its members, in contrast to other **lects** defined by

other, non-geographical criteria like social class or age. Chambers and Trudgill (1998, p. 6) note that etymologically related dialects form a **dialect continuum**. While dialects at the edges of this area may not be mutually intelligible, they are linked by a chain of mutual intelligibility across the area (adjacent dialects are always mutually intelligible). Under this conception of dialect continuum, which has not always been accepted (see p. 55), dialect denominations are mere conventions: varieties that we wish to consider as single entities for the sake of convenience.

The term 'dialect' is often used in contrast to the concept of '**standard variety**'. A standard variety is a distinct language variety that spreads over a territory defined by political and cultural rather than linguistic criteria. Usually, the dialect of the capital of a political entity gains prestige and spreads over other regions, being superposed onto the local dialects rather than substituting them. Hence, we may speak, with Herrgen (2000, p. 1514), of a two-fold opposition:

> Unter 'Dialekt' wird daher eine muttersprachlich erworbene Varietät eines Gesamtsprachsystems verstanden, die sich durch einen Teil ihrer sprachlichen Elemente einerseits horizontal abhebt von Nachbarvarietäten anderer arealer Erstreckung, die sich andererseits vertikal abhebt von überdachenden Varietäten des Gesamtsprachsystems, namentlich von der Standardsprache.[2]

On a horizontal level, dialects are in opposition with geographically adjacent dialects; on the vertical level, they are in opposition with the standard variety. However, the vertical opposition is not binary: Niebaum and Macha (2006, p. 7) cite several models that account for different intermediate varieties between the standard variety and its most distant and most distinctive counterpart, called **base dialect**.

We follow Chambers and Trudgill (1998, p. 5) with respect to the distinction between dialect and **accent**: 'accent' refers to phonetically and/or phonologically different varieties, while 'dialect' refers to grammatically and/or lexically (as well as phonologically) different varieties.

Now that we have defined the concepts of dialect, accent and language variety, a crucial question remains: What is a **language**? There is, of course, the famous aphorism that "A language is a

2 "Thus, 'dialect' is understood as a natively acquired variety of a language system, which, through some of its linguistic properties, sets itself apart horizontally from the neighboring varieties of different areal extent, and vertically from the 'roofing' varieties of the language system, specifically, the standard language."

dialect with an army and a navy."[3] This aphorism suggests an equivalence between 'language' and 'standard variety': indeed, a local variety becomes a standard variety by the means of political action.

However, we rather want to follow the approach of Chambers and Trudgill (1998): they find it useful to regard dialects as **dialects of a language**. In that sense, the term 'language' would encompass the area of a standard language and all dialects spoken therein. In fact, as a standard variety spreads out, the local dialects become increasingly dependent on it: neologisms and innovations are carried over from the standard variety to the dialects. The standard variety becomes **autonomous**, the dialects **heteronomous**. We can thus define 'language' as "a variety which is autonomous together with all those varieties which are dependent (heteronomous) upon it"(Chambers and Trudgill 1998, p. 11). This subordination of the dialects also accounts for their relatively little overt prestige and their mainly oral use.

We would like to add some remarks about the function of dialects. It may be surprising that standard varieties do not emerge spontaneously but need (at least initial) political impact. It may also surprise why dialects keep being used vigorously even in regions where autonomous standard varieties are available. While it is evident that one of the main purposes of language is to allow communication, it is less commonly admitted that language also is about identification with certain communities and isolation from others. Or as the French dialectologist Séguy (1973b, pp. 27-28) puts it:[4]

> La fonction des dialectes est double, et chacune des deux sous-fonctions est contradictoire à l'autre. La première est d'assurer la communication linguistique entre les groupes humains, la seconde est de permettre à ces groupes de se différencier.[5]

3 This citation has been published for the first time by the linguist Max Weinreich, but its origin remains unclear. It has been suggested, albeit without good evidence, that Maréchal Hubert Lyautrey originated the phrase. See http://en.wikipedia.org/wiki/A_language_is_a_dialect_with_an_army_and_navy (accessed 8.8.2008) for a discussion.

4 See Auer, Hinskens, and Kerswill (2005, p. 4) for similar positions on that issue by Schuchardt and de Saussure.

5 "The function of dialects is double, and each of these sub-functions is contradictory to the other. The first one is to ensure the linguistic communication between the groups of people, the second one is to allow these groups to differentiate."

While these two functions are contradictory, they can be accommodated thanks to the redundancy of language: it allows dialects to be different without being unintelligible.

1.2.2. Variables, variants and isoglosses

Dialectological research is essentially concerned with the study of the geographical distribution of linguistic phenomena. At the core of this research lies the distinction between the concepts of 'variable' and 'variant'.

Chambers and Trudgill (1998, p. 50) define the **linguistic variable** as "a linguistic unit with two or more variants involved in covariation with other social and/or linguistic variables". In dialectology, variables covariate with geographical location. Hence, a **dialectological variable** can be viewed as any linguistic phenomenon whose realization varies along the geographical axis. Its different realizations are called **variants**.

For example, "vowel length before *b, d, g* in an open syllable" is a variable. In region *A*, it is realized with short vowels, and in region *B*, with long vowels. Hence, this variable has two variants.

The distinction between variables and variants follows from the fact that we intend to describe different dialects. In a model that covers a single language variety, there should only be one variant per variable, which makes their conceptual distinction irrelevant.

The different variants of a variable can be delimited by a line on the map. Such a line is called **isogloss**. In our example, there is only one isogloss, namely the one that separates region *A* from region *B*. If the isoglosses of different variables overlap in the same area, one speaks of an **isogloss bundle**. However, isoglosses often will not strictly overlap, but will criss-cross the area of study seemingly randomly. Therefore, the utility of this concept has been questioned (Chambers and Trudgill 1998, ch. 7).

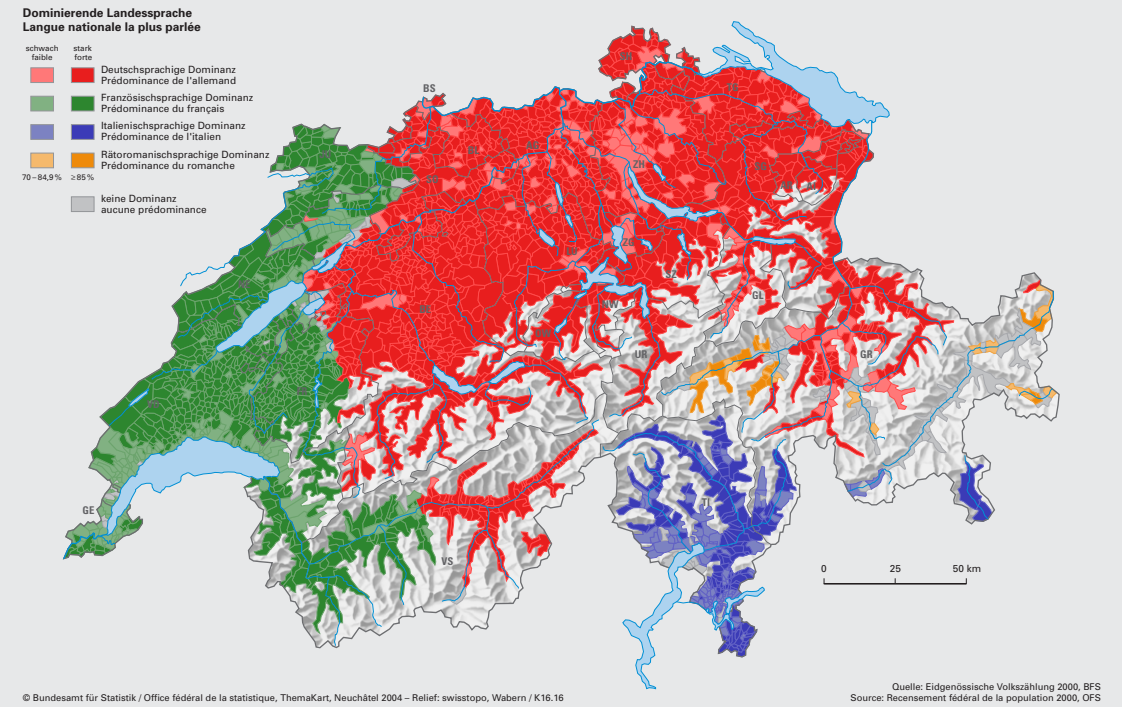


Figure 1.1.: Predominant language per municipality, as of 2000.

1.2.3. Languages in Switzerland

In spite of its small size, Switzerland exhibits a rather complex linguistic situation. The Swiss constitution defines German, French, Italian and Romansh as official languages. Their geographical distribution is shown in Figure 1.1,[6] and described in detail by Bickel and Schläpfer (2000) and Lüdi et al. (1997).

The German language area is the largest (63.7% of the Swiss population) and roughly covers the two Eastern thirds of the Swiss territory north of the Alps.[7] It is characterized by the opposition between Standard German and the local dialects, and by the diversity of the latter. We will come back to these phenomena in subsequent sections.

6 Map source: http://www.bfs.admin.ch/bfs/portal/de/index/regionen/thematische_karten/maps/bevoelkerung/sprachen_religionen.parsys.0002.PhotogalleryDownloadFile2.tmp/k16.16s.pdf, accessed 30.9.2008.

7 Statistical language data stem from the 2000 population census http://www.bfs.admin.ch, accessed 30.9.2008.

The French language area is located in the Western third of Switzerland and accounts for 20.4% of the Swiss population. The German-French language border does not follow physical or political boundaries: the three cantons of Berne, Fribourg and Valais are officially bilingual. Moreover, there are bilingual cities like Fribourg and Biel/Bienne. The French language area is quite homogeneous; under the pressure of neighboring France's language policies, local dialects have died out in most regions of French-speaking Switzerland. Relics of dialects remain in some Valais valleys. However, regionalisms subsist on the lexical and phonetic level, and marginally also on the grammatical level (see for example Knecht 2000).

The southern canton of Ticino is officially italophone. Italian is the mother tongue of 6.5% of the Swiss population, which includes numerous immigrants from Italy scattered all over Switzerland. Ticino is located in the traditional area of Lombard dialects. These dialects are still used quite commonly in familiar contexts by the local population, but migration and media consumption are pushing forward Standard Italian. Still, dialect use remains more popular in Ticino than in the bordering regions of Italy (Lurati 2000). The German-Italian language border follows the main ridge of the Alps.

The canton of Grisons features a complicated linguistic setting. There are three Italian-speaking valleys in the South of the canton that exhibit the same characteristics as Ticino. In the north of the canton, including its capital Chur, German and Swiss German dialects are exclusively used. The remaining areas of the canton are traditional Romansh language areas, but they have been germanized to various degrees. In all of these areas, the use of German is so widespread that there are no more monolingual Romansh speakers today. The linguistic fragmentation of Romansh parallels the topographic fragmentation of the alpine valleys of Grisons: there are five different idioms (i.e. regional dialects with proper orthographic norms), but local dialects often deviate from these regional idioms. Only recently, standardization efforts have led to a common standard language, *Rumantsch Grischun*.[8]

Besides the four official languages, other languages can be heard frequently in Switzerland, as

8 See Haiman and Benincà (1992) and Liver (1999) for a detailed presentation of Romansh. Our own recent computational linguistic work on lexicon induction for *Rumantsch Grischun* is described in Scherrer and Cartoni (2012).

migrants account for 21.1% of the permanent Swiss population.[9] Sign languages for the deaf, with different varieties according to the oral language areas, also exist. English is taught as a foreign language in all regions of Switzerland and is increasingly being used in multinational enterprises and institutions.

1.2.4. German dialects

The dialect landscape found in the German-speaking countries is difficult to describe because of the absence of clear and objective boundaries. Nevertheless, a tentative classification (inspired by Niebaum and Macha 2006, 217ff.) based on some pervasive linguistic features can be drawn.

The most important criterion for this dialect classification is the **Second Germanic consonant shift**, a series of consonant alterations that took place between the third and the fifth century. While the consonants remained unchanged in the Low German areas (Northern Germany, Netherlands, Scandinavia, including the homelands of the Anglo-Saxons), the alterations applied to varying degrees in the Southern part of the German territory.

The key phenomena of the Second Germanic consonant shift may be illustrated by means of English and Standard German examples: English reflects the Proto-Germanic (unshifted) consonants, while Standard German is, historically speaking, a Southern variety that applied most (but not all) consonant alterations (Lehmann 1992, pp. 125-129). The main phonemes involved in this shift are the plosives [p], [t] and [k]. In word-medial (1a) and word-final (1b) contexts, these plosives have become fricative consonants. But in word-initial (2a) and geminated word-medial (2b) contexts, [p] and [t] changed into affricates [pf] and [ts].[10]

(1) a. ho**p**e — ho**ff**en
wa**t**er — Wa**ss**er
ca**k**e — Ku**ch**en
b. u**p** — au**f**
i**t** — e**s**

9 Data from 2007, see http://www.bfs.admin.ch.

10 Standard German did not change [k] in this context; examples are given in parentheses.

boo**k** — Bu**ch**

(2) a. **p**an — **Pf**anne
tongue — **Z**unge
(**c**oal — **K**ohle)
b. ho**p** — hü**pf**en
si**t** — si**tz**en
(li**ck** — le**ck**en)

These examples might suggest a binary situation, in which some language varieties (like Standard German) fully applied the consonant shift and some language varieties (like Standard English and Standard Dutch) did not. Yet the situation is overly simplified by the development of standard varieties. In fact, Germanic dialects applied the consonant shift very unevenly. Hence, the degree of application of the Second Germanic consonant shift allows us to divide the German dialect landscape in three main dialect areas (see Figure 1.2): **Lower German** in the North (not affected by the consonant shift), **Middle German** in the center (partially affected by the consonant shift), and **Upper German** in the South (fully affected by the consonant shift). The Second Germanic consonant shift is also called **High German consonant shift** because it occurred in the **High German** dialect area; 'High German' is a hypernym for the Middle and Upper German dialects.[11]

The three main dialect varieties are each divided into a Western and an Eastern group, yielding six dialect groups:

West Lower German is spoken in Lower Saxony, the Hamburg region and parts of the Netherlands.

East Lower German is spoken in the Northern half of the former German Democratic Republic and was also in use in the former German territories in Pomerania and East Prussia. Linguistically, it is distinguished from the western variant by its verb plural endings

11 As mentioned above, the standard variety of German emerged from Southern, High German varieties, which did apply the consonant shift. The High German origin of Standard German also accounts for the fact that in Swiss language use, the term *Hochdeutsch* now unambiguously refers to Standard German, although the Swiss dialects are High German varieties as well.

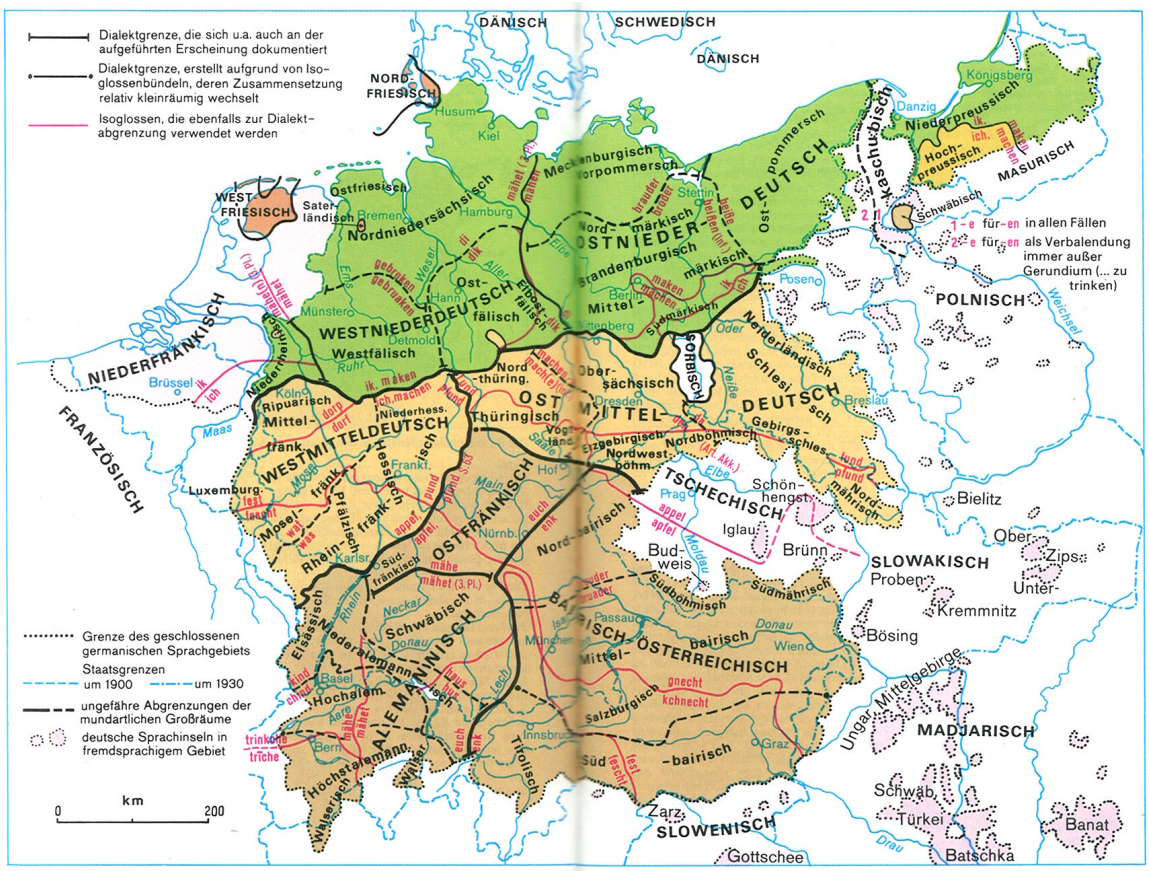


Figure 1.2.: Main divisions of the German dialect landscape as of 1900. Lower German dialects are shown in green, Middle German in orange, Upper German in brown. Map source: König 2007, pp. 230-231.

containing the consonant [t].

West Middle German is used along the Rhine between Mannheim and Cologne, with incursions into Netherlands, Belgium and Luxembourg. It is characterized by an incomplete application of the High German consonant shift: initial [p] is not shifted to [pf], final [t] is not shifted to [s].

East Middle German is spoken in Thuringia and Saxony and was spoken in the former German territories of Silesia. In these dialects, initial [p] is shifted to [f] (for example, *Ferd* instead of *Pferd* 'horse').

West Upper German dialects are found in Baden-Württemberg, Alsace, Switzerland and the westernmost parts of Bavaria and Austria. One of their distinctive features is the use of [l]+*V* as a diminutive suffix (in contrast to suffixes based on [k] or [x] in the North).

East Upper German dialects are also called **Austro-Bavarian**. Accordingly, this area covers most of Bavaria, Austria, and the Italian region of South Tyrol.[12] In these dialects, the diminutive form also contains [l], but in final position. Moreover, there are particular pronoun forms in the second person plural.

All dialect groups can be divided further. We will focus on a more detailed presentation of the West Upper German dialects here – these are the dialects in which we are interested in our work. Three of the five subgroups of West Upper German are spoken in Germany only: Swabian, East Franconian and South Franconian. **Low Alemannic** is used in the Western half of Baden-Württemberg (i.e. in Baden) as well as in Alsace. **High Alemannic** is used in German-speaking Switzerland and its neighboring regions of South Alsace and South Baden. Vorarlberg (part of Austria) and Liechtenstein contain areas of both varieties. Alemannic in general is characterized by a more widespread use of monophthongs instead of diphthongs more northwards. High Alemannic is the only dialect group that shifts initial [k] to [kx] or [x], as the ultimate corollary of the High German consonant shift.

12 A single Swiss village, Samnaun, is also part of the Austro-Bavarian dialect area; see below.

1.2.5. The dialectological division of Swiss German dialects

Lötscher (1983, pp. 137-184) distinguishes (Plain) High Alemannic, spoken in the lower, Northern parts of German-speaking Switzerland, from **Highest Alemannic** dialects, spoken in the alpine regions in the South. The alpine dialects are more conservative and contain many relics that have been abandoned in the plateau dialects. For example, alpine dialects show more monophthongs (*nüü* instead of *nöi* 'new', *schnyye* instead of *schneie* 'to snow'), have maintained predicative adjective inflection (*er isch alte* instead of *er isch alt* 'he is old') and, in some regions, three forms (instead of two or one) in verbal plural inflection. Other differences are found in the lexicon.

The second division is an East-West opposition, with Bern German as the prototypical Western dialect, and Zürich German as the prototypical Eastern dialect.[13] On the phonetic level, Western dialects are characterized by more open vowels (*Bètt* instead of *Bett* 'bed', *hùbsch* instead of *hübsch* 'beautiful').[14] On the morphological level, Western dialects have two verbal plural forms, while Eastern dialects only have one. In parallel, Western dialects use the second person plural as a polite form (as in French), while Eastern dialects use the third person plural form (as in Standard German). Some lexical differences are also quite prominent (W *Matte* vs. O *Wise* 'meadow', W *Z(w)ibele* vs. O *Böl(l)e* 'onion'). Many of these isoglosses roughly follow the course of the Aare and Reuss rivers and the Napf and Brünig mountains, hence the names of **Aare-Reuss Barrier** (Russ 1990, p. 367) or **Brünig-Napf-Reuss Line** (Haas 2000, p. 83). However, not all isoglosses coincide, such that the Aargau-Luzern area can be viewed as a transition zone. This main border, which roughly coincides with the eastern border of the former State of Berne, has been shown to live on as religious and cultural frontier (Hotzenköcherle 1984).

The combination of the North-South division and the East-West division results in a quartering of the Swiss German dialect landscape. Several minor divisions complement this quartering. The Basel region shares some common characteristics with other Low Alemannic dialects like Alsatian, e.g. word-initial lenition (*Daag* instead of *Taag* 'day') or word-initial [kʰ] instead of [kx].

13 Throughout this thesis, we adopt the convention to use the German designation of the cantons, even if English names exist, and even if the French name is more common: 'Zürich' instead of 'Zurich', 'Luzern' instead of 'Lucerne', 'Freiburg' instead of 'Fribourg'. See also A.1.1.

14 The use of diacritics in dialect spelling is discussed in Section 3.2.

North-Eastern dialects show other particularities, e.g. extended monophthongization (*Bòmm* instead of *Baum* 'tree').

The Graubünden canton shows an interesting setting. This originally Romansh-speaking area has been germanized by three different population groups, leading to distinct dialect areas. The Walser migrations started in the 9th century and led groups of people from the Bern region into the Wallis, further on to Italy, and then back north to Graubünden and Western Austria. As a result, the Walser settlements in Graubünden share many features with the Wallis dialects. A second immigration wave followed the Rhine valley upstream, accelerated by the influx of reconstruction workers after a devastating fire of the city of Chur in the 15th century. This dialect, called *Bündnerrheintalisch*, in combination with Romansh substratum influence, has become the prototypical Graubünden dialect. The Samnaun valley has been germanized by Bavarian dialect speakers moving up the Inn valley from Austria. As the only non-Alemannic dialect area in German-speaking Switzerland, it is not covered by Swiss German dialectological research, and will not be covered in our work either.

More fine-grained dialect divisions are usually based on political boundaries (cantons). These frontiers have become less important in the last decades, and as a consequence, some dialect differences are leveling out. Christen (1998a) observes that dialect differences are more easily maintained in grammar than in the lexicon. The lexicon tends towards a uniformization among the dialects on the one hand, and between the dialects and Standard German on the other hand. This means that dialect-specific words become increasingly rare. This goes in parallel with the declining importance of agriculture and its highly specialized, dialect-specific vocabulary (object loss). Neologisms are frequently taken over from Standard German, in most cases with phonological and morphological adaptation. In contrast, grammar (phonology, morphology, syntax) as well as functional and high-frequency words largely resist to uniformization tendencies, as they do generally not hinder successful communication, but rather provide the speakers an element of identification with regional traditions.

1.2.6. Diglossia

We have already mentioned the complex relations between the standard variety and the local dialects, and we evoked the possibility of a vertical continuum between them in Section 1.2.1. Such continua exist in parts of the High German dialect area, but not in German-speaking Switzerland, where the Alemannic dialects are clearly distinct from the standard variety.

The socio-linguistic configuration of German-speaking Switzerland is commonly presented as a model case of **diglossia**. This term has been introduced by Ferguson (1959) to describe an environment where two language varieties are used in a complementary way in functionally different contexts. Usually, the two varieties are used according to the social context in which the interaction takes place. Ferguson distinguishes thus a 'high' variety for formal contexts and a 'low' variety for informal situations with familiar interlocutors. The high variety usually corresponds to a standard variety, the low one to the local dialects. However, this distinction does not depict the Swiss German reality adequately: the choice of a variety is not influenced by the social affiliation of the interlocutor or the degree of formality, but exclusively by the medium of language transmission. Dialects are used in all oral communication situations, in familiar contexts as well as in interaction with political, religious or other authorities. Standard German is used for any written texts. This type of diglossia, depending solely on the communication medium, is called **medial diglossia**.

This configuration implies that every speaker who is regularly confronted to the two medial contexts is competent in both language varieties. Indeed, Swiss Germans are able to use dialects in spoken contexts and Standard German in written contexts. They are also able to make exceptions to this rule for practical reasons: prepared speeches may be held in Standard German because the written draft has been composed in Standard German; important television programs may use Standard German to extend their coverage to non-dialect speakers; discussions in the Swiss Federal Parliament avoid dialect use in order not to exclude Swiss French and Swiss Italian members; immigrants and tourists from Germany may prefer using Standard German. Despite these exceptions, the rule of a medial diglossia can still be maintained.

However, recent developments somewhat challenge this conception, as the local dialects are increasingly being used in their written form. This tendency has started with dialect literature

in the 1980s and has thrived recently in informal uses of electronic communication means. Scharloth (2004) estimates that 58% of E-mails and 75% of SMS messages written by Swiss Germans are in dialect. The diglossia in Switzerland is thus neither defined by the purely medial criterion nor by the traditional distinction of a high and a low variety, but by a complex interaction of social, medial and practical factors.

At the same time, many Swiss Germans refute the conception of heteronomous dialects subordinated to Standard German. Having grown up with dialects, they perceive Standard German as a foreign language that has to be studied at school and whose perfect mastery is difficult to attain. In support of this view, linguists point out that Swiss dialects in fact are full-fledged languages with different registers and a proper prestige. In the end, this debate not only concerns the question of the linguistic autonomy of the Swiss German dialects, but also shows that the domains of dialect use have become so large that the concept of **collective bilingualism** now seems more appropriate than the concept of diglossia (Hägi and Scharloth 2005). At least, this debate confirms that the limit between dialects and Standard German is perceived as a very clear one and not at all as a continuum.

To sum up, we can define Swiss German dialects according to two criteria, a linguistic one and a socio-linguistic one. Swiss German dialects are High Alemannic or Highest Alemannic dialects which are in a relation of medial diglossia or collective bilingualism with Standard German.

1.2.7. Impacts on cross-lingual, multi-dialectal sentence generation

In the previous sections, we have related the Swiss German dialects to their neighboring language varieties. Here, we would like to discuss the impact of these relations on our project of a multi-dialectal computational model of Swiss German. The discussion will focus on two aspects. On a practical level, we will explore what groups of users might be interested in such an application, and what kind of tasks they might want to perform with it. On a theoretical level, we will ask what phenomena of linguistic competence should be reflected by our model.

We view our model as **cross-lingual** in the sense that it describes the relation between two vertically opposed language varieties, Standard German and Swiss German. We qualify our

approach as **multi-dialectal** in the sense that our model is intended to cover the whole range of Swiss German dialectal variation.

Practically, the cross-lingual model amounts to a **translation** system. Potentially, such a system may be conceived for three translation settings: (i) from Standard German into a Swiss German dialect; (ii) from a Swiss German dialect into Standard German;[15] (iii) from a Swiss German dialect into another Swiss German dialect. In the following paragraphs, we discuss the relevance of each of these settings from a theoretical and a practical point of view.

The great majority of adult Swiss Germans are regularly confronted to spoken contexts (where dialect use is required) as well as to written contexts (where Standard German is preferred). They can thus be considered bilingual. Moreover, they are passively pluri-dialectal – they understand other Alemannic dialects without problems, although they generally do not use them actively. From a Swiss German's point of view, an automatic translation tool may not seem useful, neither to translate into or from another Swiss German dialect, nor to translate from or into Standard German. Nevertheless, it is interesting to study the various language competences of the "native Swiss German", exemplified in the following paragraphs, in order to reproduce them in a cross-lingual model.

If a speaker of an Eastern Swiss German dialect is confronted to Bernese German, she will soon notice that in some words where she would expect the sound [l], she will in fact hear an [ʊ]-like vowel or an [w]-like semi-vowel. More exposure to Bernese will allow her to detect the exact phonetic contexts in which this phenomenon occurs. She will generalize this information and memorize it, so she will be prepared to this peculiarity of Bernese the next time she hears it. Moreover, she will be able to roughly tell the origin of an unknown speaker who uses this feature of [l]-vocalization. Gradually, our speaker builds up an implicit knowledge of the linguistic relationship between the dialects and of their respective geographical extensions.

Similar phenomena occur when Swiss children, who usually grow up with dialects, learn Standard German at school. There, they may be taught that dialectal [uː] is pronounced [au] or [aʊ] in Standard German, and they will memorize this rule and apply it – sometimes even in situa-

15 In a procedural approach like the one we pursue, the direction of translation is important. Cases (i) and (ii) thus need to be represented by different systems.

tions where they shouldn't. Likewise, an adult Swiss German might read the Standard German words *Einlagensicherungen* 'deposit insurances' or *Symmetriebrechung* 'symmetry breaking' in a newspaper[16], and naturally "convert" them into his local dialect (e.g., *Iilagesicherige* and *Symetriibrächig*) when talking to his colleagues, even if he has never heard these words before. The successful transfer of words and utterances from one variety to another, as well as the occasional overgeneralization errors, suggest that the two varieties are strongly linked in the speaker's brain.

To sum up, Swiss Germans have an extensive knowledge about the differences between the language varieties they are confronted with. This knowledge may be implicit and passive, as in the case of interdialectal differences, or active and explicit (i.e. learnt at school), as in the case of Standard German. We argue that understanding utterances of another dialect, or producing utterances in another variety, are in fact translation problems. The modeling of these problems is our foremost goal, independently of the precise source and target language varieties chosen. Of course, translating from Bernese to Eastern Swiss German is easier than, say, translating from English to French: word pronunciations differ only slightly, and the general syntactic structures are similar too. If one ignores the thorny issue of interdialectal variation, translating between Swiss German dialects could thus be considered a very simple problem, which might be tackled by adapting existing machine translation techniques to closely related and geographically confined language varieties. Such a model might help to gain a better understanding of the Swiss Germans' linguistic competence.

Native speakers of French and Italian roughly make up a quarter of Switzerland's population. These people learn Standard German at school (in its written and spoken form), but usually have little knowledge of Swiss German dialects. Therefore, they are practically excluded from discussions where a Swiss German majority imposes dialect use. Moreover, the lack of any standardization of Swiss German dialects may unsettle native French speakers accustomed to strong language planning authorities. Such users may be interested in a system that translates from French or Italian directly into a Swiss German dialect. But it may be more useful to build on the existing Standard German competences of these users and to focus on the differences between Standard German and Swiss German. There are, in our view, two main goals of such a system. First, it visualizes the contrast between a Standard German sentence and its multiple

16 NZZ Online, 7.10.2008

translations in Swiss dialects and thereby raises awareness about the variety and variability of the Swiss German dialects. Second, it can be viewed as an interactive educative tool for people wanting to learn dialects by building on their knowledge of Standard German.

Such a multi-dialectal approach is very much in line with recent teaching material for Swiss German, for instance in the language course *Chunsch druus?* (Müller et al. 2009), aimed at the Swiss French and Swiss Italian population as well as at migrants. In contrast to older teaching material, *Chunsch druus?* uses material from four prototypical dialects (Basel, Bern, Zurich and St. Gallen dialects) and discusses the systematic differences between them. In this way, the authors hope to raise the learners' awareness for dialectal variation. They argue that knowledge of the four selected dialects allows to infer the features of many other Plateau dialects and also helps in understanding the more "exotic" alpine dialects.[17]

A lot of German citizens come to Switzerland as tourists or immigrants. Most of them hail from regions where dialect use is marginal. Therefore, they use a standard-like variety of German. Swiss Germans will understand them easily, but they may not be able to understand the Swiss Germans, as they are not accustomed to dialects.[18] Of course, Swiss Germans can switch to Standard German in order not to disrupt the communication, but they often feel uncomfortable doing so. At the same time, Germans may feel uncomfortable imposing an unnatural language variety to their interlocutors. In this situation, a system that translates between Standard German and Swiss German dialects may help Germans to improve their knowledge of the Swiss German dialects. From our point of view, dialect knowledge facilitates successful social integration.

Let us sum up five combinations of source and target varieties for cross-lingual multi-dialectal modeling.

1. **Translating from one Swiss German dialect to another** allows us to model the passive pluri-dialectal knowledge of the native Swiss Germans, but does not seem to be of much practical use.

17 See http://www.schulverlag.ch/platform/content/element/10353/9278_FAQ.pdf, accessed 18.5.2012.

18 For instance, television interviews held in Swiss German dialect are subtitled or dubbed when broadcasted by Germany's television channels.

2. A system that **translates from a Swiss German dialect to Standard German** would be a practical interactive tool for Germans and non-dialect speaking Swiss citizens confronted to Swiss German dialects. Such a system could also model the acquisition of Standard German by dialect speakers.

3. **Translating from Standard German to a Swiss German dialect** may have educational uses for Germans and non-dialect-speaking Swiss citizens, but is less aimed at an interactive use. Still, it would be modeled after some particularities of the pluri-dialectal language competence of the native Swiss Germans.

4. **Translation from a Swiss German dialect to a third language** like French or English might appeal to third language speakers not fluent in Standard German. Two approaches, a direct one and an indirect one, are conceivable. The first approach consists of developing a distinct transfer module that transforms the Swiss German word order directly into French or English word order, and replaces Swiss German words by French or English words. This approach can be justified for French on linguistic grounds: Swiss German dialects have borrowed much more French words than Standard German did, and some syntactic features of Swiss German are parallel to French ones. However, creating new transfer modules is labor-intensive, and an indirect approach may be more suitable. Such an approach would decompose the translation task into two subtasks, using Standard German as a pivot language. In this case, we could use the model of case two to translate dialects into Standard German, and apply a state-of-the-art machine translation system to translate from Standard German into French or English. We believe that despite the greater complexity, the latter approach would yield better results and still need less development efforts.

5. The **translation from a third language to a Swiss German dialect** is analogous to the fourth case. As above, a direct and an indirect approach can be conceived. Such a system might have educative and practical (i.e. in interaction) uses for speakers not fluent in Standard German.

This thesis will focus on the third setting. We develop a procedural translation system that considers Standard German as a reference variety from which the Swiss German dialects are derived.

It makes sense that this reference is a standardized and well described language variety.

According to the medial diglossia, it would be most intuitive to create a multi-modal text generation system that involves written Standard German and spoken Swiss German. While this represents a long-term goal of our research, we will focus here on written representations of Swiss German dialects.

In conclusion, our project is intended to achieve several objectives. It is, first of all, a translation application. That is, any person wishing to translate Standard German words or sentences into a Swiss German dialect will be able to use it. Related tasks like treebank transduction (Section 6.6), dialect identification (Section 6.5) and dialectometrical analyses (Chapter 7) are carried out thanks to the techniques and data made available for the translation application. Our model also has an educative value, in the sense that it informs non-dialect speaking people about the Swiss German dialects, their linguistic features and their geographic distributions. Finally, it is designed to model some of the specific cognitive tasks that Swiss Germans perform in their rich linguistic environment.

1.2.8. Some words about language change and borrowing

It is safe to assume that the differences between Standard German and Swiss German are essentially due to language change. Three scenarios may be distinguished, all of which can be found in the linguistic material:

1. Standard German has innovated, while Swiss German has maintained an older state.

2. Swiss German has innovated, while Standard German has maintained an older state.

3. Both Swiss German and Standard German have innovated, but in a different way.

Differences among the different Swiss German dialects may be explained in the same way: Some Swiss German dialects have innovated, while others have maintained an older state. The Standard German solution may be the one used in the innovative Swiss German dialects, the one used in the conservative dialects, or a different one altogether.

The present-day dialect landscape is the result of multiple successive steps of language change of the kinds described above, which have occurred in different areas. This matter of fact has troubling consequences on the modelling of text translation from Standard German to Swiss German. Imagine that a common variant A yields innovation B in Standard German, but is temporarily maintained in Swiss German and later on replaced by innovation C. In this case, we need to posit a translation rule $B \to C$, without having access to variant A, which had given rise to both B and C. One thus has to keep in mind that the translation rules do not necessarily reflect a historical reality. A translation rule $X \to Y$ may correspond to either one of the historical evolutions $X \rightsquigarrow Y$, $Y \rightsquigarrow X$, or $(Z \rightsquigarrow X \wedge Z \rightsquigarrow Y)$.

This picture is complicated by loanwords and grammatical borrowings. Because of the cultural and linguistic proximity, new words are constantly introduced into Swiss German by taking the Standard German model and adapting it phonetically. In some cases, a Standard German word is re-borrowed into Swiss German, although a variant of the same lexeme already exists. For example, the Standard German word *Kunst* 'art' has evolved into the Swiss German variants *Chunscht, Chuuscht* or *Chouscht* by regular language change. These variants have shifted meaning to 'tiled stove', and are becoming obsolete with the loss of the object.[19] To convey the meaning of 'art', Swiss German speakers have re-borrowed the same lexeme from Standard German and partially adapted phonetically, yielding *Kunscht* (Haas 2000, p. 99). When translating, we do not have access to the history of each word. Therefore, it is difficult to model such haphazard phenomena reliably.

1.3. Outline

In this introductory chapter, we have presented the goals and challenges of our research. Furthermore, we have discussed our object of study, the Swiss German dialects, and their context.

19 The *Idiotikon* (Staub et al. 1881-, vol. 3, col. 368) explains this semantic shift as follows: "Die 'Kunst' bestand wahrsch[einlich] darin, dass das Herdfeuer nicht nur besser als im frühern Herd [zusammen]gehalten, sondern auch zur Erwärmung der Stube nutzbar gemacht wurde; daher die anfängliche Bezeichnung 'Holzersparungskunst'."
("Probably, the 'art' consisted not only in holding together the fire better than in the earlier hearth, but also in making use of it to heat the living room; hence the initial name 'art of saving wood'.")

Thereby, we have motivated our project to generate Swiss German text of multiple dialects on the basis of Standard German text. In the following chapters, we will examine different aspects of this work.

Chapter 2 presents the history and epistemology of the two research fields that our work relies on: dialectology and human language technologies. After discussing some central pieces of closely related research in detail, we gradually extend our focus to review related work in different adjacent fields.

In Chapter 3, the two linguistic atlases of Swiss German dialects – the main data sources of our work – are presented. Their content, along with other characteristic phenomena of Swiss German, has been compiled in the form of a comparative, multi-dialectal grammar. For reasons of space, this grammar is located in Appendix A.

The key idea of our work is to use maps from linguistic atlases to parametrize the translation procedure according to the desired target dialect. Chapter 4 explains how the maps are processed in order to be suitable for this task. It starts by an introduction to Geographical Information Systems, the software used for this map processing. We then discuss the most adequate representation type of dialectal phenomena for a translation system, and present the processing steps that lead to this representation.

The translation procedure consists of hand-written rules that transform Standard German data into Swiss German data. The architecture of this translation procedure is presented in detail in Chapter 5. First, we deal with the transformation rules required by word-by-word translation. Then, we show how this approach can be extended with rules that take into account the syntactic structures of the sentences. Throughout this chapter, we aim to motivate the decisions about particular technical solutions with linguistically relevant examples. Moreover, we describe how the maps are connected with the transformation rules.

Chapter 6 evaluates the performance of the translation system on different tasks. In a nutshell, we show how well Standard German text can be translated into Swiss German, and how well the origin of a Swiss German dialect text can be predicted on the basis of the maps attached to the transformation rules.

Chapter 7 is not strictly related to the generation of dialect text. Instead, its aim is to apply existing analysis and visualization techniques from dialectometry to Swiss German data (i.e., the maps extracted from the linguistic atlases, as described in Chapters 3 and 4). The results shed new light on the Swiss German dialect landscape from different perspectives.

We conclude this thesis in Chapter 8.

2. Related Work

2.1. Introduction

As explained in the Section 1.1, our work combines methodological approaches from dialectology and from human language technology. Hence, we start this chapter with historical overviews of these two disciplines. In Section 2.2, we present the historical development of dialectological research, and more particularly, of dialect geography. In Section 2.3, the major historical breakthroughs and applications in the field of human language technology are presented and related to the NLP-CL dichotomy sketched out in Section 1.1 above.

These two historical overviews set the stage for more thorough discussions of relevant recent research. In Section 2.4, we select five papers that are crucial to our work. They all shed light on different aspects of multi-dialectal text generation.

The remainder of this chapter is divided according to different methodological approaches. Section 2.5 reviews methods of string similarity used in machine translation and lexicon induction, assuming that such methods work best between closely related languages or dialectal varieties. Section 2.6 presents the most important dialectometrical studies. We also discuss other work that involves computational processing and/or geographical analysis of dialect data (Section 2.7). A short section (2.8) is dedicated to the computational analysis and generation of spoken dialect. Section 2.9 presents computational approaches to historical data, which have similar properties to dialectal data. Finally, Section 2.10 reviews NLP tools for Standard German, on which we rely for various aspects of our dialect translation model.

Our work draws on previous research in two different and complementary ways. On the one

hand, we use, develop and extend methods that have been proposed to computationally describe dialects, their variation patterns and their relation with a standard variety. This research stems from the fields of theoretical linguistics, computational dialectology, and natural language processing. We review this work for its methodological value in this chapter. On the other hand, we rely on previous studies of the dialectological and socio-linguistic particularities of language use in German-speaking Switzerland. Our interest lies in linguistic atlases, dialect grammars, case studies on specific linguistic phenomena, and dialect lexicons. We use this work as data sources rather than as methodological inspiration. Accordingly, this research is described in Chapter 3 and in Appendix A.

2.2. A short history of dialectology

We have chosen to present the principal areas of research and the main tenets of dialectology in the form of a historical synopsis. According to the subject of this thesis, the synopsis will focus on German dialectology, but also take into account ground-breaking work from other parts of Europe.

In our view, the history of dialectology can be divided into three periods. The first period, ranging from the Middle Ages until 1876, is characterized by increasing awareness and interest in dialects and by pioneering work that set the stage by defining the main areas of dialectological research. The second period is a period of big enterprises. Starting with Wenker's work in 1876, linguistic atlases and dialect lexicons are set up all over Europe with continuously refined research methodologies. In the third period, beginning in the 1950s, new linguistic theories become increasingly influent in dialectology; in parallel, the computer enables new types of analyses of dialectal data.

In this general overview, we will not mention exact references of the cited work. Instead, the reader may refer to the historical summaries in Herrgen (2000); Jones (2000); Goebl and Schiltz (2000); Niebaum and Macha (2006). For our presentation, we also borrowed some ideas from Kristol (2005).

2.2.1. The constitution of the discipline

The awareness of regionally differentiated speech practices exists since the beginning of history, as the discussion of the opposition between *shibboleth* and *sibboleth* in the Old Testament attests.[1] In ancient Greece, Alexander the Great manages the linguistic diversity of the conquered regions and pushes standardization of the Greek dialects. Descriptions of dialect differences are also a popular topic in German medieval literature. However, they remain casual and impressionistic (Herrgen 2000; Jones 2000). The development of dialectology as a scientific discipline only starts with the gradual establishment of a written standard variety of German in the 17th and 18th centuries.

The spreading of Standard German was paralleled by the introduction of compulsory education and a growing literacy rate, which in turn raised awareness for language-related questions in general and for the opposition between local dialects and the standard variety in particular. At the same time, concerns about an imminent extinction of dialects grew. In 1697, Leibniz suggested the collection of dialect vocabularies with the objective of enhancing and vitalizing the Standard German vocabulary, and with the hope that dialect expressions might yield new insights into the history of the German language.

The first local vocabulary collections were compiled in the 18th century. In the German-speaking area, the term '*Idiotikon*' has since then become customary for such collections. In Switzerland, Franz Josef Stalder published the two volumes of the *Schweizerisches Idiotikon* in 1806 and 1812, which served as an impetus for later work. The lexicon of Bavarian dialects, edited by Johann Andreas Schmeller in 1827-1837, innovated in two major domains: Schmeller used a phonetic notation and devised a special etymological-alphabetical ordering of the lemmas.

Schmeller complemented his lexicographic work with a grammar of the Bavarian dialects (1821), inspired by the German grammar by Jacob Grimm (1822), which itself represented a ground-breaking work of historical-comparative grammar. Schmeller's grammar was the first to consider dialects as an object of comparative linguistics; it also contained the first cartographic representation of a dialect area. As Herrgen (2000, p. 1517) notes, Schmeller's work combines the three main research areas of dialectology: dialect lexicography, dialect grammar

1 Judges XII, 6. The use of this term in current dialectology is discussed in detail in Section 6.5.1.

(*Dialektgrammatikographie* in Herrgen 2000) and dialect geography. Furthermore, his sociolinguistic characterizations of dialect phenomena confer his work a modern appeal. Without any doubt, Schmeller can be considered as one of the precursors of German dialectology.

2.2.2. Big enterprises

Beginning with the 1870s, several independent factors led to an increasing interest in dialect research. The unification of Central European nations facilitated such enterprises with nationally unique institutions like postal service and public schools. In turn, politicians considered dialect research a useful means to help substantiate nationalist claims. At the same time, the industrial revolution brought about efficient transport systems and massive migration. The increased mobility and the resulting social changes led to a feeling of urgency: the dialects were endangered and had to be studied before their disappearance. Finally, the Neogrammarian school was about to establish (historical) linguistics as an independent scientific discipline. The study of dialects was seen as an opportunity to refine the comparative method, to discover previously unknown etymological links and to confirm rules of language change on a more local ground.

In the domain of **dialect grammar**, 1876 marked a turning point with the work of Jost Winteler. His grammar *Die Kerenzer Mundart des Kantons Glarus* is based on Neogrammarian methodology and describes the phonetic and inflectional systems of the dialect of an Eastern Swiss German village. His work anticipates several tenets of 20th century linguistics like structural analyses and the reduction of the research object to the competence of an individual speaker. Many dialect grammars followed Winteler's methodology.

The first big enterprise in the field of **dialect lexicography** starts in 1881 with the publication of the first issue of the *Wörterbuch der schweizerdeutschen Sprache*.[2] Following the prototyping work of Stalder, from which it borrowed its short name *Schweizerisches Idiotikon*, this project aims at a comprehensive collection of the Swiss dialect vocabulary. Its methodology is inspired by Schmeller's work and by Hermann Paul's claims for a historical-comparative lexicography.

2 We find that the *Wörterbuch der schweizerdeutschen Sprache* merits the appellation 'big enterprise' since its last issue is scheduled for publication in 2020.

Dialect geography will occupy a major role in our work and will be discussed in more detail here. In this field, the beginning of the second period is usually set at 1876, when Georg Wenker started his first dialect survey in Germany. His original motivation was to prove the romanticist claim of clear-cut dialect boundaries reflecting the former Germanic clan boundaries ("*Stammeshypothese*"). However, the results quickly led him to abandon this claim (Niebaum and Macha 2006, pp. 59, 81). For his survey, Wenker sent a list of sentences written in Standard German to teachers in northern Germany and asked them to translate the sentences into their local dialect. In the following ten years, he made successive mailings in other regions which eventually allowed him to cover the entire territory of the German Empire. Wenker received 45 000 responses, each questionnaire containing about 40 sentences. This mass of data could not be analyzed adequately with the available technical and financial resources; an immediate large-scale publication of the results in maps was not possible. Wenker ended up drawing two sets of maps by hand, which were deposited in Marburg and Berlin in 1881. Nevertheless, Wenker's *Sprachatlas des Deutschen Reichs* is considered the first linguistic atlas to be published. A partial publication of his survey was finally published between 1927 and 1956 by his successor Ferdinand Wrede, under the name *Deutscher Sprachatlas* (DSA); Wenker's hand-colored original maps have recently been made available on the Internet.[3]

Wenker's work showed that the use of a postal questionnaire was problematic for several reasons. The task of translation biased the informants and often led to unnatural dialect utterances. The teachers were not trained for their task and had to devise their own transcription systems. The origin, social affiliation and age of the informants (which were not necessarily the teachers) were not reported to Wenker. Under these circumstances, the collected data were extremely difficult to interpret.

At the same time, a fierce dispute on the nature of dialect boundaries broke out in France. The leading dialectological school of France, led by Paul Meyer and Gaston Paris, claimed that France was covered with a single dialect continuum without any clear-cut internal boundaries. This view was probably biased by the French Revolution's *égalité* motto and the political need for national unity at the time of the French-German war of 1871. Therefore, they could not accept Graziadio Isaia Ascoli's claim that France was not only divided in two dialect groups (*langue d'oïl* in the north, Occitan in the south), but in three (adding *franco-provençal* in the

3 This project is called *Digitaler Wenker-Atlas* and is available on http://www.diwa.info.

south-east). Meyer and Paris maintained that every definition of a dialect was a *definitio nominis*, an arbitrary conception of mind (Niebaum and Macha 2006, p. 81). Hence, they considered that the question of the number of dialect groups in France was devoid of sense.

In 1896, the Swiss romanist Jules Gilliéron set out to prove the claims of his preceptors Meyer and Paris, by creating a dialect atlas of France. Following the criticism of Wenker's methodology, he radically changed the method of data gathering. He designed a questionnaire in which responses could be elicited, and thus partially avoided the problem of translation. He then chose a fieldworker, Edmond Edmont, and trained him to select informants, conduct interviews and record the data in a consistent phonetic notation. From 1896 to 1900, Edmont cycled through France and conducted 700 interviews. This method yielded less, but qualitatively better data than Wenker's. The efficient organization of the project led to the immediate publication of the thirteen volumes of the *Atlas linguistique de la France* (ALF) between 1902 and 1910. It stimulated a lot of follow-up work. However, Gilliéron's atlas invalidated his working hypothesis by showing that several distinctions between northern and southern varieties coincided on a *croissant*-like line ranging from Bordeaux to the Alps. Once more, an extensive field study led to refute an earlier claim, even if Gilliéron's result finally went into the opposite sense than Wenker's…

The next milestone in the history of dialect geography was the AIS (*Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale*), conceived in 1911 by two Swiss disciples of Gilliéron, Karl Jaberg and Jakob Jud. The most important innovation of their work was the acknowledgement of the close relation between words and things. In that sense, the AIS is, as its title indicates, a linguistic and an ethnographic atlas. In contrast to previous work, Jaberg and Jud granted more freedom to the investigators. They were not confined to asking the exact questions given in the questionnaire, but could demand further precisions about a subject, note hesitations or comments of the informants, and leave out subjects that did not correspond to the ethnographic reality of the given region (Kristol 2005).

The last of the major enterprises in dialect geography is also the most relevant one for our work. The *Sprachatlas der deutschen Schweiz* (SDS) was initiated in 1935 by Heinrich Baumgartner and Rudolf Hotzenköcherle. According to Goebl and Schiltz (2000, p. 2355), the SDS represents

"eine gelungene Synthese zwischen den Atlastraditionen der Romanistik und Germanistik".[4] Besides the particular attention paid to ethnographic phenomena, it retains the interview as the method of data collection. However, the rather small investigation area of German-speaking Switzerland allowed a network of inquiry points whose density approached the one of Wenker's atlas. The major innovation of the SDS lies in the cartographic presentation: instead of displaying the pronunciations literally in the maps, the SDS uses symbols and legends. This prevents overloading the maps and guarantees an unprecedented readability. The data collection was undertaken between 1939 and 1959, and the publication of the 8 volumes lasted from 1962 until 1997 (see Chapter 3). Several subsequent atlas projects in Anglo-Saxon countries relied on the experiences of the AIS and the SDS.

The *Deutscher Wortatlas* (DWA) was intended to build a bridge between purely lexicographic work and dialect geography. Initiated in 1939, it built on the methodology and the network of the DSA, but set its accent on word geography. The 22 volumes of maps were published between 1951 and 1980.

2.2.3. New theories and new technologies

After the Second World War, dialectological research underwent four important, albeit not radical, changes.

The first change affected the scope of dialect studies. In most Central European countries, large-scale dialect atlases and lexicons were already published or in the works. The methods of data collection and presentation had been constantly refined and had reached a crest. Hence, new projects reduced their scope on smaller areas which allowed qualitatively better results with a denser network of informants. This tendency, initiated by the SDS, is described by Hotzenköcherle (1962, pp. 2-3) as a shift from *Grossraumatlanten* to *Kleinraumatlanten*:

> Der Landesatlas oder Großraumatlas zielt wesensgemäß auf den großräumigen Überblick über die weiten Gebiete einer Landessprache […]. Der Regional- oder

4 "a successful synthesis between the Romance and German atlas traditions"

> Kleinraumatlas zielt in seinem von vornherein kleinern Rahmen auf etwas ganz anderes: auf die Erfassung der Feinstruktur seines Gebiets […]. Differenziertheit und Präzision begründen recht eigentlich die Daseinsberechtigung des Kleinraumatlasses gegenüber dem Großraumatlas.[5]

As in dialect geography, this shift was also visible, for the same reasons, in the field of dialect lexicography.

The second change lay in the relationship between dialectology and linguistics. Since the period of the Neogrammarians, the contacts between linguistics and dialectology had been diminishing. In his effort to define a structuralist approach to language, Saussure (1916, p. 41) had excluded dialect-related issues from linguistics:

> Enfin tout ce qui se rapporte à l'extension géographique des langues et au fractionnement dialectal relève de la linguistique externe. Sans doute, c'est sur ce point que la distinction entre elle et la linguistique interne paraît le plus paradoxale, tant le phénomène géographique est étroitement associé à l'existence de toute langue; et cependant, en réalité, il ne touche pas à l'organisme intérieur de l'idiome.[6]

This reduction of language to a homogeneous object of study and the ensuing focus on idealized rather than raw language data surely was necessary for the foundation of linguistics as a scientific discipline, but durably alienated dialectology from structural linguistics:

> At its worst, there has been a kind of mindless friction between the two groups, with the dialectologists scorning linguists as "abstractionists" who deal in "hocus-pocus" rather than real language data, and the linguists dismissing dialectologists as "mere

5 "The national atlas or large-area atlas inherently aims at the large-scale overview of the major regions of a national language […]. The regional atlas or small-area atlas, due to its smaller scale in the first place, aims at something radically different: the representation of the microstructure of its area […]. Differentiation and precision intrinsically establish the right to exist of the small-area atlas compared to the large-area atlas."

6 "Finally, everything that relates to the geographical spreading of languages and dialectal splitting belongs to external linguistics. Doubtless the distinction between internal and external linguistics seems most paradoxical here, since the geographical phenomenon is so closely linked to the existence of any language; but geographical spreading and dialectal splitting do not actually affect the inner organism of an idiom." (Translation by Wade Baskin, Columbia University Press, New York, 2011.)

> butterfly collectors" who get so entangled in the bushes that they cannot see the trees, let alone the forest. (Chambers and Trudgill 1980, p. 17)

This situation changed in the 1950s. Weinreich (1954) introduced the concept of structural dialectology, which aimed at overcoming a major criticism against traditional dialectology: dialectal forms were studied in isolation, not in relation with opposing forms in the same dialect or in neighboring dialects. The concept of **diasystem** allowed Weinreich to study the systematic correspondences between different dialects with a structuralist approach. At the same time, he lay the base for a psychologically plausible model of a bi-dialectal speaker.

A further step was the adoption of generative grammar for dialectal studies. In that framework, dialect forms are viewed as the result of transformations applied to an underlying form which is common to all dialects. Dialect maps will thus represent distributions of transformation rules rather than distributions of surface forms (see Section 2.4). While the first works on generative dialectology were essentially concerned with phonology, linguists soon found that the generative framework could be extended to a general model of a dialect speaker. This led to the concept of **polylectal grammars**, grammars that incorporate more than one language variety and thereby account for the fact that speakers of different dialects rarely have difficulties communicating with each other (Bailey 1973).

Besides its methodological innovations, generative linguistics was also responsible for a general shift of interest towards syntax-related issues. In dialect geography, where the atlases of the 'big enterprise' period rarely had described syntactic phenomena, more recent projects integrated syntactic parameters. Milestone projects in this area are the Syntactic Atlas of the Dutch Dialects (SAND) (Barbiers, Cornips, and Kunst 2007), and the Syntactic Atlas of Northern Italy (ASIS) (Benincà and Poletto 2007). For Switzerland, the *Syntaktischer Atlas der Deutschen Schweiz* (SADS) (Bucheli and Glaser 2002) is intended to fill the gaps left by the SDS (see Chapter 3). In parallel, the newly acquired dialect data led to the creation of a new field of generative linguistics, the study of **syntactic microvariation**. Its aim is to contribute to the definition of the universal properties of the human language by studying "the patterns, loci and limits of syntactic variation" (Barbiers, Cornips, and Kleij 2002, p. 2) on the basis of dialect data.

The third change concerned the object of study of dialectology and was brought about by

demographic changes and the creation of sociolinguistics. The social changes induced by the industrial revolution had affected urban and suburban areas first, but speakers of traditional dialects could still be found in rural areas. Hence, until the 1950s, dialectologists relied on a population of non-mobile, older, rural males (subsumed under the acronym *NORM*), considered as the most conservative group of dialect speakers. With this selection, researchers could focus on areal language change (**diatopic change**) without getting entangled in non-areal parameters like social affiliation, gender, age, or the influence of the standard variety. With the non-mobile population slowly dying out in the post-war years, *NORM*s became increasingly difficult to find.

At the same time, the neglect of urban areas in dialect research was perceived as a drawback, and the interaction between the standard varieties and the local dialects was found to be an interesting area of research. These new interests were triggered by Labov's seminal work on socially conditioned speech patterns in urban environments (Labov 1966). His work led to the creation of sociolinguistics as a distinct branch of linguistics with its proper research goals and methodology. Sociolinguistics rubbed off on the branch of 'urban dialectology' in Anglo-Saxon countries, but its influence remained more contained in Continental Europe.

The demographic changes and the new "socio-pragmatic tendencies" (Goebl and Schiltz 2000) led to the transition from a monodimensional to a **pluridimensional dialectology** (Herrgen 2000). Instead of focusing on diatopic variation and abstracting away from other parameters, dialectologists began to brave the complex social realities of dialect use by studying dialect change, urban dialect mixtures and the continuum between standard varieties and local dialects.

The fourth change is related to the introduction of electronic data processing. Dialectology is one of the most data-oriented branches of language studies. Therefore, the availability of computers for data processing was bound to have an impact on the discipline. At first, computers were merely used as a tool to manage the raw data and to draw maps. While this simplified and sped up the publication process of dialect atlases, the computerization yielded no immediate methodological innovations.

In the 1980s, a technique called **dialectometry** was introduced by the romanists Jean Séguy and Hans Goebl (see Section 2.6 for a detailed review of dialectometric approaches, and Chapter 7 for an application to Swiss German data). They developed numerical measures of the linguistic

distance between two dialects which allowed them to detect and visualize spatial patterns in dialect data. Due to the complex calculations involved, dialectometry was the first major innovation in dialectology that depended entirely on the availability of computers. Dialectometry quickly became a valuable tool for the global analysis of dialect systems. However, the dialectometric approach remains purely descriptive and does not implement procedural models. In this respect, it is different from most other work in Computational Linguistics.

2.2.4. Summary

We have presented dialectology as a set of three branches: dialect grammar, dialect lexicography and dialect geography. The former two branches share most of their methodological approaches with the branches of grammar and lexicography dealing with standard varieties. In contrast, dialect geography focuses on the spatial distribution of lexical and grammatical data and can be viewed as the application of geographical data representation techniques to language. Our thesis draws heavily on previous work in dialect geography, which is why we presented its history in more detail.

In our historical overview, we distinguished traditional from more modern approaches in dialectology. Traditional projects were essentially monodimensional and were not influenced by theoretical linguistics; some of these projects are still underway today. Projects using a modern approach try to respond to some criticisms of the traditional school by integrating recent theoretical advances in linguistics, sociolinguistics and other domains. Our work can be called 'meta-dialectological' in the sense that it relies extensively on existing resources. Even if the resources we intend to use are essentially of the traditional type, the computational orientation and the generative conception of our model are compatible with the modern paradigms.

2.3. A brief history of Human Language Technologies

Since the beginning of the computer era, the processing of textual data was regarded as one of its main uses. Textual data processing does not *per se* imply the use of linguistic knowledge,

but some linguistically interesting applications have come up very quickly. One of them was **machine translation** (MT). In the context of the Cold War in the 1950s, machine translation was considered a cost-effective way of obtaining intelligence information in the Western as well as the Eastern block. At that time, many different approaches were explored; however, the development of MT systems was mainly done by engineers rather than linguists or translators. Therefore, the first models were based on crude word-by-word translation, underestimating the linguistic complexity of the translation task. This first work introduced the engineering-oriented NLP approach (as sketched in 1.1). Not much later, linguists began to show interest in MT and argued for the integration of syntactic rules in the translation process, in line with the concurrent interest in syntactic theories. This focus on linguistic adequacy led to the creation of the CL paradigm. However, the integration of syntax did not improve the systems as expected. The development costs were very high, and the computing power available at that time was largely insufficient to obtain satisfying performance, so that human translation was found to be cheaper and at the same time more appropriate. These severe criticisms led to a temporary abandon of MT, and researchers turned to the more fundamental domains of syntactic analysis and artificial intelligence (Léon 2000a).

The main goal of syntactic analysis, or **parsing**, is to break linguistic utterances up into their components and to determine their grammatical structure. Parsing techniques were pioneered by Chomsky's early work in the 1950s and 1960s. His **generative transformational grammar** allowed to describe the syntactic structure of a sentence as a tree which is built with the help of a set of rules. While this approach was appealing to linguists because it incorporated and explained fundamental properties of human language (like recursion), its procedural character was also regarded as a reasonable basis for computational implementations. On a more theoretical level, Chomsky introduced a hierarchy of formal grammars and studied their mathematical properties. This work paved the way for the development of programming languages, as well as for the implementation of generative grammars for natural languages.

The theory of formal languages not only served as a theoretical underpinning for parsing techniques, but also for **morphological analysis**. Until today, morphology tools continue to be implemented with finite-state transducers, which define a specific type of formal language (Wehrli 1997, ch. 5).

The syntactic analysis of natural language soon proved to be difficult. For instance, the parsing algorithms had to be extended to accommodate the inherent ambiguity of natural language. In fact, most sentences can be assigned several syntactic structures, and parsers are expected to retrieve them all. In order to handle these ambiguities, several efficient algorithms, based on the paradigm of dynamic programming, were proposed in the late 1960s (Wehrli 1997).

At the same time, researchers in the new field of **artificial intelligence** thought about the possibility of a computer "understanding" language. The core of this approach, inspired by Montague's work, was to create a mapping that translated syntactic constituents into a formalism of semantic representation, based on logic (Moeschler 2007). The formal representation of meaning has not yet been resolved in a satisfying manner for general-purpose systems, but it led to the creation of more task-oriented, limited-domain applications like **language understanding**, **language generation**, **dialogue systems**, and to some extent **information retrieval** (Cole et al. 1997). However, these applications are rather marginal to our own work, and we will not provide further details on their history here.

During the 1970s, few advances were made in the field of machine translation. Vauquois (1976) summarized the different approaches in use at that time in his famous triangle (see Figure 2.1). His classification is based on the three subtasks of translation: analysis (depending only on the source language), transfer (depending on the source-target language pair), and generation (depending only on the target language). Word-by-word translation systems (**direct translation**) do not have specific analysis or generation modules; the substitution of the words is done entirely by the transfer component. More sophisticated systems analyze the source text to create a structural representation of it; the **transfer** module transforms it into a structural representation of the target text, which is then handed over to the generation module to build the target text. According to the level of sophistication, these structural representations contain syntactic and/or semantic information. However, it has been argued that models relying on transfer modules are not viable for multilingual systems. In fact, a system containing n languages would require $n * (n - 1)$ transfer modules to account for all language pairs. The **interlingua** approach was designed to overcome this limitation by discarding transfer modules altogether. Instead, the source text is analyzed into a language-independent, generic structure (an *interlingua*), which is then used in the generation process to build the target text. While theoretically appealing, this model suffered from two major stumbling blocks. First, truly language-independent interlinguas

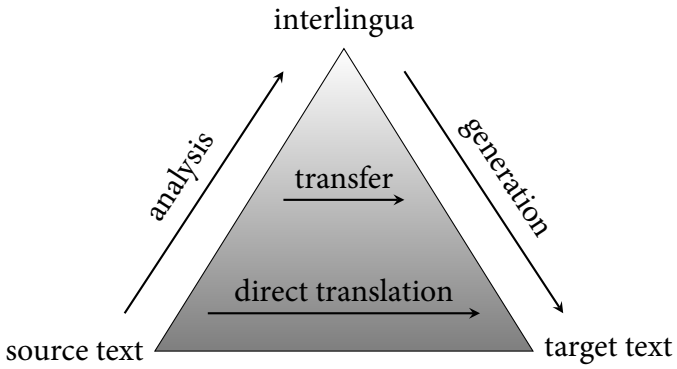


Figure 2.1.: Vauquois' triangle symbolizes different translation strategies. The lower left corner represents the source text, the lower right corner the target text. Upwards arrows represent linguistic analysis of the source text, downwards arrows represent generation of the target text, and horizontal arrows represent transfer from one language to another.

are difficult to define. Second, obtaining a complete interlingua-based representation from a source sentence would require the resolution of all potential ambiguities, which is considerably more difficult than focusing on the differences between a given language pair. These difficulties have favored MT systems that operate on the lower half of the triangle.

In the 1980s, most impetus for machine translation research came from Japan, incited by its increasing trade relations with Western economies. In the West, only a few systems (for instance, *Systran*) attained commercial productivity after long development (Léon 2000b).

At that time, research in parsing increasingly dealt with European languages. Some of these languages use quite complex morphological features, which the concept of **unification grammar** allowed to integrate into the parsing algorithm. Several unification-based frameworks were developed in the 1980s, e.g. Lexical Functional Grammar (LFG, Bresnan and Kaplan 1982), or Head-driven Phrase Structure Grammar (HPSG, Pollard and Sag 1994). In addition to unification, these frameworks were designed to keep the parsing algorithm separate from the actual linguistic data (i.e., the lexicon and the grammar rules). As a consequence, linguists without much programming knowledge were enabled to work efficiently on their respective language modules. These improvements reduced an important bottleneck that had hampered parser development:

> The strength of unification grammar formalisms lies in the advantages they offer for grammar engineering. Experience has proven that large grammars can be specified, but that their development is extremely labor-intensive. Currently, no methods exist for efficient grammar engineering. This constitutes a serious bottleneck in the development of language technology products. The hope is that the new class of declarative formalisms will greatly facilitate linguistic engineering and thus speed up the entire development cycle. (Uszkoreit and Zaenen 1997, p. 101)

In the debate on the computational properties and complexity of human language, Shieber (1985) was responsible for the proverbial fifteen minutes of fame of Swiss German dialects in the domain of computational linguistics. He showed that Swiss German cross-serial constructions (the word order discussed as "third constructions" in Appendix A.4.6, pages 375ff.) could not be analyzed by a context-free grammar. As a result, more powerful ('mildly context-sensitive') grammars have been proposed for natural language parsing. One such grammar formalism is *Tree Adjoining Grammar* (TAG, Vijay-Shanker and Joshi 1988).

Independently of parsing and machine translation, the field of **speech processing** became popular the 1980s, following work on signal processing. Its two main applications are **speech recognition** – the transformation of an acoustic stream of speech signals to a sequence of written symbols (phonemes or graphemes) –, and **speech synthesis** – the transformation of a sequence of written symbols to an acoustic stream of speech signals. Both applications started without using any linguistic information, but integration of linguistic clues was soon found to enhance adequacy and performance.

Speech recognition was at the origin of a fundamental methodological and epistemological change that came to be known as the "statistical revolution". Human speech is characterized by enormous variation: speakers have different accents, different pitches due to different physiological conditions, and they produce different kinds of phonetic assimilation. Moreover, noise and low-quality transmission channels such as telephone lines alter the speech signal. Humans are usually able to filter out this variation and to recover a generic (phonemic) description of the speech signal, but phoneticians have struggled to manually devise filtering rules precise enough for speech recognition. The solution to this dilemma lay in the use of stochastic learning algorithms that automatically discover the regularities between the acoustic signals and the

corresponding phonemes. Given a large corpus of speech data and its transcription, the learning algorithm creates and parameterizes a Hidden Markov Model (HMM). At the end of the training phase, the HMM allows to find the most likely transcriptions of a given speech signal. The resulting transcriptions are then filtered on the word level and on the sentence level, according to statistical regularities in the structure of the target language. These filters are commonly called **language models**, and they can also be learned with the help of text corpora. In an independent evolution, psycholinguists also discovered statistical models of language perception and learning algorithms. Some of their models, based on the connectionist paradigm, were also used in speech processing instead of the more common HMMs (Zue and Cole 1997).

The use of statistics and learning algorithms immediately produced encouraging results. There are three main reasons for this success. First, the impossibility of devising symbolic rules for an inherently continuous (non-discrete) phenomenon like speech triggered the search for other methodologies. Second, the computers had become sufficiently powerful to handle the complex calculations of the learning algorithms. Third, the availability of large training corpora eventually permitted the application of the learning algorithms to real language data.

At the beginning of the 1990s, the statistical revolution spilled over to machine translation. The advantages of the statistic paradigm (fast development, no need for linguistic experts) as well as the moderate success of traditional MT models based on extensive linguistic knowledge encouraged many researchers to develop learning algorithms for **statistical machine translation** (SMT). They argued that the regularities contained in corpora of human-translated text (**parallel corpora**) could be extracted automatically if they were large enough (Brown et al. 1993). The corpora made available by multilingual institutions like the Canadian or the European Parliament allowed to confirm these hypotheses. In particular, lexical correspondences and "word reordering patterns" could be learned from the corpora. As in speech recognition, language models filter out the most ungrammatical target text candidates, on the basis of the frequency of word sequences. In this early work, linguistic clues like morphological features or syntactic constituents were completely ignored. Methodologically, this approach represented thus a regression to the bottom of Vauquois' triangle, but it considerably sped up the development without much affecting the quality of the translations.

Meanwhile, traditional MT frameworks continued to be maintained and developed, with two

notable improvements. The first one, originated in Japan and called **example-based machine translation** (EBMT), integrates corpora of correct human translations. The translation process draws on these resources and recombines them according to syntactic patterns (Carl and Way 2003). The second improvement enhances the interactivity between the MT system and its users. It acknowledges that semantic ambiguities cannot be resolved automatically and in consequence relies on the competent human user to resolve them (Boitet 1997). However, both tendencies were somewhat overwhelmed by the success of the statistical paradigm.

In the 1990s, some parsers achieved sufficient performance to successfully annotate raw text corpora with syntactic information. With the help of human post-editors, parsed corpora (**treebanks**) were released for several languages. These treebanks in turn inspired researchers to apply statistical methods also to the tasks of part-of-speech tagging and parsing. **Part-of-speech** (POS) **tagging** consists of annotating each word of a sentence with its grammatical category (its part-of-speech tag); grammatically ambiguous words are disambiguated by looking at their sentential context.

While statistical part-of-speech tagging quickly yielded good results, the statistical approach did not seem promising for parsing until the introduction of lexicalized grammars (Collins 1997). In lexicalized parsing, the word forms themselves, not only their part-of-speech tags, guide the parser in determining the best attachment sites and types. Lexicalization allows to take into account collocational phenomena as well as morphosyntactic criteria which could not be modeled otherwise in the relatively simple context-free grammars used for statistical training.

While the developments of the 1980s led to an organizational separation between programmers and linguists, the rule-based systems still required the availability of competent linguists to create the rules. This changed with the statistical revolution where the linguists' and translators' work was "outsourced" to the corpus-building institutions. Following this separation, research is increasingly confined to improving the computational and statistical properties of the learning algorithms, while few efforts are directed at the creation and improvement of the linguistic resources. (This is not surprising given the enormous costs involved, and given the insecurity caused by diverging linguistic theories.) Another risk of this separation is that less important languages with few available resources are being excluded from current language technology.

Today, the limited use of linguistics in statistical models is generally seen as a drawback. Hybrid approaches combining statistical learning with explicit linguistic knowledge have been advocated for a decade, but they have not had a lasting impact so far.

2.4. Central pieces of related research

In the proceedings of the 1979 meeting of the ACL, a paper entitled "Prospects for computer-assisted dialect adaption" asked the following question:

> When a document is to appear in several dialects or closely related languages, there are many practical reasons for adapting it from one to another rather than preparing separate translations. However, manual adaption can be tedious, errorful, and requires a bidialectal adaptor (often unavailable) and/or qualified linguist (if available, very expensive). Computer-aided adaption might be an alternative, but is it feasible to write a computer program which contributes enough to be worth the bother and cost? (D. J. Weber and W. C. Mann 1979, p. 109)

The authors transform a text written in one dialect of Quechua into another dialect, word by word. Adaptation rules operate on the phonological, lexical and grammatical levels. This early piece of research is surprisingly similar to the goal pursued in our own work: we derive Swiss German dialect text from a single "glorified" dialect, Standard German. We also use regular sound correspondences to adapt words, and resort to lexical transformations when the former are insufficient. Morphological and syntactic rules account for grammatical changes.

We have not found any follow-up work by the abovementioned authors. However, their paper suggests that they consider the five Quechua dialects as clearly distinct and internally homogeneous. This contrasts with our approach, which relies on dialectological resources and does not presuppose any particular delimitation of dialect areas.

A couple of more recent papers occupy a central place in our work. We would like to present them in the remainder of this section.

Forst (2002) proposes a machine translation system from Standard German to Zürich Swiss German. His system is based on LFG. Rules for lexical and syntactic transfer are built manually on the basis of a Zürich dialect grammar.

While the use of an existing framework like LFG has undeniable advantages in terms of theoretical validity and available tools, we see some shortcomings in using this framework for dialect translation. The close relationship between Standard German and Zürich German is not reflected in the implementation. In other words, the nature of the translation rules is the same for an English-Chinese MT system than for a Standard German-Swiss German MT system. In particular, the translation lexicon is manually populated with bilingual entries, without taking into account the phonetic correspondences between the source and target varieties. Such a lexicon contains highly redundant information, but at the same time is not capable of correctly translating out-of-vocabulary words. Furthermore, Forst (2002) assumes an internally coherent "Zürich German", which is a simplification of the dialectal reality: even if Zürich German forms a more homogeneous area than other dialects, there is considerable internal variation (A. Weber 1948, pp. 19-22).

The machine translation system by Forst (2002) constitutes one of the few applications of NLP methods to Swiss German, focusing on selected syntactic features of a single dialect. However, the monograph (in fact, a Master's thesis) contains few methodological innovations that would be worthwhile pursuing in our work.

Chiang, Diab, et al. (2006) and Rambow et al. (2006) present novel research on dialect parsing. Their goal is to develop a statistical parser for the Levantine Arabic dialect (LA) by leveraging existing resources for Modern Standard Arabic (MSA). Methodologically, they combine data-driven approaches on the MSA side with explicit linguistic modeling of the relation between MSA and LA. Three approaches are pursued:

Sentence transduction uses hand-written rules to transform an LA sentence into an MSA sentence, and parses the result using an MSA parser trained on an existing MSA treebank.

Treebank transduction uses transformation rules to turn all sentences of an MSA treebank into LA sentences. The resulting LA treebank is used to train an LA parser.

Grammar transduction starts with a grammar trained on an MSA treebank and uses hand-written rules to transform the grammar rules directly. The result is a synchronous grammar (Chiang and Rambow 2006) that combines MSA grammar rules with LA grammar rules and that allows to parse sentences from either variety.

All approaches allow to obtain parse trees for LA sentences that are improved with respect to the baseline, i.e., parse trees obtained by running an unmodified MSA parser on LA sentences. The grammar transduction approach has been found to work best. This work shows that describing differences between two related linguistic systems often is easier than describing the two systems separately, in particular when one of the systems is a standardized variety for which large resources are available. Our own work on Swiss German syntax (Chapters 5.4 and 6.6) is inspired by the treebank transduction approach.

Another study on dialect parsing is reported by Vaillant (2008a); Vaillant (2008b). He presents a multi-dialect grammar for a group of French-based Creole languages of the West-Atlantic area. His grammar consists of hand-crafted rules within the Tree Adjoining Grammar (TAG) framework (e.g., Vijay-Shanker and Joshi 1988). Each rule contains a numeric parameter that specifies in which dialect(s) it is valid. This approach allows to distinguish a grammar core that is common to all dialects from a set of surface subsystems that are valid only in part of the dialects. This proceeding, called factorization, is worthwhile only if the common kernel can cover a substantial part of the grammar. This is usually the case when dealing with dialects or closely related languages. It allows economic and fast development of grammars and parsers for language families with scarce resources, since identical parts of the grammar only have to be encoded once.

To sum up, the main innovation of Vaillant's work is its ability to deal with multiple dialects in a flexible yet linguistically elegant way. He conceives the different dialects as discrete entities which can be clearly distinguished and referred to with a numeric parameter. This view is justified for Caribbean creoles spoken on different islands:

> Dans le cas qui nous a occupé, nous avons affaire à des langues insulaires, pour lesquelles, malgré une variabilité géographique interne, la question de la délimita-

> tion des variétés principales au sein d'une famille ne fait pas l'objet de polémique.[7] (Vaillant 2008b, footnote 1)

However, this conception cannot be maintained for dialect areas that lack major topographical and political borders, such as part of German-speaking Switzerland. We resort to using dialectological probability maps for this purpose.

Vaillant's multidialectal approach recalls earlier theoretical work in **generative dialectology** (Veith 1970; Veith 1982). Generative dialectology differs from generative linguistics by its inherently comparative approach:

> Das entscheidende Problem [der generativen Dialektologie] ist die Bewältigung der Vielzahl von Systemen in Beschreibung, synoptischer Darstellung und Erklärung. Geht man davon aus, daß Wenker und Wrede zwischen 1877 und 1933 Wenkersatzübertragungen aus etwa 50 000 Orten sammelten […] und daß damit fast jede deutschsprachige Siedlung Zentraleuropas erfaßt war, so wird deutlich, welche Leistung eine g[enerative] D[ialektologie] letzten Endes zu erbringen hätte: Sie müßte in der Lage sein, bis zu 50 000 Dialektsysteme […] vergleichend zu beschreiben […].[8] (Veith 1982, p. 277)

The solution to this problem lies in the use of a **reference system** from which the dialect systems are derived by transformation rules. These rules only replace those elements and structures in those dialects that differ from the reference system. For example, the following rule indicates that the lexeme *Töpfer* 'potter' of the reference system *B* is realized as *Häfner* in the dialect systems D_{33333} to D_{46999} (Veith 1982, p. 280):

(1) $\#\text{Töpfer}\#_{B} \rightarrow \#\text{Häfner}\#_{D_{33333-46999}}$

7 "In our case, we are concerned with insular languages, for which, despite internal geographical variation, the question of the delimitation of the principal varieties within a family is not subject to controversy."

8 "The crucial problem [of generative dialectology] is the handling of a multitude of systems in terms of description, synoptic representation and explanation. Assuming that Wenker and Wrede collected Wenker-sentence translations from about 50 000 places between 1877 and 1933 […] and that this collection covered nearly every German-speaking settlement in Central Europe, it becomes clear what an effort a g[enerative] d[ialectology] would eventually have to render: it should be able to comparatively describe up to 50 000 dialect systems […]."

A single lexicon containing such rules may thus fully describe the 50 000 German dialectal varieties. Veith further analyzes sound change in terms of phonetic feature transformations and uses morphosyntactic features in morphological transformation rules. These rules apply in cascade. The order of the rules – a thorny issue in rule-based NLP systems – receives a precise interpretation here: it corresponds to the historical order in which dialect changes have propagated.

The main contribution of generative dialectology is to view dialect change as processes, or rules. This conception is in line with our approach: we take Standard German as a reference system and build rules that specify which linguistic elements change in which dialects (see Section 1.2.8). By re-interpreting dialectological maps of single words as maps of general transformation rules, we adopt Veith's conception of "Sprachkarten als Regelkarten".[9]

2.5. Closely related languages and cognate words

Dialectal varieties can be viewed as closely related languages. Closely related languages are characterized by a high proportion of **cognate word pairs**, i.e., word pairs that share a common etymology and similar surface forms. This characteristic has been leveraged in recent research on machine translation.

Simard, Foster, and Isabelle (1992) have found that cognate word pairs improve the sentence alignment in a parallel corpus. They consider two words as cognates if their first four characters are identical. More commonly, cognates are used for word alignment and the resulting constitution of bilingual translation lexicons. For example, Koehn and Knight (2002) are able to find cognate word pairs in monolingual (not parallel) corpora. They use a measure called **longest common subsequence ratio** that counts the length of the longest common character sequence of two strings, divided by the length of the longer of the two strings. Yzaguirre et al. (2000) also show improved alignment performance for related languages by integrating word similarity measures.

9 "language maps as rule maps"

G. S. Mann and Yarowsky (2001) use Levenshtein distance thresholds to identify cognate word pairs. Levenshtein distance (Levenshtein 1966) counts the number of edit operations to transform one string into the other; allowed edit operations are (i) the insertion of a character, (ii) the deletion of a character, and (iii) the substitution of a character by another. G. S. Mann and Yarowsky (2001) have found that word pairs having a Levenshtein distance of less than 3 are generally cognates. In their work, they present a variant of Levenshtein distance that counts substitution operations from one vowel to another vowel only as 0.5, on the grounds of the intuition that vowels are less stable across languages than consonants. Furthermore, they propose another model that learns weights for every edit operation on a small corpus of cognate word pairs; in this way, the definition of cognate pairs can be adapted to a particular language pair according to the historical sound and spelling changes that have separated the two languages from their common origin. Mackay and Kondrak (2005) and Kondrak and Sherif (2006) present several similarity algorithms for cognate identification; they present different learning models to induce similarity measures.

In earlier work (Scherrer 2007a; Scherrer 2008b), we have experimented with different cognate identification methods to induce a translation lexicon between Standard German and Bernese Swiss German. Besides several flavors of Levenshtein distance, we used a stochastic transducer and a rule-based system containing the most common phonetic correspondences between the two language varieties. While this research showed promising results, it proved difficult to extend to the entire dialect area of German-speaking Switzerland. Some of our recent dialectometric work relies on cognate identification methods as well (see Section 7.3.5).

String distance algorithms (such as Levenshtein distance) have also been used for related tasks like letter-to-phoneme conversion (Jansche 2001; Jiampojamarn, Kondrak, and Sherif 2007), transliteration (Sherif and Kondrak 2007), dialectometry (see Section 2.6 below), and diachronic phonology (Bouchard-Côté et al. 2007).

An interesting approach to closely related languages is character-based machine translation. Most common statistical machine translation systems translate words and sequences of words (commonly called **phrases**). Character-based machine translation uses the same methods and tools, but operates on a lower level of linguistic analysis: it translates characters and sequences of characters (morphemes or words). Obviously, this approach only works between two related

languages, where the structures of the words as well as the syntactic sentence structures are similar. Tiedemann (2009) views character-based machine translation as a special case of transliteration: "In translation, the general assumption is that, similar to the transliteration task, many correspondences between lexical items of related languages can be explained on the character level."

Vilar, Peter, and Ney (2007) pioneered character-based machine translation and applied it to Spanish-Catalan. While the character-based approach did not outperform the word-based approach, the former correctly translated out-of-vocabulary words. If training corpora are small, this benefit should not be underestimated. Tiedemann (2009) finds similar results for Norwegian-Swedish. Common machine translation metrics like BLEU (Papineni et al. 2002) or translation error rate (TER, Snover et al. 2006) are problematic in this context because they evaluate the translation of a word on a binary scale (correct or wrong). Therefore, Tiedemann (2009) tentatively proposes to measure translation quality in terms of longest common subsequence ratio. Our proposal to use phonetic rules to locally transform sounds in Standard German words follows the same idea.

2.6. Dialectometry

Dialectometry uses statistical and mathematical methods to analyze aggregated dialect data. In other words, data containing dialectal variation patterns of several linguistic phenomena are brought together so that a single synthetic picture of the dialect landscape emerges. Following pioneering work by Séguy (1973a), dialectometry has been constituted as a scientific discipline by Goebl (1982); Goebl (1984). He introduced several visualization techniques: (i) **similarity maps** show the similarity of all other data points with respect to one fixed data point; (ii) **parameter maps** combine similarity maps for all data points, so that transition zones form "valleys" and stable zones form "plateaux"; these may be drawn in the form of digital elevation models; (iii) **interpoint maps** represent the distance between adjacent areas by the thickness of the separation line. He also proposed weighting techniques like *Gewichtender Identitätswert* (GIW, Weighted Identity Value), whose rationale is that agreement among rare variants is less probable than agreement among frequent variants and that the former should be weighted more

heavily in the computation of linguistic distances (Goebl 1984, ch. 3). As a romanist, Goebl mainly worked on Italian and French dialect data. Similar dialectometric studies were done early on with German and English material (Goebl and Schiltz 2000, p. 2362; Lee and Kretzschmar Jr. 1993). Kelle (2001) analyzes a small sample of Swiss German dialect material.

The Salzburg school of dialectometry founded by Goebl uses discrete nominal units of measurement: the recordings at two inquiry points are either identical or different. However, single words may exhibit different independent phenomena, and they should all be captured by dialectometric measures. Kessler (1995) uses Levenshtein distance to quantify the linguistic distance of two word recordings made at different inquiry points. This line of research has been continued by Heeringa (2004), who experimented with different types of string distance algorithms. These techniques were applied to Dutch (Heeringa 2004), Norwegian (Gooskens and Heeringa 2004), German (Nerbonne and Siedle 2005), and Bulgarian (Osenova, Heeringa, and Nerbonne 2009) dialect data.

Rumpf et al. (2009) challenge the view that the aggregation of large numbers of linguistic phenomena is required for obtaining interesting dialectometric results. They use methods from spatial statistics to analyze single linguistic phenomena (see Section 4.6.1 for a detailed discussion and application of this work). Their follow-up work (Rumpf et al. 2010) reintroduces aggregation, albeit in a different way: whereas classic dialectometric work measures similarity of locations on the basis of the distribution of linguistic variants, they measure similarity of linguistic phenomena on the basis of their respective geographical distribution.

In Chapter 7, we review some of the most popular dialectometric analysis and visualization techniques and apply them to our Swiss German data.

2.7. Computational dialectology

Apart from dialectometry, computers are used in various other settings to analyze dialect data. Several studies deal with the acquisition of written dialect data. In recent years, electronic media like chat, SMS, blogs and micro-blogging services have proved interesting for the study of dialects. Indeed, many users write in their dialect on these informal data channels. Siebenhaar

(2005) examines data from different regional chat rooms of German-speaking Switzerland. He retrieves some patterns of regional dialect variation, even if dialect spelling and writing is quite inconsistent among users. As part of the international *sms4science* project, SMS messages have recently been collected in Switzerland (Stähli, Dürscheid, and Béguelin 2011). A large part of these messages are written in Swiss German dialect, but they have not been analyzed yet. The collection of written dialectal data poses various challenges in terms of dialect identification, spelling normalization, and morphosyntactic annotation. Diab et al. (2010); Benajiba and Diab (2010) discuss the methodological choices made in the COLABA project covering blog posts of multiple Arabic dialects.

Eisenstein et al. (2010) use data from Twitter to infer geographical patterns of lexical variation. On the basis of Twitter posts from different parts of the USA,[10] they create topic models and region models using sophisticated machine learning techniques. They find, perhaps unsurprisingly, that the use of sports team names varies a lot across regions (e.g. *Celtics* is predominant in Boston, *Lakers* in Los Angeles), as well as toponyms (like *Hollywood, Oakland*). More interestingly, some slang terms were found to have strong regional biases. These results seem rather shallow given the complexity of the model; it remains to be seen if data from dialect regions with more internal variation than American English yield more interesting results.

Another issue that has recently received attention is the digitization and georeferencing of existing dialect lexicons. Wandl-Vogt (2006b) and Wandl-Vogt (2008) present current work on the Lexicon of Bavarian Dialects of Austria. The lexicon entries were initially provided with heterogeneous indications about their geographical distribution (municipality, valley, district). These indications are digitized and georeferenced so that they can be displayed on maps. Landolt (2007) reviews other issues of current lexicography and makes proposals for the digitization of the Swiss German *Idiotikon*.

10 A lot of Twitter posts written on mobile devices are marked with the GPS coordinates of the author.

2.8. Spoken dialect

In general, dialects are spoken, not written. It is therefore logical that the phonetic and prosodic properties of specific dialects have been studied in the fields of speech recognition and speech synthesis. Although we do not make use of spoken dialect data in our work, we shortly review some of the most relevant research here.

Siebenhaar, Forst, and Keller (2006) study the prosodic parameters of two Swiss German dialects required for the development of speech synthesis systems. They analyze interviews conducted with Zürich and Bern dialect speakers and conclude that rhythm and intonation differ between dialects. These differences are then used to build specific text-to-speech systems for the two dialects.

Pucher et al. (2010) build synthesis systems for Standard Austrian German and the Viennese German dialect. They adopt the framework of HMM-based speech synthesis, which simplifies specific parameterizations, e.g., for speaker adaptation and emotional speech. In contrast to these adaptations on the suprasegmental level, dialect adaptation also requires modifications in the phone inventory. Pucher et al. also investigate interpolation techniques to model the vertical continuum between Standard Austrian German and the Viennese dialect.

Biadsy, Hirschberg, and Habash (2009) deal with a related problem: dialect identification of spoken data. They use phone recognizers to obtain a phonetic transcription of the spoken data, and then determine the source dialect on the basis of the phone distribution patterns in this transcription. They test their model on five Arabic dialects. Since a Standard Arabic phone recognizer does not contain all dialectal phones, Biadsy, Hirschberg, and Habash (2009) use several phone recognizers (for English, Hindi, Japanese, etc.) in parallel. We borrow some ideas from this dialect identification model in Section 6.5.

2.9. Computational processing of historical texts

Swiss German dialect texts are characterized by a high degree of diatopic variation and spelling variation. These two types of variation also appear in historical texts, i.e. texts that have been written before an official standardized variety has come into effect.

One possible approach to the computational analysis of historical text is normalization: different spelling variants of a word are reduced to a single canonical form, which may or may not correspond to the modern, standardized form. For example, Baron, Rayson, and Archer (2009) present a normalization tool based on spelling correction techniques and explore the potential of this tool on texts from Early Modern English.

Another approach is to manually annotate historical corpora, so that these annotations can be used to train NLP tools like part-of-speech taggers and parsers. Gleßgen and A. Stein (2005) present some recent work on Old French data.

Höfler and Piotrowski (2011) describe the construction of a corpus containing historical Swiss legal texts from the early Middle Ages up to 1798. These texts are written in various historical and regional varieties of German, French, Italian, Romansh and Latin. Piotrowski and Senn (2012) derive a controlled vocabulary from the indices of this corpus. However, they only use modern index terms and avoid the historical terms because of their high degree of variation. Future work is intended to normalize different historical spellings of related terms.

2.10. Natural Language Processing of Standard German

Compared to other languages, Standard German has proven rather challenging from a Natural Language Processing point of view. Among the problems, one may mention the richness of the inflectional morphology, the productive compounding mechanism, the presence of discontinuous constituents, and the word order that is relatively free and differs from other commonly used Western languages. In this section, we discuss the application of morphological analyzers, part-of-speech taggers, parsers and machine translation systems to German.

Mahlow and Piotrowski (2009a) review four morphological analyzers for German. Three of them achieve similar performance levels: about 60% of the tested words yielded all correct analyses and no incorrect analyses. In general, they also perform lemmatization and compound splitting. In our research, we will use *Morphisto* (Zielinski and Simon 2009). It is based on the Stuttgart finite state tools (Schmid 2005), an extended morphological grammar coming with the latter (Schmid, Fitschen, and Heid 2004), and a large-coverage lexicon based on the *DeReWo* word list (Institut für Deutsche Sprache, Programmbereich Korpuslinguistik 2007). The advantages of *Morphisto* compared to other morphology tools are its open-source availability, its speed, and its use of the *Stuttgart-Tübingen Tag Set* (STTS, Thielen et al. 1999), which has become standard for German NLP. Nevertheless, Mahlow and Piotrowski (2009a) point out some drawbacks of *Morphisto*: it is very memory-intensive to compile, and it performs less well than its competitors. This is mainly due to the way *Morphisto* analyzes compound words. It returns the citation form of every component of a compound, from which it may be difficult to derive the actual lemma. For example, *Zeitungsleser* 'newspaper reader' is split into *Zeitung+lesen+er*, whereas one would rather obtain the complex lemma *Zeitungs+leser*.

Morphological analyzers output all morphotactically possible readings. The task of disambiguating this output with respect to the sentential context is known as part-of-speech tagging. Statistical POS taggers are based on Markov models: they look at the two or three preceding words and their (previously estimated) POS tags to derive the POS tag for a given word. T. Brants (2000) presents such a tagger, called *TnT*. It is available with a pre-trained model for German. Schmid (1995) discusses *TreeTagger*, a similar tagger that uses decision trees for estimating the probabilities. It is available as well with a pre-trained model for German. In addition to part-of-speech tagging, *TreeTagger* is able to perform lemmatization. The performance of both taggers is comparable and lies at about 96% accuracy.

Statistical parsers have to be trained on treebanks. For German, three treebanks have been built: *Negra* (Skut et al. 1997), TIGER (S. Brants et al. 2002), and *Tüba-D/Z* (Hinrichs, Kübler, et al. 2004). Kübler, Hinrichs, and Maier (2006) have shown that the treebank annotation style has a large influence on parsing quality; parsers trained on *Tüba-D/Z* systematically outperform parsers trained on *Negra*. These findings have been replicated in Kübler (2008). This shared task on parsing German features three state-of-the-art parsers: the Berkeley parser (Petrov and Klein 2008), the Stanford parser (Rafferty and Manning 2008), and *MaltParser* (Hall and Nivre

2008). The first two are available for download with a pre-trained German grammar. In contrast to other statistical parsers, these three tools do not require part-of-speech tagged input.[11]

While lexicalization has improved statistical parsing of English, it has not yielded the expected benefits for German parsing, largely because of the larger morphological complexity of German, which leads to a higher type/token ratio and in turn to sparse data problems. Dubey (2005) explores several alternative approaches; he uses suffix analysis, smoothing techniques and treebank transformations with success.

Besides the abovementioned parsers, we present another small sample here. *BitPar* (Schmid 2004; Schmid 2006) is an efficient statistic parser that comes with a German grammar trained on the TIGER corpus. Sennrich et al. (2009) present a hybrid parser that combines a rule-based grammar with a probabilistic disambiguation system and relies on a separate morphological analyzer. It obtains similar accuracy figures to *MaltParser*. Foth, Daum, and Menzel (2004) developed a parser based on weighted constraint dependency grammar using hand-written grammar rules and statistical helper modules.

Fips (Wehrli 2007) is a rule-based multilingual parsing engine. Its German module contains a morphological analyzer that includes compound splitting (Ulmann 1995). The parser can also be used for part-of-speech tagging, albeit with inferior performance compared to a statistical part-of-speech tagger (Scherrer 2008a). The aim of *Fips* is to build linguistically satisfactory and multilingually valid sentence structures consistent with recent work in Generative Grammar. This aim is at the same time a strength and a drawback: its analyses use tag sets and constituent types that do not coincide with those used in the existing German treebanks, which complicates its comparison and evaluation. We have attempted such a comparison in terms of part-of-speech tagging, lemmatization and morphological analysis, but not in terms of parsing (Scherrer 2008a).

Machine translation into and out of German has also proved difficult. One difficulty concerns the large proportion of compound nouns used in German text, many of which do not figure in the translation lexicon. Koehn and Knight (2003) and Popovic, D. Stein, and Ney (2006) address this issue. Another problem is the largely differing word order between German and

11 However, *MaltParser* works much better when using pre-tagged input data (J. Hall, p.c.).

English (the language from which or into which German is most commonly translated). In statistical machine translation, this issue is dealt with by reordering the source text. For example, when translating from English to German, one first transforms the English sentences so that their word order resembles German word order, and then translates the words contained in the reordered English text (Nießen and Ney 2004; Collins, Koehn, and Kučerová 2005). Such systems rely on a parser either on the source or on the target side in order to perform these word order transformations. Besides compounding, inflection is another source of data sparseness problems and consequently poor translation quality. Fraser (2009) uses an intermediate representation called "simplified German" that only contains word stems. Inflection markers are added explicitly to these stems in a separate step.

These pre-processing and post-processing methods are effective, but linguistically not very satisfactory; ideally, a single model should account for all operations involved in translation. Machine translation systems based on hand-crafted transfer rules offer such a single model: the source text is parsed and disambiguated, then transferred on the syntactic and lexical level in order to create the target text. *Its-2* is such a model and is presented in Wehrli, Nerima, and Scherrer (2009). However, the transfer approaches often lack the wide coverage obtained by statistical models trained on large corpora. Another possibility is the use of synchronous grammars, which learn the syntactic structures of both languages at the same time, but generally require treebanks for both languages. This line of research has been explored by Chiang (2010) and is at the heart of the grammar transduction approach for Arabic dialects discussed above.

The issues described above are less critical when translating from Standard German to Swiss German dialects. While word order changes, discrepancies in the morphological inventory and different compounding elements exist, they can be dealt with more easily. Nevertheless, we rely on a linguistic (morphological and syntactic) analysis of the Standard German input to resolve these issues.

2.11. Conclusion: machine translation with scarce resources *vs.* machine translation for closely related languages

As we have mentioned earlier, we build a set of algorithms that adapt Standard German words and sentences so that they become Swiss German words and sentences. Essentially, this can be viewed as a case of machine translation. Given these settings, two recent approaches are of particular interest to us: machine translation with scarce resources, and machine translation for closely related languages.

The training of standard SMT systems requires large corpora of parallel text. Such data are only available for a limited number of language pairs. However, there is a growing (civil and military) demand for MT systems dealing with languages of secondary economic importance. The answer to this demand usually consists of low-quality systems that can be implemented easily and quickly while covering the most basic needs, i.e. getting the gist of otherwise incomprehensible texts. One particular feature of such systems is that their training algorithm can do without (or with very small) genuine parallel corpora – this is, basically, the **scarce resource** approach to machine translation.

At first sight, such an approach seems ideal for Swiss German. In a diglossic society, parallel corpora hardly exist because the need for translation between the two language varieties is limited. However, even monolingual text corpora are hard to find, as the amount and quality of available written data is affected by the mainly spoken nature of dialects and by the absence of binding orthographic rules. Anyway, there is a dilemma associated to the use of textual data to train statistical models. If we want to draw an adequate picture of the Swiss German dialect landscape, we would need to build several models (for several dialects), each model requiring its own training corpus. It seems extremely unlikely that sufficient data be available for a dozen different models. Otherwise, we could combine all the data to build one single model of an "average Swiss German". Besides the interesting discussion that this concept would call for, we would miss our goal of formalizing the diatopical variation. But there is a way out of this dilemma: Swiss German dialects have been studied extensively by dialectologists, and their results merit to be included in our model; this is difficult to achieve with current SMT frameworks.

Already before the statistical revolution, computational linguists have been attentive to the etymological and typological distance between the languages of a MT system. Simply put, the more the languages differ, the higher up one has to go on the Vauquois triangle (Figure 2.1) to maintain constant translation quality. In other words, closely related languages allow to obtain higher translation quality with simpler methods: they are likely to possess similar grammatical and lexical inventories, similar multi-word expressions and a similar word order. In contrast to the models for scarce resource languages, the aim of MT systems for **closely related languages** has always been high quality rather than simplicity of implementation. The reason for this is simple. For instance, any native speaker of Italian is able to decipher Spanish texts and to get the gist out of them. A low-cost, low-quality MT system would not provide any further benefit.

The different Swiss German dialects, as well as Standard German, are closely related language varieties. As in other situations of translation between closely related languages, we consider high-quality machine translation based on the integration of explicit linguistic information as the most relevant approach. In particular, and contrary to many current approaches, we believe that the syntactic analysis of the source text yields important clues that improve the result of the translation. For example, the Standard German (synthetic) preterit tense must be translated in Swiss German as a (compound) perfect tense. The lexico-syntactic analysis of the source text detects preterit forms, provides the position on which the participle has to be inserted as well as the type of auxiliary verb to be used. The need for a deep analysis of the source text is the main reason why we focus on Standard German as a source language: as shown in Section 2.10, many such tools currently exist for Standard German. One particularity of closely related languages is their large proportion of cognate word pairs (see Section 2.5), which can be modeled by regular phonological correspondences.

To sum up, our approach is more inspired by the "closely related languages" setting rather than the "scarce resources" problem. It follows from this choice that we favor a machine translation model that relies on explicit linguistic information. However, this does not mean that our approach completely forgoes the use of statistics and probabilities. In fact, many dialect features geographically overlap and form areas of concurrent use. Such situations are best implemented in a model based on probabilistic rules. The probabilities will not be estimated from text corpora (because such corpora are not available), but rather from the results of dialectological fieldwork.

3. Data sources

3.1. Overview

The development of a multi-dialectal model for word and sentence generation requires a thorough knowledge of the linguistic particularities of the modeled language varieties. This chapter deals with the different data sources that describe the Swiss German dialects.

The dialects of German-speaking Switzerland have been described from various perspectives. General linguistic descriptions of Swiss German (Haas 2000; Baur 1983; Lötscher 1983; Rash 1998) provide a good starting point, but they mostly focus on an "average" Swiss plateau dialect[1] and forgo a detailed description of the inter-dialectal differences. Dialect grammars, such as those existing for the dialects of Zürich (A. Weber 1948), Basel (Suter 1976), Luzern (Fischer 1989) and Bern (Marti 1985), describe particular dialects in much detail, but do not adopt a comparative approach and do not cover the whole Swiss German dialect area. Case studies, like Penner (1995) for syntax, provide additional information about specific phenomena, albeit without geographic references. Publications in dialect geography, such as the dialect atlas SDS (Hotzenköcherle, Schläpfer, et al. 1962-1997), as well as the synopsis by Hotzenköcherle (1984), provide detailed geographical variation patterns of selected linguistic phenomena. The selection criteria of the included phenomena are based on dialectological interest, not – as we would prefer for our application – on frequency and salience. Some recent case studies drawing on

1 The Plateau (*Mittelland* in German) commonly refers to the densely populated plain reaching from Freiburg in the West to the Bodensee in the East of Switzerland. It is bordered in the North by the Jura mountain chain and in the South by the Prealps and the Alps. In our use of this term, we include the Northern Jura regions and partially the Prealps, but sometimes exclude the Freiburg dialects which go with the Southern Alpine dialects.

data from the (yet unpublished) syntactic atlas SADS complete this picture.

The ideal foundation of our implementation should be something like a comparative multi-dialect grammar of Swiss German. It should be **comparative** in the sense that Swiss German features are compared with their Standard German counterparts, and **multi-dialectal** in the sense that it it intends to describe the linguistic variables according to the entire Swiss German dialect continuum. As said above, such a document does not exist. We therefore compiled our own multi-dialect grammar from the mentioned data sources. For reasons of space, the dialect grammar is presented in Appendix A.

For most dialectal phenomena described in our grammar, their geographical distribution is presented in more or less detail in one of the Swiss German dialect atlases SDS (*Sprachatlas der deutschen Schweiz*) and SADS (*Syntaktischer Atlas der deutschen Schweiz*). Therefore, we discuss the history and methodology underlying these atlas projects, as well as the criteria that we used to select a subset of their content. The SDS is presented in Section 3.3, the SADS in Section 3.4.

An important issue in our work are spelling conventions. We deal with written dialect data, and dialect spelling is not standardized. In fact, the granularity of our implementation is defined by the granularity of the spelling system that we intend to use. Section 3.2 details the spelling conventions that we use throughout our work. In particular, the examples in Appendix A are formatted according to these guidelines.

3.2. Spelling

Our model deals with written Standard German and written Swiss German, rather than with phonetically transcribed Standard German and phonetically transcribed Swiss German. This choice simplifies some tasks, but creates additional difficulties as well. Therefore, it is important to look closely at the spelling system for Standard German, and at the various ways of spelling Swiss German.

3.2.1. Standard German spelling

In Standard German spelling, the grapheme-phoneme correspondence is quite simple, in comparison with other languages like French or English. However, some particularities are worth pointing out (Dieth 1986, pp. 13-14):

- Vowel length is inconsistently spelled: by duplicating the vowel (1a), by adjoining a *h* (1b), by adjoining a *e* to *i* (1c), or simply by the absence of doubling of the following consonant (1d):

  (1) a. Saal 'hall'
  b. zahlen 'to pay'
  c. viel 'much'
  d. malen 'to paint'

- The diphthong qualities are not written consistently. *au* is not pronounced [au], but rather [aʊ]; *äu* is not pronounced [ɛu], but rather [ɔʏ]. The grapheme *ei* is not pronounced [eɪ], but [aɪ]. This is particularly problematic because in Swiss German both [aɪ] and [eɪ] occur and need to be distinguished.

- Standard German spelling does not distinguish open and closed vowels. As a rule of thumb, long vowels are closed, while short vowels are open.

  (2) a. lesen [eː] / Bett [ɛ] 'to read / bed'
  b. Not [oː] / Gott [ɔ] 'emergency / god'
  c. Nöte [øː] / Götter [œ] 'emergencies / gods'

- Double consonants serve to mark the shortness of the preceding vowel rather than the length (or gemination) of the consonant itself.

- Aspiration of voiceless plosives is not written:

  (3) a. Kind [kʰ] 'child'
  b. Tau [tʰ] 'dew'

- Word-initial *st* and *sp* are pronounced [ʃt] and [ʃp].

3.2.2. The Dieth spelling rules

With the first wave of Swiss German dialect literature and dialect philology at the end of the 19th century, different authors have come up with different spelling conventions. On the one extreme, writers took the Standard German spelling as a model, with few adaptations required by major phonetic differences. This approach is still very popular in the Bern dialect tradition (Marti 1972). On the other extreme were the defenders of a strictly phonetic approach to spelling. Under this approach, the surface forms resulting from phonetic and prosodic assimilations were written, with the result that such texts unsettled many readers (who were accustomed only to Standard German writing).

As a compromise between these two approaches, the phonetician Eugen Dieth conceived a set of spelling conventions that would be easy to read, yet allow fine-grained phonetic distinctions to be expressed clearly. After a first edition in 1938, his proposals were re-edited nearly 50 years later (Dieth 1986). The Dieth spelling system has been fairly successful; it is used in most scientific accounts of Swiss German dialects, and also by a large number of dialect writers. However, it is not taught to native dialect speakers, so that the spelling used in blogs or e-mails may or may not follow this system. Usually, laymen dialect writing is closer to Standard German spelling than the Dieth guidelines suggest.

The Dieth convention uses Standard German spelling as a starting point, but deviates where it is inconsistent or lacks the precision needed for the description of the various Swiss dialects. For example, vowel length is uniformly indicated by letter doubling; *h* cannot be used to indicate length, and *ie* refers to a diphthong rather than to a long *i*. While Standard German does not systematically distinguish geminated fricatives from simple ones, this distinction is maintained in the Dieth system by doubling the consonants (*ss*, *ff*) or by underlining complex graphemes (*ch*, *sch*); underlining is optional.

In contrast to Standard German, the openness of Swiss German vowels does not depend on their length. In the Dieth system, vowel quality is marked with diacritics. A grave accent denotes

open vowels (as in *è, ì, ò, ȍ, ù, ǜ*), while a tilde denotes nasal vowels (*ẽ, õ*).[2] Acute accents can be used as stress marks. However, the use of diacritics is not compulsory. Dieth (1986, p. 15) suggested a "narrow" variant of his spelling (*enge Dieth-Schreibung*) that included diacritics for phonetic precision, and a "broad" variant (*weite Dieth-Schreibung*) which would not use diacritics and would be sufficient in contexts where phonetic details were not required.

3.2.3. Conclusion

Like most other scientific work in Swiss German dialectology, we adopt the Dieth spelling rules. We use the broad variant without diacritics. While this decision was a difficult one for personal reasons,[3] its practical advantages are apparent: vowel quality is hard to detect in the Standard German source, and absent from most of the data that we plan to evaluate our system on. The distinction of vowel qualities would therefore be very costly to implement, but only yield marginal benefits (at least as long as we do not aim to build a dialect speech synthesis system). Note however that diacritics may be used in Appendix A to help the reader understand the relevant phonetic issues.

3.3. *Sprachatlas der deutschen Schweiz*

3.3.1. History of the project

The project of a dialectological atlas for German-speaking Switzerland was initiated in 1935 by Heinrich Baumgartner (Bern University) and Rudolf Hotzenköcherle (Zürich University).[4] In several aspects, that time period was propitious for such an undertaking. Dialectological re-

2 Dieth defines the letter *y* as an alternative spelling for closed *i*. For reasons of consistency, we do not use *y*.

3 The lack of discrimination between [o] and [ɔ] (in Standard German spelling) managed to puzzle me in the earliest attempts of acquiring writing skills. This early metalinguistic experience probably has contributed to my interest in linguistics some 15 years later.

4 This section is largely based on the historical account given in Trüb (2003).

search had been done for 50 years in German-speaking Switzerland, but mostly in monographic form rather than as a systematic overview (with the exception of the *Idiotikon* project, Staub et al. 1881-). With the previous realization of dialect atlases in France, Germany and Italy, a consensus on its form and content had emerged. Moreover, the 1930s still featured a traditional agrarian society where social migrations were rare and local dialects well preserved.

The following years were dedicated to the planning of that Herculean project, to be named *Sprachatlas der deutschen Schweiz* (abbreviated SDS). A fine-grained network of inquiry points was defined: on average, every third village of German-speaking Switzerland was to be visited. In general, two informants were solicited in every inquiry point. A questionnaire consisting of 4000 questions was elaborated, based on earlier dialectological work. After some test interviews in 1939, the questionnaire was reduced to about 2500 questions. The informants' answers were directly transcribed by the fieldworker, yielding about 250 pages of notes per informant.

The fieldworkers' notes contain several types of data. The **primary material** consists of the first answer given by the informant to a precise question of the questionnaire. **Secondary material** is obtained by reflection on the first answer given; for instance, the informant might correct himself, give information about older or more recent forms, or compare his pronunciation with the one in neighboring villages. **Spontaneous material** is not related to the particular question being asked, but consists of spontaneous utterances deemed relevant by the fieldworker. Spontaneous material was most tedious to be post-processed, because it had to be linked to the relevant questionnaire page. In addition to linguistic data, the notebooks also contain the place and conditions in which the inquiry took place, as well as factual information, e.g. about the architecture of the houses or about the use and look of certain tools.

During World War II, recordings could only be carried out when the fieldworkers were on leave from military service. The pace was accelerated after the end of the war, so that 625 villages were visited until 1958. Most recordings took place in winter, when the informants were less busy with farm work. Working conditions were difficult, as guesthouses were not available in all villages. The explorers traveled by train and bus, sometimes by bicycle. Some villages were accessible only on foot or on skis.

While the recordings were still underway, the material was sorted and prepared for publication. The graphic design of the maps was defined, as well as the number and content of the different

volumes. In this stage, the project leaders decided to use symbols in the map. In earlier work, the exact pronunciations had been written directly onto the map, but the dense network of inquiry points in the Swiss German project would have led to massively overloaded and unreadable maps. Examples of these two map types can be found in Figures 3.1 and 3.2.[5] On the downside, the use of symbols required a preliminary interpretation of the raw material, in order to create classes to be represented by distinct symbols. This greatly increased the work load.

All maps were drawn by hand by a graphic designer. This allowed for variability in the layout of the maps, so that the most adequate design could be chosen for each phenomenon. For instance, isogloss lines and hachures were used instead of symbols in simpler cases. Different aspects of a phenomenon are illustrated by overlaying different cartographic techniques: symbols can be combined, symbols can be drawn in different colors, and symbol layers can be combined with hachure layers. Iconicity was respected as far as possible, so that similarity in symbols corresponds to similarity in the linguistic material. Quantitative differences were encoded by special symbols consisting of empty and full squares. Nevertheless, the classification of the raw data into symbol types sometimes commanded difficult choices that could not always be supported by safe etymological evidence (Trüb 2003, p. 46).

The fieldworkers used a specific transcription system, which was based on the Latin alphabet and additional diacritics for fine-grained distinctions. This system continued to be used in the first published volumes, where phonetic details were crucial. In the volumes dealing with lexical types, the more coarse-grained, but more easily readable Dieth spelling system was used instead.

In 1962, only three years after the last recording, the first volume of the SDS was published. After many personal changes, the last and eighth volume was achieved in 1997. Table 3.1 lists the publication dates. All in all, the eight volumes contain 1548 maps on a total of 1926 pages. The volumes are described as "in der Form und im Gewicht noch handlich"[6] (Trüb 2003, p. 37). We do not completely agree with this characterization: an open volume is more than a meter long. The format of one page is larger than A3, which makes it impossible to reproduce these

5 Throughout this thesis, SDS maps are cited by mentioning the volume in Roman numerals and the page number in Arabic numerals.

6 "still handy in form and weight"

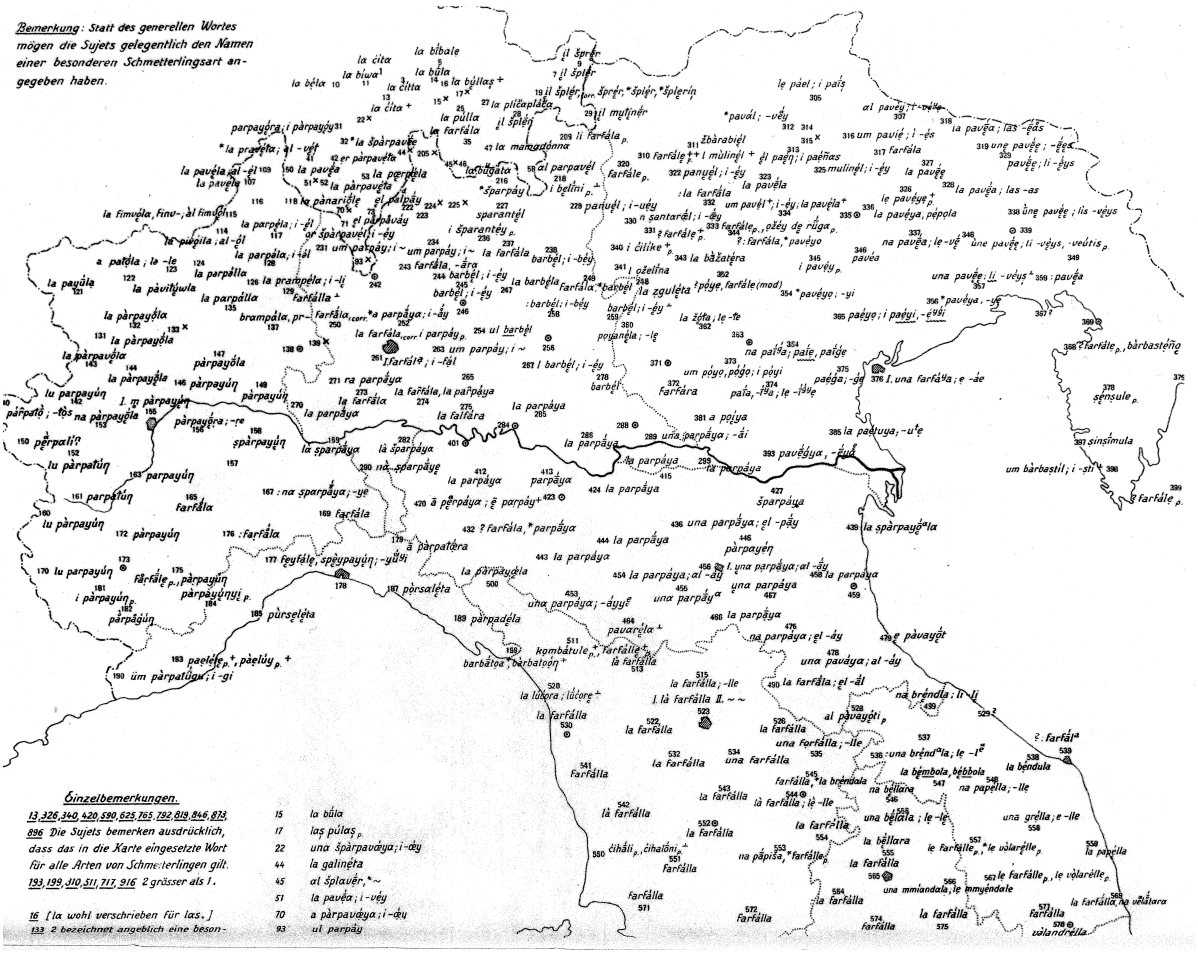


Figure 3.1.: Detail of map 480 from the Linguistic Atlas of Italy and Southern Switzerland (AIS, Jaberg and Jud 1928-1960), presenting the dialectal variants of the word 'butterfly' (*farfalla* in Standard Italian). It shows the problem of overcharging, which would only have been exacerbated with the denser network of the SDS inquiry points.

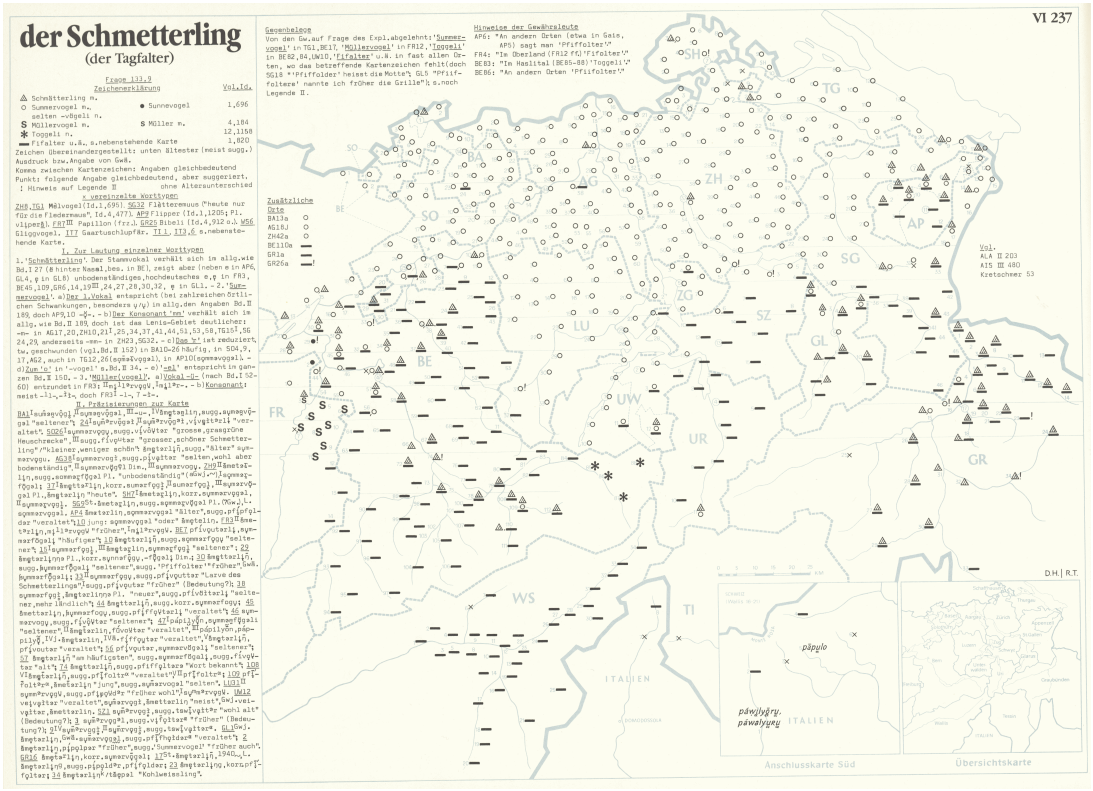


Figure 3.2.: Map SDS 6/237, presenting the dialectal variants of the word 'butterfly' (*Schmetterling* in Standard German).

Volume	*Year*	*Content*	
I	1962	Phonetics	*(Lautgeographie)*
II	1965	Phonetics	*(Lautgeographie)*
III	1975	Morphology	*(Formengeographie)*
IV	1969	Lexicon	*(Wortgeographie)*
V	1983	Lexicon	*(Wortgeographie)*
VI	1988	Lexicon	*(Wortgeographie)*
VII	1991	Lexicon	*(Wortgeographie)*
VIII	1997	Lexicon	*(Wortgeographie)*

Table 3.1.: Publication dates and contents of the eight SDS volumes.

maps on standard photocopiers or scanners.

3.3.2. Contents

The SDS volumes represent information from three major linguistic levels: phonetics and phonology, morphology and morphosyntax, and the lexicon (see Table 3.1).

In dialectology, it is difficult to distinguish between phonetic and phonological phenomena; sound oppositions that have phonological status in one dialect may not in another. For example, Western dialects distinguish open *ì* from closed *i*, so that the words *Ris* [riːs] 'rice' and *Rìs* [rɪːs] 'giant' are distinct. Eastern dialects do not make the distinction between these two vowels, so that the two cited words are homophonous. German dialectology usually employs the term *lautlich* to refer to phonetic as well as phonological phenomena. For simplicity reasons, we will subsume these phenomena under the term *phonetic*.

Volume 1 of the SDS contains 124 maps for vowel quality phenomena. These include *umlaut* phenomena, diphthongization of monophthongs, and monophthongization of diphthongs. Volume 2 contains 83 maps related to vowel quantity phenomena, and 92 maps dealing with consonant changes. Volume 3 of the SDS deals with different aspects of morphology. Nearly 100 maps are dedicated to verbs (regular verbs, auxiliary verbs, modal verbs and short verbs). Determiners, pronouns and numerals take up another 67 maps, while noun inflection and

derivation is explained through 37 maps. 7 maps deal with adjective inflection. 7 further maps are related to some syntactic phenomena, which are now largely superseded by the much more complete SADS dataset (see Section 3.4 below).

The volumes 4-8 of the SDS are dedicated to lexical phenomena and cover the following semantic fields (Trüb 2003, pp. 98-101): the human body and its expressions, kinship terms, function words and particles (volume 4); community and social traditions, clothing, food (volume 5); time expressions, weather, terrain and localization, plants, fruits and vegetables, wild animals (volume 6); containers, housework, buildings (volume 7); livestock and pets, farming and forestry (volume 8). Part of this content deals with everyday vocabulary, while other parts reflect the tools and methods used in agricultural life in the 1940s.

3.3.3. Kleiner Sprachatlas

Some ten years after the end of the SDS project, one has to admit this atlas is not very well known among a large audience. Several reasons may have contributed to this: the long period of publication, the cost of the volumes (more than 1000 francs for the whole collection), the handling difficulties mentioned above, and the detailed conception of the maps and legends, more adapted to an academic public. Therefore, it was decided to edit a reduced, "popular" version of the SDS. This project, called *Kleiner Sprachatlas der deutschen Schweiz* (KSDS) (Christen, Glaser, and Friedli 2010), aims to present 120 mostly lexical maps in a modern graphical design in a single, standard-size paperback book.

The KSDS maps break with the SDS tradition: instead of placing a symbol at each inquiry point, different color surfaces represent the areas of occurrence of the different variants. Hachures represent overlap. Figure 3.3 shows an example.[7]

7 Map from http://ofv.ch/index.php?&ID=kleiner_sprachatlas, accessed 11.4.2011.

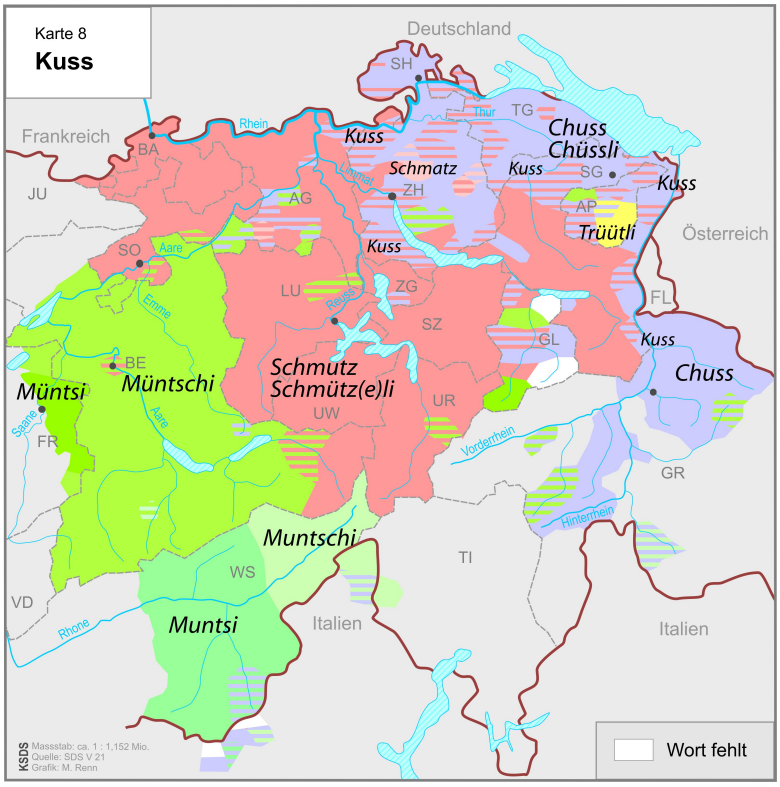


Figure 3.3.: Map designed for the KSDS showing dialectal variants of the word 'kiss'.

3.3.4. Using SDS data for our work

As mentioned above, the SDS consists of 1500 hand-drawn maps. In order to use this material for our work, it has to be digitized and processed. Given the time and workload constraints, it is impossible to obtain a full coverage of SDS data with all the details of the original maps. However, we argue here that full coverage is not necessarily required. This section thus describes the selection criteria, and also raises some potential incompatibilities of the original data presentation and how we resolved them.

Type frequency One important criterion in selecting maps for digitization is their frequency of use when analyzing a general-language dialect text. Phonetic and morphological maps, by definition, have a higher **type frequency** than lexical maps: a phonetic feature may occur in several different words of the lexicon, while lexical maps are restricted to one lexical item and possibly its compounds. For example, phonetic knowledge about word-initial *k* (SDS II/94) allows us to infer the correct pronunciation of words like *Kind* 'child', *Kirche* 'church', *krank* 'sick', *Kerze* 'candle', *kurz* 'short'. Likewise, the third plural verbal ending patterns (SDS III/33) are valid for all regular verbs and will thus occur repeatedly in a corpus. In contrast, the different lexical variants of the word 'potato' (SDS VI/202) are not generalizable to other words, and therefore have a smaller impact on the overall system performance. For this reason, all of our work is biased towards phonetic and morphological phenomena – in other words, towards volumes 1 to 3 of the SDS.

Multiple maps describing the same phenomenon One major difficulty is that we cannot establish a bijective relation between SDS maps and linguistic phenomena. The underlying idea is that every linguistic phenomenon should be represented by one transfer rule that transforms its Standard German realization to different dialectal realizations.

In this section, we deal with the case where several maps represent different distribution patterns of a single phenomenon. This problem mainly occurs in phonetics, where a given phenomenon does not apply in the same manner to different words, even if they contain the same phonetic context. As an extreme example, there are 40 maps dedicated to the phenomenon of vowel

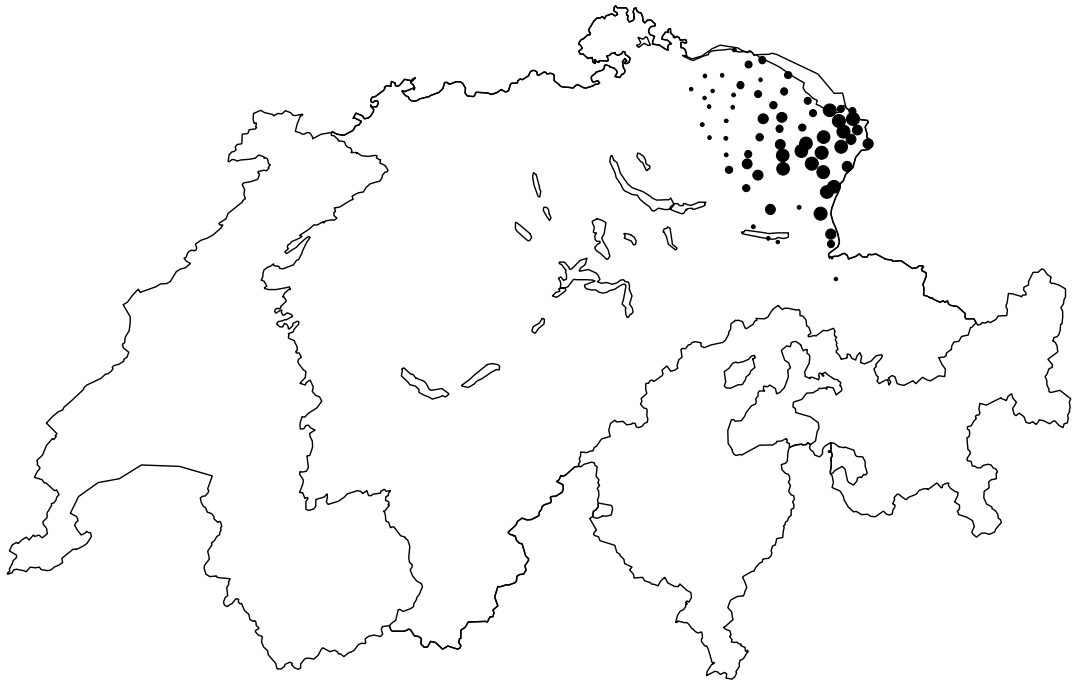


Figure 3.4.: Digital map reconstructed from four different word maps in SDS II/93. The maps represent the shift from *a* to *ä* in the words *Gras* 'grass', *Arbeit* 'work', *sparen* 'to save (money)' and *Axt* 'axe'. The size of the circles represents the number of words in which the shift occurs.

lengthening in an open syllable (SDS II/1 - II/44).

Ideally, one would use all available cartographic material and devise rules that are restricted to the relevant lemmas. Besides the prohibitive workload required for this task, this would miss the important goal of generalization: even if the different maps do not coincide completely, they overlap in some regions and share some distribution patterns. Another solution would be to aggregate all maps of a linguistic phenomenon into one probabilistic map. We pursued this approach for some phenomena where the number of maps was lower than half a dozen and where the additional workload could be justified (see Figure 3.4 for an example). In the other cases, we have chosen a somewhat less rigorous way of selecting maps. We tried to find a subset of "maximally representative" maps. This notion is rather intuitive and relies on the one hand on the number of similar maps, and on the other hand on the perceived type and token frequency of the comparison words mentioned in the legend of the map. With this method, we

reduced the 40 maps of the example given above to a subset of 7 maps. These can be linked to 7 rules that are distinguishable by phonetic context and lexical features.

Maps describing multiple phenomena The opposite phenomenon also occurs. One map may contain information for several independent phenomena. Take as an example map III/72 (see Figure 3.5). It represents several aspects of the plural form of the verb *schlagen* 'to beat':

- The shape of the symbol represents, among others, the number of syllables of the form. This is crucial for morphology: one-syllable stems use the short verb endings, while two-syllable stems use regular verb endings (see Appendix A.4.6, pages 360ff.).
- The color of the symbol indicates if the stem vowel has undergone *umlaut* mutation (yielding *ä, ö, e*) or not (yielding *a, o*).
- The dotted surface represents a general phonetic phenomenon called *Entrundung* (see Appendix A.2.1, page 298), which further transforms the forms.
- The green isogloss lines represent different inflection paradigms for the three persons of plural (see Appendix A.4.6, pages 357ff.).

These four phenomena are considered separate and yield four independent transfer rules.

Precision of the original maps We have chosen to use the broad Dieth spelling conventions (see Section 3.2. We have argued that not all phonetic nuances need to be covered when working with written dialect data. In contrast, some phonetic maps of the SDS are based on the fine-grained transcription system of the original explorer notes. The editors recognize that this approach risks creating what they call *Exploratorenlandschaften* (fieldworkers' landscapes):[8] if different fieldworkers perceive the same sound differently, the final map represents the distribution of fieldworkers rather than the distribution of objective phonetic differences.

8 For example, in the legend of SDS II/161.

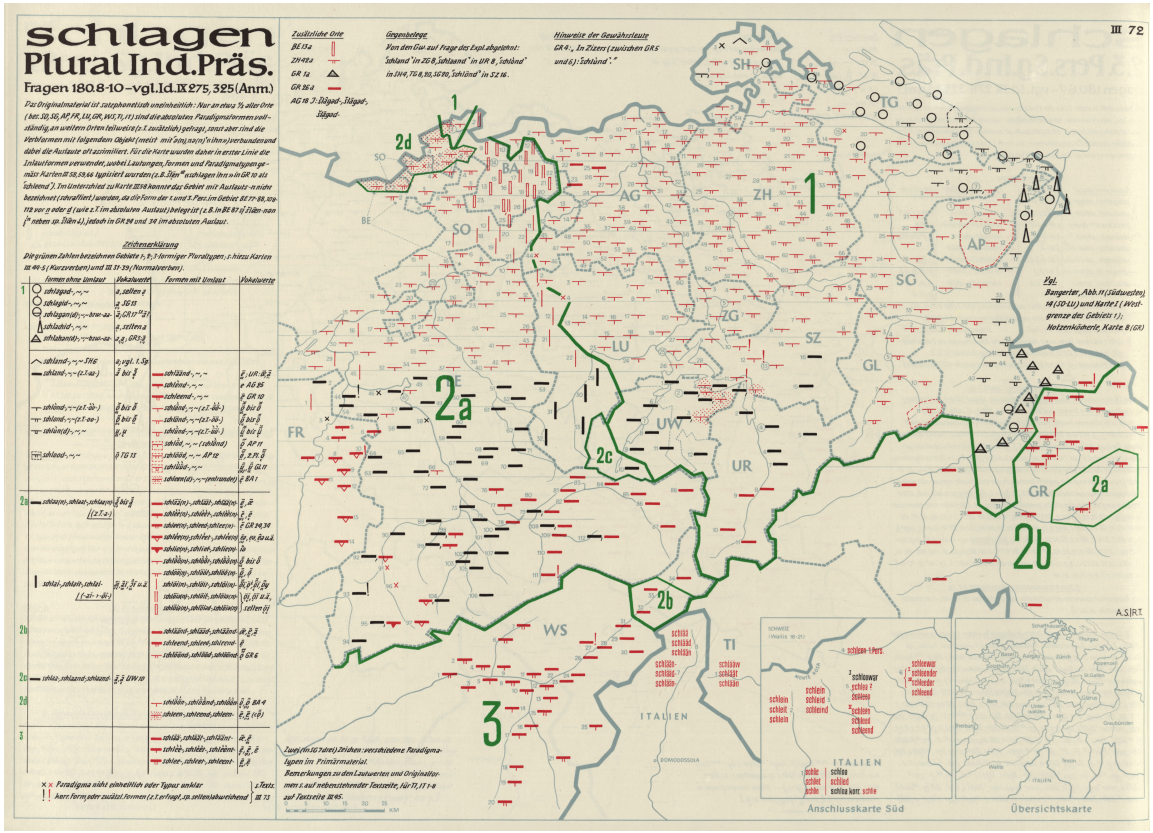


Figure 3.5.: Map of the inflected plural forms of the verb *schlagen* 'to beat' (SDS III/72).

Figure 3.6.: Legend for map SDS I/50. Seven symbols (the rightmost two symbols are combined with a hachured background) are used to represent phonetic differences in the *u/o* sound of the Swiss German word *Chuchi* 'kitchen'.

As a result, some phonetic maps of the SDS are too detailed for our purposes. We chose to collapse several phonetic variants into a single Dieth grapheme. For example, map I/050 distinguishes seven vowel qualities (see Figure 3.6). We interpreted the variants symbolized with one bar as *u*, and the variants symbolized with two bars as *o*. The choice of the cutoff line was set based on the native speaker's intuition. We concede that this is problematic, but we could not find a straightforward, more objective solution. The least we could do is to document these choices (see Appendix B).

Lexical maps sometimes show the opposite phenomenon: the editors may have chosen to merge several slightly different pronunciations and use a single symbol. Map SDS V/112 shows the most common greeting terms. It combines the variants *grüessi, grüessdi, grüessech* 'hello' into a single symbol, and the variants *grüezi, grüezdi, grüezech* into another symbol. The underlying thought was that the map should represent the distribution of the verb variants *grüez-* vs. *grüess-*, but not the form of the pronominal clitic (*-di, -i, -ech*). However, the greeting term is completely lexicalized, and the surface form seems more important to us than the underlying decomposition. Therefore, we wanted to reflect the form of the clitic in our map. Fortunately, in this case, the legend provides more detailed information about the clitic distribution, even if this complicates the digitization process.

Diachronic evolution As explained above, the data collection for the SDS took place more than 50 years ago. It can thus be expected that the dialects have changed since then, and that at least some of the maps do not reflect present-day dialect features. Two comparative inquiries may shed some light on the diachronic evolution that the Swiss German dialects have undergone during the last 50 years.

Christen (1998b) compares contemporary verbal morphology with SDS data. Her work shows various phenomena of dialect convergence, but this convergence is all but directed to one specific dialect. Moreover, the variant towards which the dialects converge is often not the one that most resembles Standard German. My personal intuition confirms this tendency. As an example, the third person singular conjunctive form of the verb *to be* is *seg* in my idiolect, and I have observed that this form is very common across Eastern Switzerland. However, the corresponding SDS map (III/51) shows this form only in a single inquiry point in Aargau, more than hundred

kilometers away from my hometown.

On the lexical level, common knowledge is that particular Swiss German lexical items disappear in favor of phonetically adapted variants of the Standard German terms (Christen 1998a, pp. 59-60). An online survey conducted in 2008 by researchers from Zürich University confirmed this hypothesis.[9] Dialect speakers all over German-speaking Switzerland were asked to indicate their preferred dialect word for eighteen frequent lexical items for which SDS data were available.

The results for the Standard German word *Butter* 'butter' are characteristic. In the corresponding SDS map (V/179), the variant *Anke* was the unique answer in all but the Northeastern parts of Switzerland; in the latter region, *Schmalz* was used instead. The variant *Butter* was used in very few data points. Half a century later, *Schmalz* has nearly disappeared; the area of *Anke* has receded; the Standard German borrowing *Butter* is the default variant in the Northeast, but is also used as a common alternative in former *Anke* regions.

Similar results are reported for a characteristic item of Swiss cuisine: in the SDS map (V/197), the term *Röschti* was limited to Bern; in many regions, a descriptive noun phrase like *praate Härdöpfel* 'roasted potatoes' or *gchochet Härdöpfel* 'cooked potatoes' was used instead. At present, the simplex term *Röschti* is used all over Switzerland, except in regions where other simplex terms had been available before (*Brägel* in FR, *Brausi* or *Bräusi* in AG and LU).

Despite these evident cases of diachronic change, many SDS maps still seem to be valid without major modifications, especially on the phonetic level. In the absence of a systematic collection of recent data, there is not much we can do about outdated material. One strategy is to sort out SDS maps that do not correspond to the perceived reality anymore. Another strategy would be to simulate the hypothesized mechanisms of dialect change in the map interpolation process (the latter is described in Section 4.6.2). For example, variants occurring in regional capitals might have a higher chance of survival than small-scale variants occurring in socio-economically less important regions.

9 Results are available on http://www.ds.uzh.ch/Forschung/Projekte/Schweizer_Dialekte/.

Object loss A related problem that occurs with lexical SDS entries is object loss (*Sachverlust*). Large parts of the lexical section deal with the tools and methods used in agricultural life in the 1940s. Some of these tools have partially or completely disappeared since (e.g. different types of "wooden backpacks", SDS VII/41-72), and the words disappeared with them. Likewise, the invention of the washing machine and of the refrigerator made several washing and food preservation techniques obsolete (SDS V/183-186, VII/22-29, VII/87-89). As a result, several maps are completely obscure to me, because I neither understand the referent (as described in the legend of the map) nor the dialectal word variants. Such maps are discarded from further analysis.

Criteria for lexical map selection We have already stated above that we prefer phonetic and morphological maps due to their inherently higher type frequency. However, lexical variation is still important in present-day dialects, and some lexical maps of the SDS are still worth integrating into our system. In the following, we list some criteria that are used to select the relevant lexical maps:

- The lexical item should have a high token frequency.[10] Function words like adverbs and prepositions have higher token frequency than lexical words like nouns and verbs and should therefore be preferred. Moreover, function words often show an idiosyncratic behavior that makes them difficult to model with phonetic transformations.

- The concept that the lexical item denotes is still known and in current use.

- The concept is expressed by a simplex lexical item in Standard German. This criterion is important for translation: there is no use to implement a lexical item that will never occur because there is no equivalent word in Standard German from which it could be translated.

- The mere phonetic adaptation of the Standard German lexical item is perceived as infelicitous by native dialect speakers.

10 Frequency information about words may be collected from word lists like Ruoff (1981), which contains data of spoken Southern German.

- There is a complete SDS map for that lexical item.

3.3.5. Summary

This section presented one data source in detail: the *Sprachatlas der deutschen Schweiz*. It discussed its history, its contents, and the problems that arise when using this material for our work. Currently, we have selected 59 maps for phonetic phenomena, 120 maps for morphological phenomena, and 27 maps for lexical phenomena.

3.4. *Syntaktischer Atlas der deutschen Schweiz*

3.4.1. Methodology

While the SDS data draws a fairly complete picture of the dialectal distribution of phonetic, morphological and lexical variants, it contains few information about syntactic properties. Although the existence of syntactic variation is acknowledged by the SDS editors, only about 10 maps in volume III deal with syntax. The reason of this lack is mainly to be found in the difficulties of eliciting specific syntactic structures in the interviews with the informants.

Dialectological interviews are usually organized according to semantic fields. This allows the fieldworker to easily elicit a set of carefully chosen semantically related words and expressions. These provide the raw material for phonetic, morphological and lexical variation patterns. With this strategy, all types of variation are ultimately grounded in the words and expressions linked to a semantic field. This is difficult to achieve with syntactic constructions, which occur much more freely.

Arguing that syntactic differences among Swiss German dialects are existent and relevant, that syntactic phenomena have been dealt with in other recent atlas projects, and that the field of microsyntax (i.e. the small-scale variation of syntactic parameters) has become popular in recent years, a research project to systematically collect syntactic material and to publish its results in

atlas form was started in 2000 under the direction of Prof. Elvira Glaser at Zürich University. The project became to be known as *Syntaktischer Atlas der deutschen Schweiz*, abbreviated SADS (Bucheli and Glaser 2002). Its organization differs in several respects from the earlier SDS project in order to take into account the specificities of syntax and the technological advances in data processing. Several innovations are worth mentioning.

Written questionnaires The SADS editors distribute written questionnaires instead of sending fieldworkers out to perform oral interviews. This obviously allows to gather the linguistic material much faster (all informants can fill out their forms at the same time) and cheaper (no transportation and accommodation expenses for fieldworkers). Informants may also find it easier to reflect more thoroughly on the proposed answers when they are not under time pressure.

On the downside, the editors have less control and can provide less guidance on the informants' work. Some 30% of potential informants did not send back the questionnaires. About 5% of the informants were unable to fill out the questionnaire in a useful manner (see Bucheli and Glaser 2002). The use of written questionnaires thus requires more thorough post-processing in order to filter out unusable responses.

Several informants per inquiry point The drawbacks of the written questionnaires call for a higher degree of redundancy. Instead of asking one or two persons per inquiry point as in the SDS, the SADS editors opted for a larger number of informants, with an average of 7 informants per point. The SADS database lists 1331 female informants (42%) and 1854 male informants (58%).[11] Hence, if one informant fails to give the expected answer, another one will provide it.

Even if the handling of these additional data is facilitated by the use of written questionnaires and electronic databases, it was decided to cut down the density of the inquiry point network. While the SDS relied on nearly 600 points, the SADS only uses 344 points.

11 Figures as of 18.11.2010.

Several batches of questionnaires Written inquiries are less spontaneous and interactive than oral interviews: a fieldworker can dig deeper if something seems interesting or unclear. In order to obtain a minimal level of interactivity with the SADS informants, it was decided to split the questionnaire into several batches. In this way, a subsequent batch could be adapted in function of the insights obtained from the previous batch – clarification questions could be added, additional solutions could be proposed. Splitting the questionnaire probably also improved the alertness of the informants. One batch is designed to be completed in about 45 minutes. Filling out the whole questionnaire at once would have required several hours. One is easily convinced that this would have put off a lot of informants in the first place. As said above, this proceeding also allowed to adapt the list of informants. If an informant handed in an unsatisfactory first batch, he was not sent any further batches.

Various tasks Three different tasks are proposed to elicit syntactic information. In all tasks, the questions are embedded in a little story that sets the context and is supposed to facilitate the understanding.

One question type is **translation**. Here, the informant is asked to translate a Standard German sentence into his dialect. This task gives the informant a lot of liberty and few constraints; in this sense, it is the "least influenced and most spontaneous form of any indirect question type" (Bucheli and Glaser 2002, p. 61). However, informants may mirror their answer on the Standard German source or produce an unintended answer (for example, by completely paraphrasing the sentence so that the intended syntactic structure does not appear).

The second task is **sentence completion**. By providing the beginning of a sentence, the range of possible answers is more constrained than in the translation task. Nevertheless, unintended answers may still appear.

The third task consists of **check box marking**. In this case, a set of potential dialect answers are provided, and the informant has to check all answers that are acceptable in his dialect (we call this subtask **multiple choice question**). Moreover, if the informant accepts several answers, he has to indicate which one sounds the most natural to him (we coin this subtask **single choice question**). This task allows to collect two degrees of grammaticality: acceptance and preference (see also Figure 4.5 on page 120). On the other hand, it may unduly restrict the choices of the

informant: had he been asked for a spontaneous answer, he might have used a different variant that was not listed. In this case, the questionnaire authors provide an additional field where the informant can add another variant that corresponds more closely to his intuitions. To sum up, this task mostly avoids unintended answers, but still manages to obtain precise distinctions.

Most syntactic phenomena were investigated through several tasks, sometimes in different batches. This allowed the editors to compare the effects of the task type on the same phenomenon, sometimes at the risk of obtaining contradictory answers.

3.4.2. Contents

In comparison with other linguistic levels, syntactic differences between Standard German and Swiss German dialects are rather rare. Some of these differences are representative of the mainly spoken use of Swiss German. They do not show much inter-dialectal variation, and they are also encountered in other spoken varieties of German. Other differences are dialectological in nature, in the sense that they are specific to some subgroups of Swiss German dialects and usually do not occur outside of the Alemannic dialect group. This second type of differences constitutes the main research object of the SADS project.

While there has not been a systematic collection of syntactic differences among Swiss German dialects before the SADS, some interesting cases of syntactic variation have been mentioned in dialect grammars and dialect syntax monographs. In addition, some variation patterns are immediately perceptible by native speakers and are part of common dialect knowledge. However, there precise geographic extension of these phenomena has largely remained unknown. It is thus on the basis of these sources that the questionnaire of the SADS survey was established. As mentioned above, preceding batches of the questionnaire could unearth previously unknown variants, which could be investigated in more detail later on.

3.4.3. Using SADS data in our work

The four batches of the questionnaire contain about 120 questions. For practical reasons, not all of them could be considered for our work. A subset of SADS maps was chosen according to the following – rather technical – selection criteria:

- The phenomenon shows significant variation over the Swiss German dialect area. Extreme relict phenomena occurring only in a small number of alpine dialects were not deemed relevant enough for our purposes.

- The phenomenon is triggered by well-defined syntactic patterns that are identifiable in a (manually or automatically) annotated Standard German text. For example, a Standard German treebank like TIGER (S. Brants et al. 2002) does not distinguish between copredicative (4a) and adverbial (4b) uses of adjectives (see Appendix A.4.4, page 352). Therefore, it is impossible to perform the necessary dialectal transformations.[12]

  (4) a. *Blitzblank* hängen die Töpfe an der Küchenwand.
  'The pots are hanging sparkling clean on the kitchen wall.'
  b. *Häufig* hängen die Töpfe an der Küchenwand.
  'The pots frequently hang on the kitchen wall.'

- The syntactic description of the Swiss German variants is clear and allows the implementation of a transfer rule. For instance, the replacement of inflected verb forms by a *tun* + infinitive periphrase seems to depend on multiple morpho-phonological and pragmatic factors whose interactions are not clear. Thus, the transfer rule for this phenomenon cannot be operationalized.

- For a given phenomenon, a single map is extracted,[13] preferring the ones from the first

12 The treebank annotations could be augmented with the help of semantic heuristics. For the example at hand, a semantic resource could easily tell that pots can be sparkling clean but not frequent. However, we did not pursue such investigations any further.

13 However, a phenomenon may be linked to a specific context. For example, prepositional dative marking works differently depending on the type of dative object (resumptive pronoun, wh-pronoun or full noun phrase). These three contexts count as three phenomena and are associated with three maps.

Question	*Phenomenon*
I/2	Prepositional dative marking
I/7	Prepositional dative marking
I/9	Verb clusters in main clauses
I/10	Article doubling/movement
I/11	Infinitival clause complementizer
I/14	Cliticization of feminine articles
I/19	Auxiliary verb clusters in subordinate clauses
I/20	Prepositional dative marking
II/6	Modal verb clusters in subordinate clauses
II/10	Article doubling/movement
II/18	Relative resumption of dative noun phrases
II/26	Double marking of wh-complementizer
II/28	Relative resumption of prepositional phrases
II/30	Proper noun as genitive attribute
II/32	Articles with proper nouns
III/7	Clitic pronoun order
IV/8	Clitic pronoun order

Table 3.2.: SADS maps selected for further analysis.

batch. This decision might be controversial: the SADS authors view the first batch as rather experimental and as a mere basis for refined follow-up questions. However, the data of the first batch give a rather good overview of the phenomena without delving into marginal contexts and variants that are less relevant for our purposes.

We chose seventeen questionnaire items that satisfied these criteria. They are listed in Table 3.2.

4. Maps

4.1. Introduction

The main part of our thesis consists of the implementation of a set of transfer rules that modify Standard German words and sentences to produce different dialectal forms. These rules are linked to maps that define their geographical area of validity. Chapter 3 presented the linguistic data on which they are based. Before we discuss their computational implementation in Chapter 5, we would like to detail the processes required for acquiring the relevant cartographic information. In particular, we present the workflow of transforming SDS and SADS raw data into electronic maps so that the transfer rules can directly access them. Most of this workflow involves a cartographic software called *Geographical Information System*. Over the course of this chapter, we will repeatedly refer to the dialectological distinction between **variables** and **variants**, defined in the introduction (Section 1.2.2).

4.2. Geographical Information Systems

A Geographical Information System (GIS) is a particular piece of software used for the creation and processing of geographical data. According to Burrough and McDonnell (1998, p. 11), a GIS can be defined as "a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes". A GIS handles geographical (spatial) objects in terms of "(a) their **position** with respect to a known coordinate system, (b) their **attributes** that are unrelated to position (such as colour, cost, pH, incidence

of disease, etc.) and (c) their spatial interrelations with each other which describe how they are linked together (this is known as **topology** and describes space and spacial properties such as connectivity […])" (Burrough and McDonnell 1998, pp. 11-12, emphasis mine). The geographical data are assembled in maps: a **map** consists of a set of spatial objects.

There are two fundamentally different types of maps: **vector** and **raster** maps. In the vector model, the spatial objects are defined as **discrete entities** with particular shapes and sizes. Three types of objects are supported: points, lines and polygons. For instance, water wells may be represented as points, rivers as lines and lakes as polygons. As explained above, these objects have spatial attributes that determine their position, size and topology, while non-spatial attributes may hold additional information like average water temperature, pollution level, etc.

Some phenomena are continuous in nature (like altitude or atmospheric pressure) and therefore difficult to represent with the discrete entities required by the vector model. The raster model (also called **continuous fields** model[1]) is more adequate for these data. A raster map consists of a regular grid of cells (called **pixels**) which all have the same size and shape. The cell size depends on the resolution of the data source, rather than on the actual size of the objects (which usually cannot be determined anyway). In the raster model, each cell may contain a single non-spatial attribute value.

Current geographical information systems handle vector as well as raster data and perform the following actions (Burrough and McDonnell 1998, p. 29):

- import spatial data in different formats,
- select and classify objects on the basis of their spatial or non-spatial attributes,
- create new entities by geometrical union or intersection of existing entities,
- derive new attributes from existing ones by mathematical operations,
- perform spatial analysis on vector data (e.g. surface or distance calculations),
- perform spatial analysis on raster data using linear algebra,
- provide digitization tools to convert printed maps into digital ones,
- convert vector to raster data by interpolation,

1 Equating continuous fields with rasters is not quite exact: as will be explained in Section 4.6.1, continuous fields can be built with a vector-based approach (TINs, Thiessen polygons).

- export spatial data in computer-readable and human-readable formats (providing large choices of symbology for the latter),
- provide programming interfaces to automate processes with scripts.

We use a GIS to import data from our atlas sources and transform them into maps suitable for the transfer rules. All of our work has been carried out using *ESRI ArcGIS 9.3*,[2] but other systems provide similar functionality.

4.3. Digitizing SDS maps

As said above, the SDS maps are designed by hand on the basis of the fieldworkers' notes. At every point of the inquiry network, the map shows one (or more) symbol(s), according to the variant(s) observed at that point. In cartographic terms, the SDS maps consist of discrete entities in point form. Figure 4.1 shows an example of an SDS map. We will use this example throughout this section to illustrate the different processing steps.

Scanning At the beginning, the printed maps have to be scanned, so that the resulting TIFF image file can be imported into the GIS for further processing. The scanning process is complicated by the fact that the maps are too large to fit in a standard size (A3) scanner. Some of the maps were photographed due to time constraints; the resulting images remained usable despite their lower quality.

Georeferencing Once the scanned file is loaded into the GIS, it has to be associated with a coordinate system. We use the *SwissGrid* coordinate system (CH1903+) in our work.[3] Georeferencing works as follows: four Swiss cities (see Table 4.1) have been selected as anchor points.

2 http://www.esri.com/software/arcgis/arcgis-for-desktop/index.html, accessed 8.8.2011.

3 *SwissGrid* is the official Swiss coordinate system. It consists of a six-digit easting value and a six-digit northing value, whose value ranges do not overlap. One unit corresponds to 1 m. See http://www.swisstopo.admin.ch/internet/swisstopo/de/home/topics/survey/sys/refsys/switzerland.html, accessed 9.8.2011.

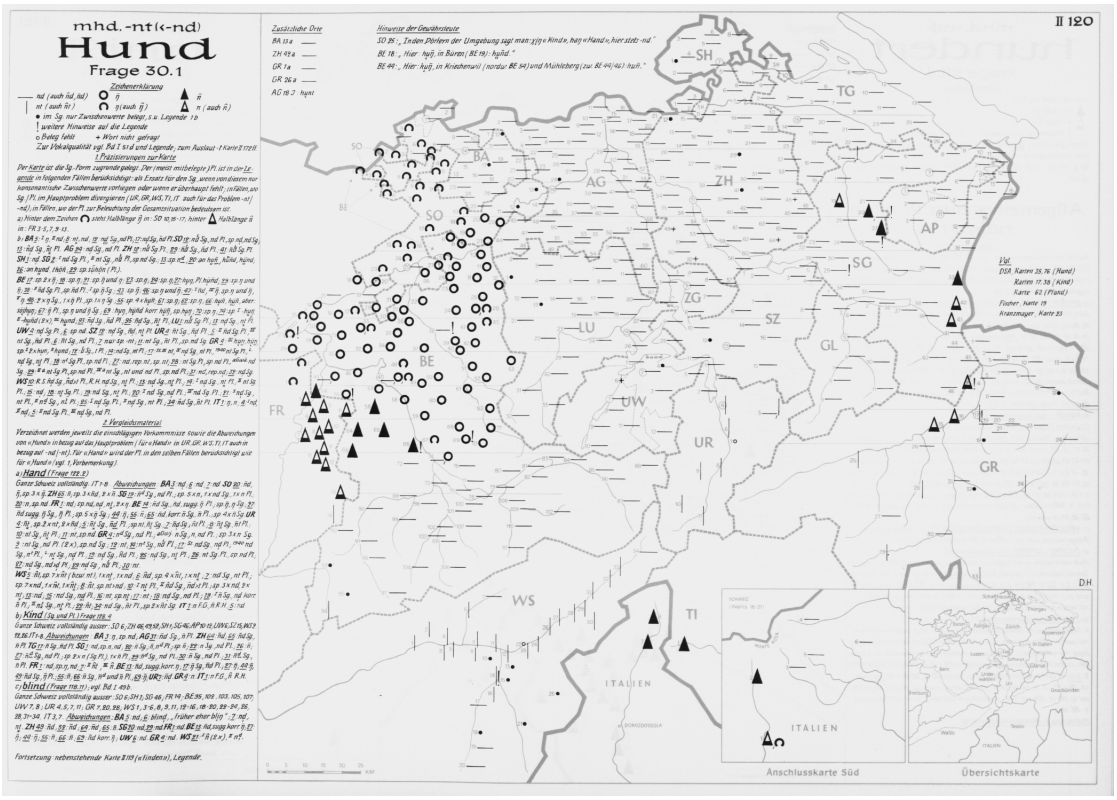


Figure 4.1.: Original SDS map for the transformation of word-final *-nd* (SDS II/120). The map contains four major linguistic variants, symbolized by horizontal lines (*-nd*), vertical lines (*-nt*), circles (*-ng*), and triangles (*-nn*) respectively. Minor linguistic variants are symbolized by different types of circles and triangles.

City	Coordinates	City	Coordinates
Basel	611 300 / 267 600	St. Gallen	746 200 / 254 600
Bern	600 100 / 199 700	Chur	759 300 / 191 100

Table 4.1.: Four Swiss cities used for georeferencing. These cities are easily identifiable on the scanned maps and are relatively far apart, achieving higher overall georeferencing precision.

They are visually identified on the map and associated with their coordinate pairs. Once this is done, all further objects created on the map are defined in terms of *SwissGrid* coordinates.[4]

Digitizing In this process, a new map layer is created for every dialectal variant occurring on the scanned map. These layers are manually populated with data points by following the visual information given by the underlying scanned map and clicking on each point on the map where the symbol occurs. Mouse clicking is guided by a "magnetic" grid of predefined locations (corresponding to the SDS inquiry points). See Figure 4.2 for a screenshot of the digitizing process. The digitization process is repeated for each dialectal variant. At the end, the different layers can be displayed simultaneously, as shown in Figure 4.3.

Recall from Section 3.3.4 that some simplifying assumptions were made to speed up the process. Variants that are used in less than ten inquiry points are omitted – many of these small-scale variants have most likely disappeared since the data collection in the 1940s. We also collapse minor phonetic variants which cannot be distinguished in the Dieth spelling system. For example, in Figure 4.2, the open and closed circles represent minor phonetic variants of the *ng* pronunciation; we treat both of them on the same layer (red dots).

Layer merging The digitization process yields one layer for each dialectal variant. The different layers are then automatically merged into a single layer with a *Python* script that interacts with *ArcGIS*. The table of non-spatial attributes of the resulting layer contains the relevant information

4 A set of SDS maps have also been scanned and georeferenced for the *DiWA* project (http://www.diwa.info/DiWA/atlas.aspx, accessed 8.11.2011). However, these maps were made available online only after our own georeferencing work was completed.

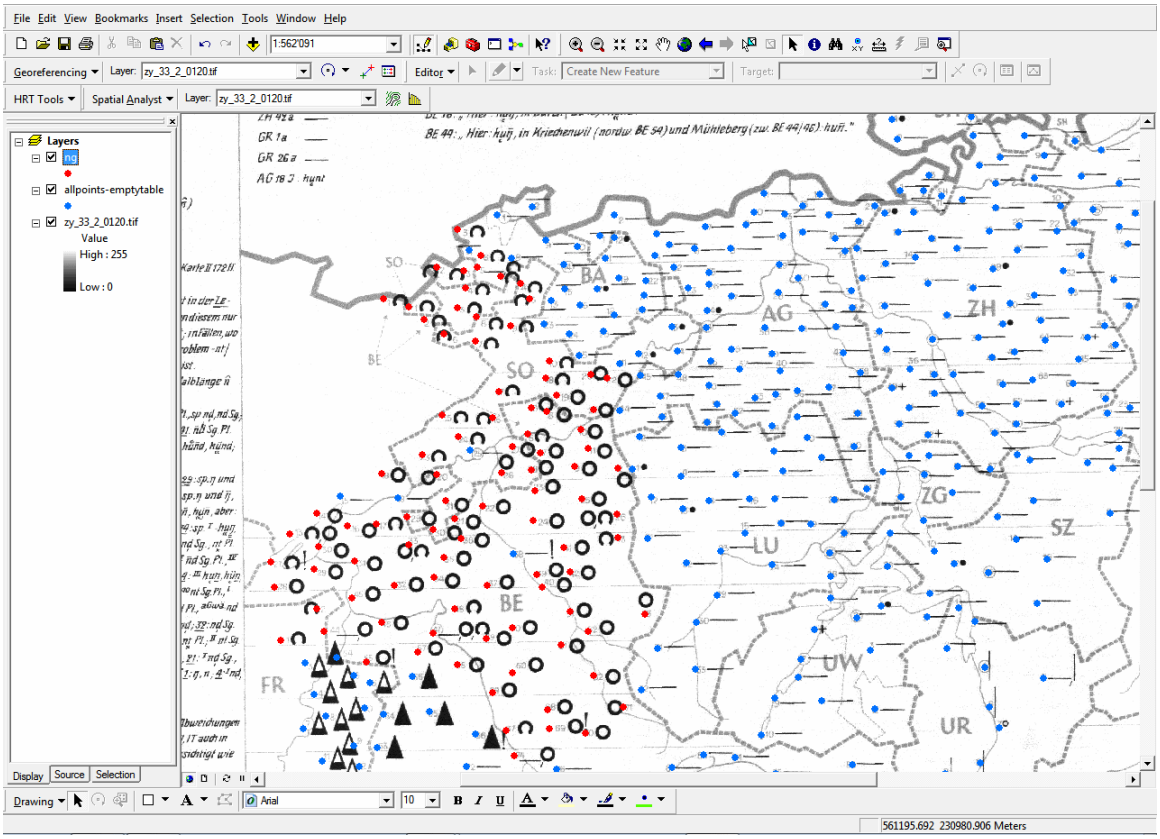


Figure 4.2.: Screenshot of the digitization process. The scanned map constitutes the background layer which is used to guide the digitization. The blue dots represent the predefined set of inquiry points; these are the only locations where adding points is permitted. The red dots are crucial information created here: they represent the inquiry points where the *ng*-variant is used. They are created by clicking on the screen next to all circles in the scanned map. The "magnetic" blue dots ensure that all red dots are placed directly upon them.

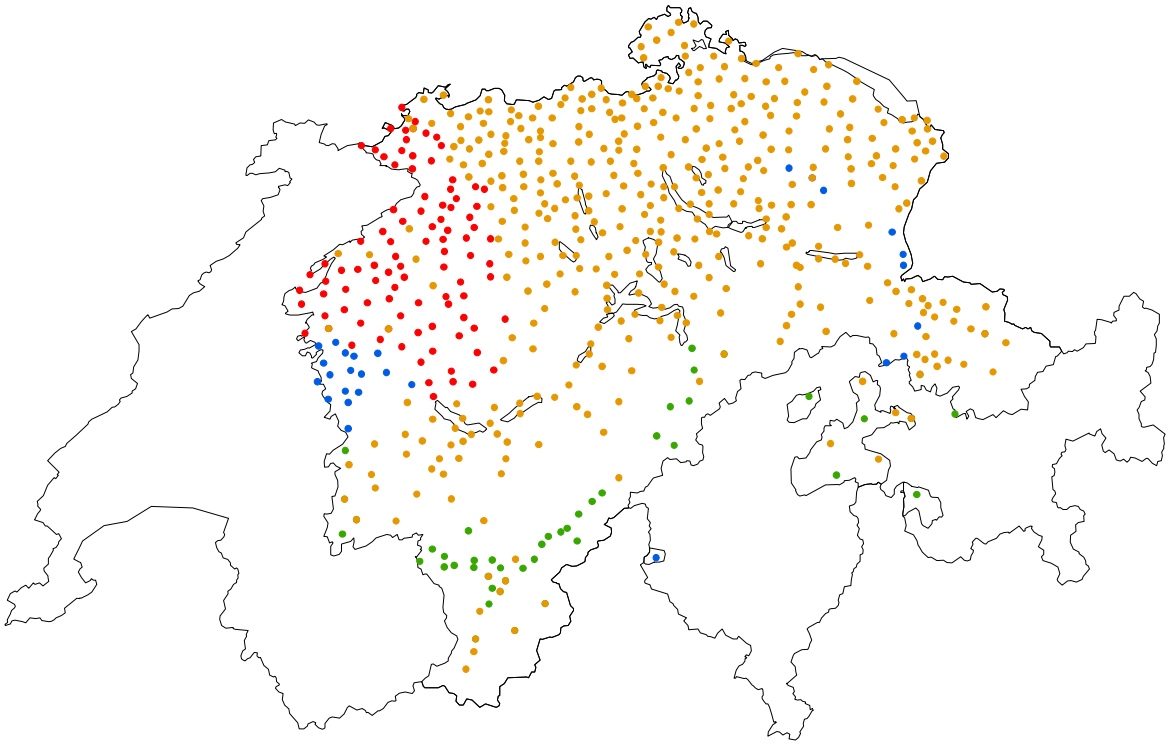
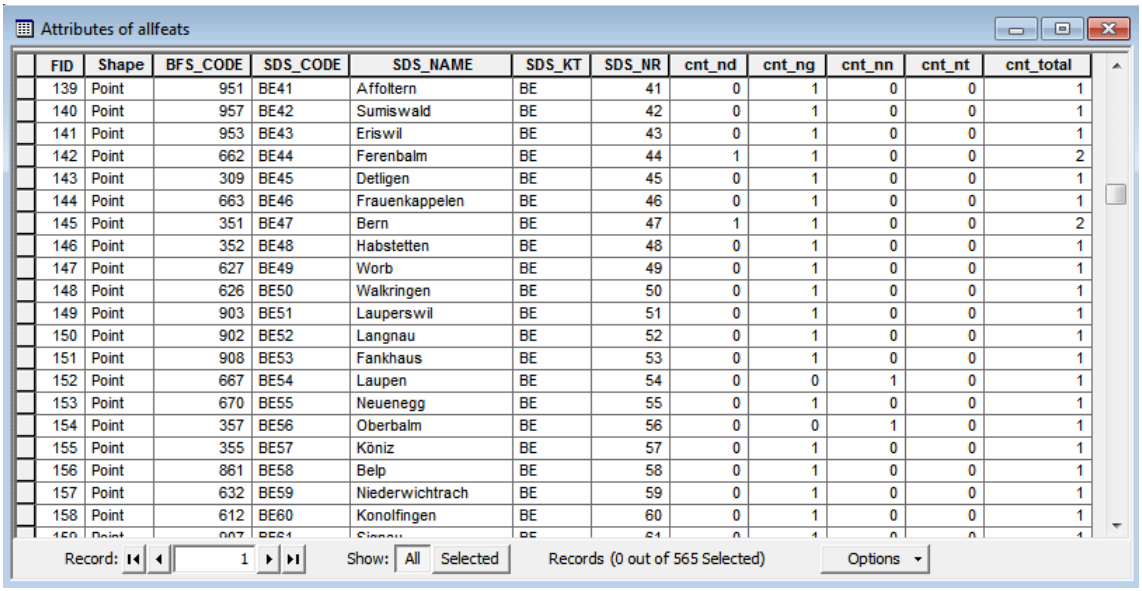


Figure 4.3.: The result of digitizing the four major dialectal variants of the *nd*-map. *nd*-variants are represented in yellow, *ng*-variants in red, *nn*-variants in blue, and *nt*-variants in green.



FID	Shape	BFS_CODE	SDS_CODE	SDS_NAME	SDS_KT	SDS_NR	cnt_nd	cnt_ng	cnt_nn	cnt_nt	cnt_total
139	Point	951	BE41	Affoltern	BE	41	0	1	0	0	1
140	Point	957	BE42	Sumiswald	BE	42	0	1	0	0	1
141	Point	953	BE43	Eriswil	BE	43	0	1	0	0	1
142	Point	662	BE44	Ferenbalm	BE	44	1	1	0	0	2
143	Point	309	BE45	Detligen	BE	45	0	1	0	0	1
144	Point	663	BE46	Frauenkappelen	BE	46	0	1	0	0	1
145	Point	351	BE47	Bern	BE	47	1	1	0	0	2
146	Point	352	BE48	Habstetten	BE	48	0	1	0	0	1
147	Point	627	BE49	Worb	BE	49	0	1	0	0	1
148	Point	626	BE50	Walkringen	BE	50	0	1	0	0	1
149	Point	903	BE51	Lauperswil	BE	51	0	1	0	0	1
150	Point	902	BE52	Langnau	BE	52	0	1	0	0	1
151	Point	908	BE53	Fankhaus	BE	53	0	1	0	0	1
152	Point	667	BE54	Laupen	BE	54	0	0	1	0	1
153	Point	670	BE55	Neuenegg	BE	55	0	1	0	0	1
154	Point	357	BE56	Oberbalm	BE	56	0	0	1	0	1
155	Point	355	BE57	Köniz	BE	57	0	1	0	0	1
156	Point	861	BE58	Belp	BE	58	0	1	0	0	1
157	Point	632	BE59	Niederwichtrach	BE	59	0	1	0	0	1
158	Point	612	BE60	Konolfingen	BE	60	0	1	0	0	1

Figure 4.4.: Screenshot of the attribute table. Each line is linked to one point on the map. The attribute table provides non-spatial information like the name of the inquiry point and the variants observed at that point. Presence or absence of a dialectal variant is indicated as 1 or 0 in the corresponding column.

about the dialectal variants. It is augmented with other information like the village name and corresponding administrative codes. Figure 4.4 shows an excerpt of the attribute table.

To sum up, the printed SDS maps have to be redrawn in digital format. This redrawing consists of several steps, of which the digitization step is the most time-consuming one. Depending on the complexity of the phenomenon, digitizing an SDS map takes between one and two hours. The result of this process is a single file consisting of a set of data points. Each point is characterized by spatial attributes (its coordinates in the *SwissGrid* coordinate system) and non-spatial attributes (the village name and code, as well as the counts of the observed dialectal variants). The resulting maps are visually checked in order to avoid obvious digitization errors. Nevertheless, digitization is a manual process through which errors and omissions may be introduced.

4.4. Importing SADS data

The SADS data is easier to process. Pending its publication in cartographic form, the underlying database is accessible online. Creating the relevant map files thus amounts to acquiring the corresponding database tables and spatially referencing the entries of the tables.

Acquiring the data The data are first downloaded in the form of HTML tables, as they are available online. Then, the tables are converted into a file format that is readable by *ArcGIS*. These conversions yield tables that relate inquiry points (represented by the name and code of the village) with the number of informants using the different variants.

Georeferencing the data The data table obtained in the previous step contains the relevant non-spatial attributes. However, it is not yet related to points on the map. This is the goal of this step. For each line of the data table, the coordinate pair of the inquiry point is looked up in a separate file, and a point is drawn at its location on the map. In contrast to georeferencing SDS data, this can be done entirely automatically.

Acceptance and preference As mentioned in Section 3.4.1, some items in the SADS survey contain two questions: one about acceptance of a given dialectal variant, and another one about the preference of that variant over the others. Both datasets are collected and mapped. However, the differences are rather small (see Figure 4.5 for an example). Acceptance maps generally have higher probabilities in transition zones than preference maps.

In contrast to the SDS maps, the SADS data is available in electronic format. Hence, the conversion can be fully automated and does not require manual digitizing.

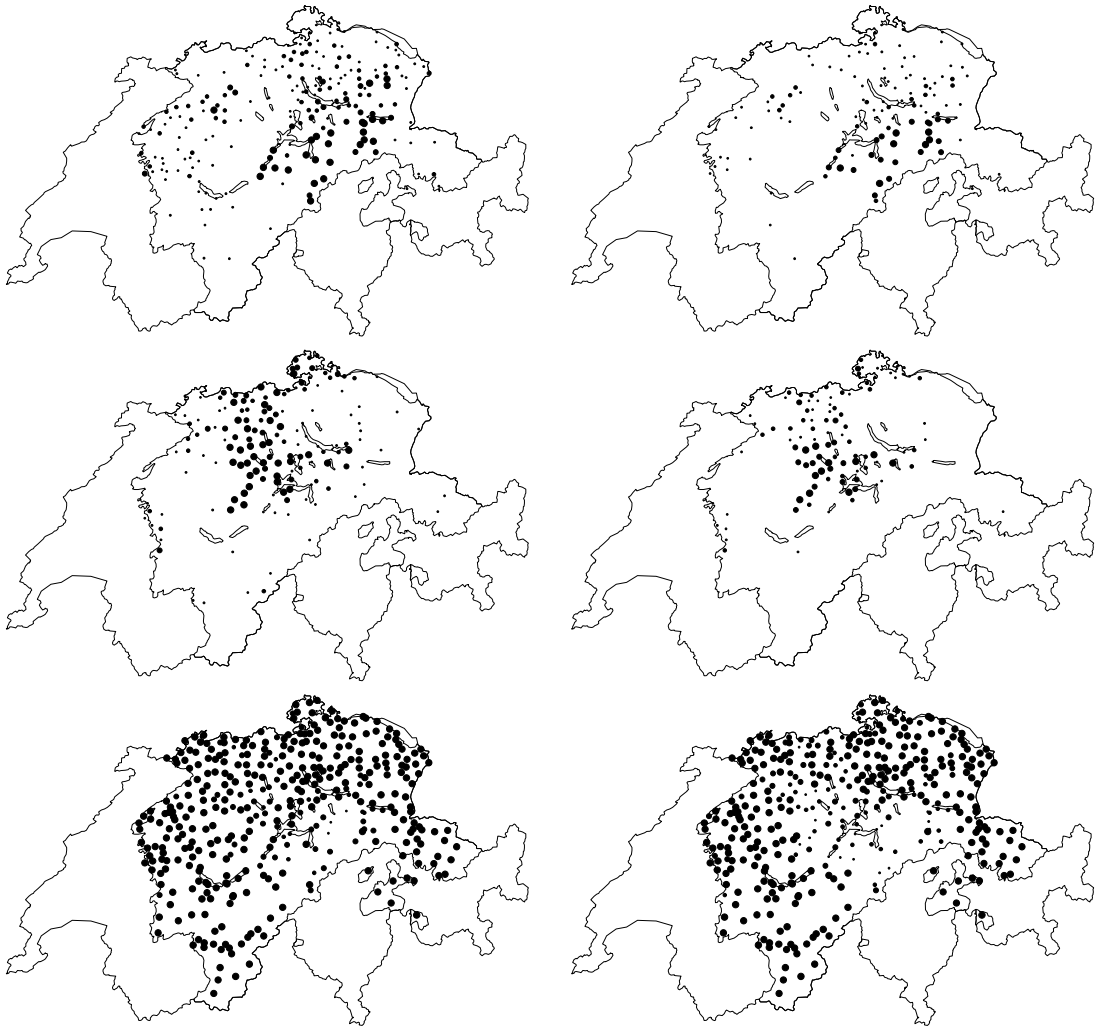


Figure 4.5.: These maps show the three variants of prepositional dative marking (see page 329; maps extracted from SADS I/7): *a*-marking in the top row, *i*-marking in the central row, and non-prepositional marking in the bottom row. The maps on the left are drawn on the basis of acceptance judgements, while the maps on the right are based on preference judgements. The size of the dots represents the proportion of informants approving the variant.

4.5. Dialectal variation: discrete or continuous?

In Section 4.2, we have insisted on two fundamentally different ways to conceive maps: the discrete entities model and the continuous fields model. The maps created from SDS or SADS sources are based on discrete entities in the form of points. Each point represents a village of the inquiry network and the dialectal variants observed at that place. Here, we argue that dialectal variation is more adequately rendered with the continuous fields model. In the following sections, we present different methods of transforming the discrete entities maps into continuous fields maps.

Throughout the history of dialectology, there has been some debate about the nature of dialectal variation (see page 55). Dialectal variation can either be discrete and follow clear-cut borders, or it can be continuous with relatively large areas of gradual change. There is evidence for both types of variation. We start with some arguments in favor of discrete dialectal variation.

- Many dialectal phenomena, especially lexical ones, are discrete in nature. When talking about a kiss (see Figure 3.3 on page 96), one must choose one of the lexemes *Kuss, Schmutz* or *Müntschi*. There is no intermediate form that could be used in a transition area.

- In many areas of German-speaking Switzerland, the settlements are not continuous; villages and towns are separated by uninhabited zones. Boundaries between linguistic variants naturally occur in uninhabited zones rather than cutting through the middle of villages. Moreover, a discrete account of dialectal variation allows for blank areas, which avoids the nonsensical task of defining the dialects of mountains.

- Dialectal isoglosses sometimes coincide with non-linguistic boundaries (mountain ridges, rivers, religious or political borders). The boundary between *Müntschi* and *Schmutz* in Figure 3.3 follows quite neatly the Brünig-Napf-Reuss line (Section 1.2.5), which is at the same time a topographical, political (Bern vs. Luzern) and denominational (protestant vs. catholic) border.

However, there is also evidence for continuous dialectal variation:

- The argument of dialects changing in uninhabited areas rather than within villages holds in a non-mobile society. The increasing mobility in the last decades has led to heterogeneous distributions even inside villages (see example (1) below).

- Many phonetic phenomena are continuous in nature. The SDS authors distinguish five degrees of openness in the *o-u* vowel change (see Figure 3.6 on page 101). Generally, such phonetic change happens gradually: between the area with most open vowels and the area with most closed vowels, there are transition zones with intermediate degrees of openness.

- Even if the phenomenon in question is not continuous in nature, there can be a gradual change in the proportion of the population using one variant or another. For example, in subordinate clauses, a finite modal verb can either precede or follow the infinitive of the main verb (see Appendix A.4.6, page 371). Four inquiry points in the canton of St. Gallen, aligned in a row from East to West and distant by 40 km, show the following distribution (SADS I/9, most natural variant):[5]

(1)

	Ebnat-Kappel	*Wildhaus*	*Grabs*	*Sevelen*
VVINF – VMFIN	0%	25%	50%	60%
VMFIN – VVINF	100%	75%	50%	40%

  It is difficult to determine a clear boundary between the two word orders. In any case, it does not completely coincide with the topographic boundary between the Toggenburg (Ebnat-Kappel, Wildhaus) and the Rhine valley (Grabs, Sevelen).

In summary, continuous dialectal variation can be interpreted in two ways: as continuous change in the realization of a linguistic phenomenon itself (as in the *u-o* example), or as continuous change in the proportion of people using discrete variants of a linguistic phenomenon (as in the verb order example). For practical reasons, continuous linguistic phenomena have been discretized by the SDS authors and even more in our own preprocessing steps. In contrast, continuous usage proportions are easily encoded in continuous field maps (see below).

5 In the following table, we use the abbreviations as defined in the Stuttgart-Tübingen tagset (Thielen et al. 1999): VVINF stands for "infinitive of a lexical verb", VMFIN stands for "finite form of a modal verb".

We prefer to view dialectal variation as a continuous phenomenon in terms of continuously changing proportions of usage. Despite the large uninhabited Alpine areas, the Swiss German data is not compelling enough to adopt a model of purely discrete dialectal variation. However, continuous variation can occur over smaller or larger distances. In the cited 'kiss' example, the transition zone between *Müntschi* and *Schmutz* is very small, whereas the transition zone between *Schmutz* and *Kuss* spreads over a wider area.

Besides these linguistic aspects, there are also more technical reasons that call for a continuous representation of dialectal data:

- For an end user, it is more appealing to use a system that is able to produce text in any dialect, not just in those that have been registered in the atlas survey.

- Surface maps are better suited for some types of evaluation. For example, Section 6.5 presents an algorithm that identifies the dialectal origin of a given text. Usually, this algorithm is not precise enough to predict a single village of origin. Rather, it allows to predict an area of origin that is composed of various zones with higher and lower probabilities. Such an area map is at the same time more meaningful and more appropriate for this particular problem.

- The essence of continuous surface maps is to predict dialectal variants in places that are not part of the atlas inquiry network. This prediction is a nice problem of artificial intelligence and machine learning (see Section 4.6.2) and is worth while studying by itself.

4.6. Tessellation and interpolation

The original SDS and SADS maps are point maps, but we have argued in the preceding section that continuous surface maps are more adequate for our approach to dialect processing. Hence, we first present different ways of deriving surface maps from point maps (4.6.1), and then show different methods of computing the attribute values of the surface map cells (4.6.2).

4.6.1. Tessellation

Continuity can be approximated in two ways in geographical information systems: as adjacent polygons of variable size (using a vector model), or as a fixed grid of raster cells (using thus the raster model). This section explains how the points of the original data set can be transformed into continuous surfaces of either of the two types. This type of transformation is called **tessellation** (Burrough and McDonnell 1998, pp. 23-25).

Thiessen polygons Probably the most popular method to transform point data into surface data is to create polygons of irregular shape around each input point, such that the borders between the polygons are situated halfway between the input points. The resulting grid is called 'Voronoi mosaic' or 'Thiessen polygons'. With this method, each input point is projected onto one polygon.

Triangular irregular networks Another approach is to create a triangular irregular network (TIN). Here, any three adjacent input points are connected between them, yielding a grid of triangles. This model is mainly used for digital elevation models, and – under the name of *finite elements modeling* – in engineering.

These tessellation techniques transform point data into polygon data, staying inside the vector paradigm. The number and size of the polygons (or triangles) depends on the number and position of the input points.

Fixed-grid raster This technique creates a raster map where all cells are of the same (rectangular) shape and size. Hence, the number of cells is independent of the number of points in the input dataset. By choosing a high grid resolution, a reasonably good approximation of a truly continuous field can be achieved.

Figure 4.6 shows examples of the three tessellation methods explained above. We have chosen to use fixed-grid rasters, for the two following reasons:

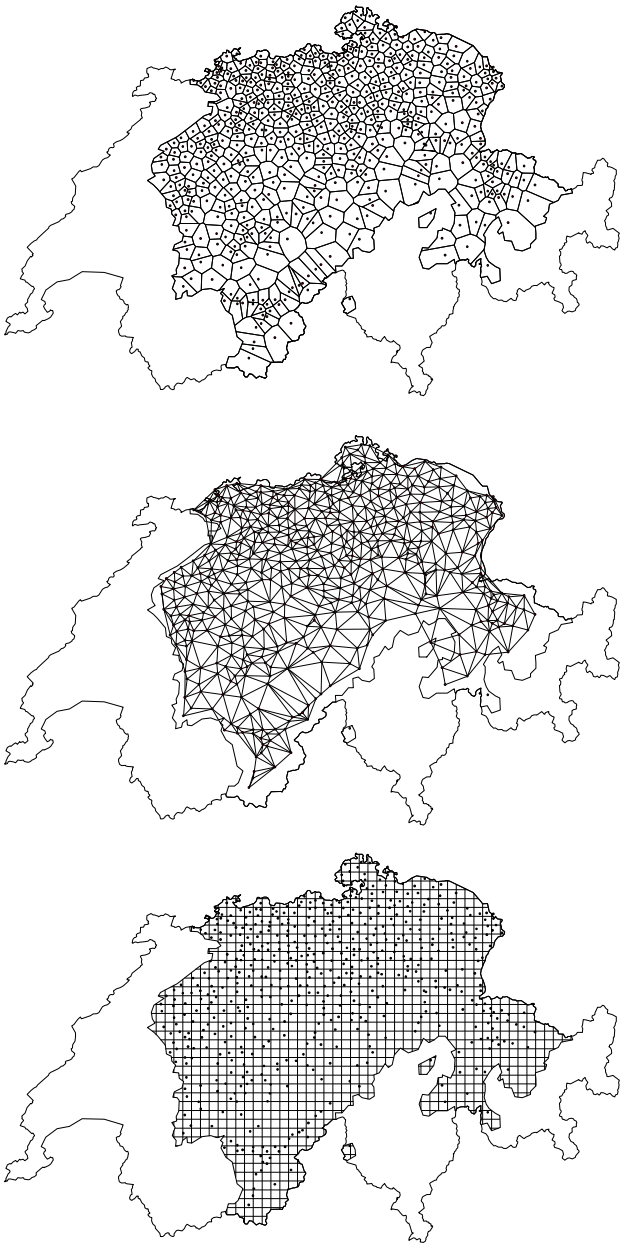


Figure 4.6.: Surface maps created with different tessellation methods on the basis of the SDS inquiry points: Thiessen polygons at the top, TIN in the center, regular raster grid at the bottom. The cells of the raster grid are independent of the number and position of the inquiry points.

- We will need to overlay maps of the two data sources SDS and SADS. Since they have a different number of inquiry points, the topology of the Thiessen polygons or TINs would vary depending on the data source. Fixed-grid rasters are independent of the number of inquiry points in the source.

- We will need to overlay maps of different phenomena, for example of different phonetic transformations occurring in the same word. Fixed-grid rasters can be interpreted as two-dimensional matrices. Map overlaying receives thus a precise mathematical interpretation as matrix addition or multiplication.

Our raster maps are generated with a cell size of 1 km^2.

Different tessellation methods define different topologies: triangles, rectangles or other polygons, with constant or variable sizes. However, they do not define the non-spatial attribute values that are associated to these cells. This is done by interpolation methods.

Note that in general, tessellated maps only contain one attribute value per cell. For our work, this means that for a linguistic variable with n variants, n tessellated maps have to be created. While the n maps share the same topology, their attribute values are calculated independently.

4.6.2. Interpolation methods

Burrough and McDonnell (1998, p. 98) define **interpolation** as follows:

> *Interpolation* is the procedure of predicting the value of attributes at unsampled sites from measurements made at point locations within the same area or region. […] Interpolation is used to convert data from point observations to continuous fields so that the spatial patterns sampled by these measurements can be compared with the spatial patterns of other spatial entities.

Interpolation methods are based on the general idea of **spatial autocorrelation**: "on average, values at points close together in space are more likely to be similar than points further apart" (Burrough and McDonnell 1998, p. 100).

In the following, we review the most important interpolation algorithms. We call the point locations at which the measurements were made *input points*, whereas the unsampled sites for which the values have to be inferred are called *output cells* or *grid cells*.

Map interpolation can also be formulated as a machine learning problem. The set of input points corresponds to the training data in a two-dimensional feature space. Each input point is a sample defined by two features, namely its x and y coordinates, and labeled with the dialectal variant(s) that occur in it. The set of output cells (or rather, the set of output points at the center of each cell) represent the test data: their coordinates are known, but their labels have to be predicted (Mitchell 1997, ch. 8).

Nearest neighbor This algorithm assigns to each output cell the value of the closest input point. If the output map consists of Thiessen polygons, this is trivial: each polygon is assigned the value of the input point located at its center. With a raster grid, the topology obtained by the nearest neighbor algorithm will approximate the Thiessen polygons. This algorithm involves no computation whatsoever of attribute values, and results in rather sharp value changes at cell boundaries.

Inverse distance weighting (IDW) This method departs from the previous one by taking into account several input points in order to compute the value of an output cell. The input points are weighted by their distance from the output cell, so that input points closer to the cell are more influential than distant points.

The value of an output cell w_{t_j} is computed as follows:

$$w_{t_j} = \frac{\sum_{i=1}^{n} w_{t_i} \cdot \frac{1}{d(i,j)^r}}{\sum_{i=1}^{n} \frac{1}{d(i,j)^r}}$$

where the t_i represent the positions of the input points and the t_j the positions of the output cells. w_{t_i} and w_{t_j} represent the attribute values of the respective points. $d(i,j)$ corresponds

to the distance between t_i and t_j.[6] r is a parameter that allows to weight the influence of the distance: the lower the value of r, the lower the influence of far-away data points. Figure 4.7 illustrates inverse distance weighting. The IDW method results in more smoothed surfaces than the nearest neighbor method.

The expression $\frac{1}{d(i,j)^r}$ in the above formula determines the weight of each given w_{t_i} with respect to distance. In spatial statistics, such an expression is called a **kernel function** (Mitchell 1997, p. 236). A more sophisticated kernel function is shown below.

Kernel density estimation Except for distance, the IDW method treats all input points equally, independently of their overall likelihood in the data set. As a result, it is rather sensible to outliers. More sophisticated interpolation methods correct this shortcoming by relying on stochastic principles. The general idea is that some input points fit better into the global picture and should therefore be trusted more than those input points that behave like outliers. This approach is particularly well suited for dialectological data, which have often been compiled on the basis of a single informant per inquiry point:

> If, for instance, a single outlier appears in an otherwise uniform area, it is very likely that, had another person been asked, he or she would have given the variant of the surrounding locations. At the same time, it is to be expected that, if enough persons had been asked in the surrounding area, a few of them would have come up with the variant of the outlier. This is a very simple example to illustrate our basic assumption: the more records of a certain variant appear in a location's environment, the more likely it is that this variant would have been given as an answer at the location itself if many people had been asked, even if the informant that actually was interviewed has given a different one. (Rumpf et al. 2009, pp. 282-283)

Thus, some amount of random fluctuation has to be assumed in dialectological data. Statistical methods allow to dissociate the random fluctuation from the "true" dialectal variation.

6 Distances are usually calculated with respect to the center of a cell.

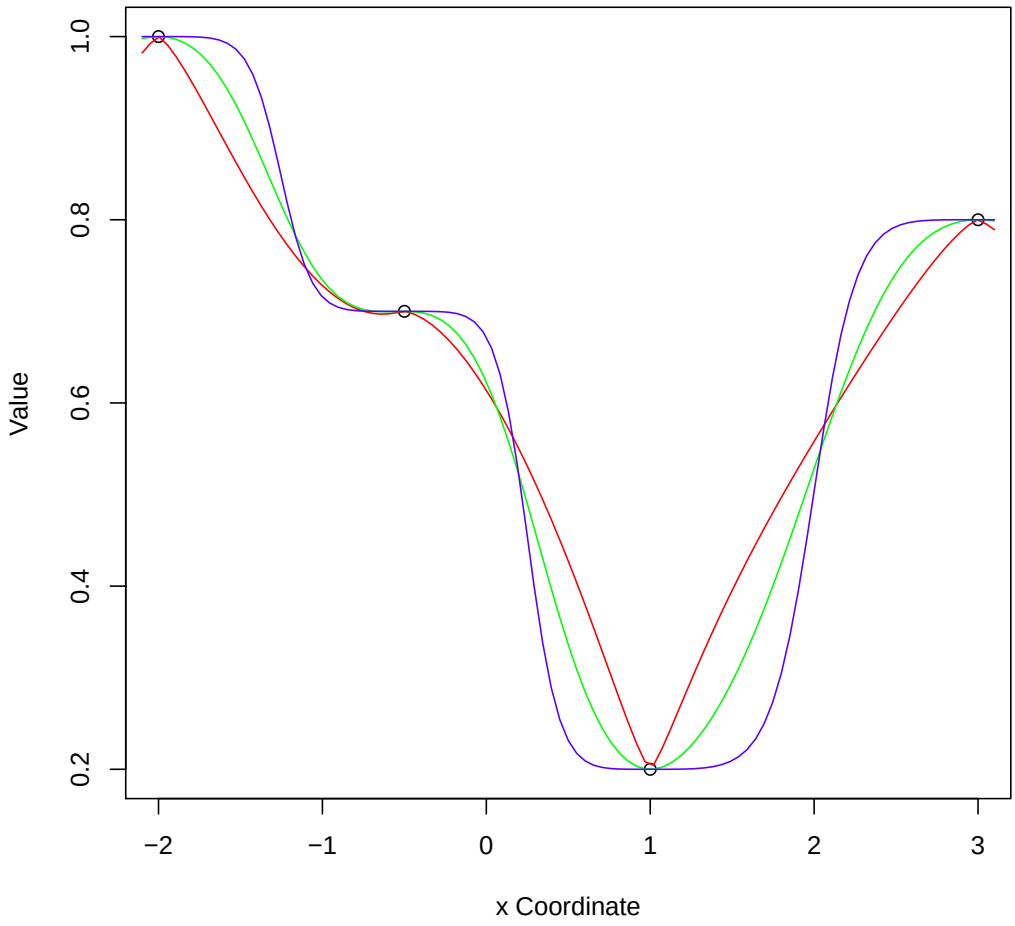


Figure 4.7.: This figure illustrates the IDW algorithm on a one-dimensional map, consisting of four input points (displayed as black circles). Their respective attribute values are displayed on the vertical axis. The IDW algorithm draws a continuous curve through the input points. The different curves represent different settings of the r parameter: red for $r = 1.3$, green for $r = 2$, blue for $r = 3$. Note that in all cases, the curve passes directly through the input points.

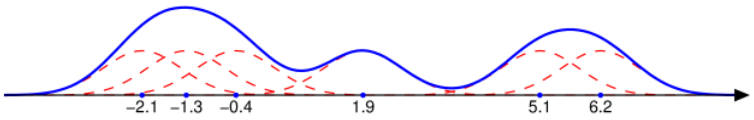


Figure 4.8.: This figure reproduces a one-dimensional map where the variant is observed at coordinates -2.1, -1.3, -0.4, 1.9, 5.1 and 6.2. A Gaussian curve is laid on top of each observation, illustrated in red. The probability density function is defined as the sum of all probability values occurring in each position. As a result, the more observations there are in the neighborhood of a point, the higher its density.

According to the technique presented by Rumpf et al. (2009), a so-called intensity field is estimated over the whole area for each variant. The intensity field of a variant represents the likelihood of that variant occurring at the respective location. It is estimated by placing a probability density distribution around each input data point. The probability density is highest right on the data point and decreases according to distance. The probabilities of adjacent data points are added, resulting in a single curve. This curve is described by a kernel function. Figure 4.8 illustrates this process. The entire process is known as **kernel density estimation**.[7]

The general formula for kernel estimation predicts a value for each output cell t_j, based on the input points t_i:

$$w_{t_j} = \frac{\sum_{i=1}^{n} w_{t_i} \cdot K(\frac{d(i,j)}{h})}{\sum_{i=1}^{n} K(\frac{d(i,j)}{h})}$$

Two parameters determine the estimation. The kernel type K is a function which determines the shape of the curve that is drawn around the observations. The bandwidth h determines the scale of the curve. Rumpf et al. (2009) propose to use the standard normal (i.e. Gaussian) kernel defined as follows:[8]

$$K(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

7 See Bishop (2006, chap. 2.5), Mitchell (1997, chap. 8) and http://fr.wikipedia.org/wiki/Estimation_par_noyau. The diagram in Figure 4.8 stems from http://commons.wikimedia.org/wiki/File:Parzen_window_illustration.png, access date 7.9.2011.

8 In contrast to the general formula found in many textbooks, Rumpf et al. (2009) use a slightly different formula that does not contain the square root in the denominator. At the moment, it is unclear if this is an oversight or if this change has been made on purpose.

The optimal value for the bandwidth depends on statistical properties of the dataset and has to be determined empirically. In order not to overload this presentation, we explain this process in a separate section below (4.6.3). Figure 4.9 illustrates the result of the kernel estimation method.

External indicator functions Some other interpolation methods based on spatial statistics allow the addition of external indicator functions. These methods are commonly subsumed under the term **kriging**. In our field, indicator functions may be used to investigate external factors of dialectal variation. For example, the size and socio-economic importance of the inquiry points could influence dialectal change over time. By including an indicator function based on population size, one could obtain a model where dialects of large cities propagate more easily than dialects of small villages.

To sum up, the judicious use of interpolation algorithms may not only allow us to fill the gaps between the inquiry points, but also help us to exclude noise from the original data, or even simulate some language-external diachronic changes. However, comparison between different algorithms is difficult: there are many parameters to take into account, and there are no universally accepted evaluation procedures or guidelines. Therefore, we simply follow the kernel density interpolation method proposed by Rumpf et al. (2009), which has been successfully applied to dialectological data.

4.6.3. Bandwidth estimation

For the kernel density estimation method, the choice of the bandwidth parameter has been found to be more important than the choice of the kernel. The bandwidth indicates how strong a location should influence its surroundings, or in other words, up to what distance the presence of a certain variant should be considered.

Recall that a kernel function predicts an attribute value from a given coordinate pair. These predicted values can be compared with the original values of the input points. For example, in Figure 4.9, the bandwidth settings 0.2 and 0.5 show near perfect fit with all four input values,

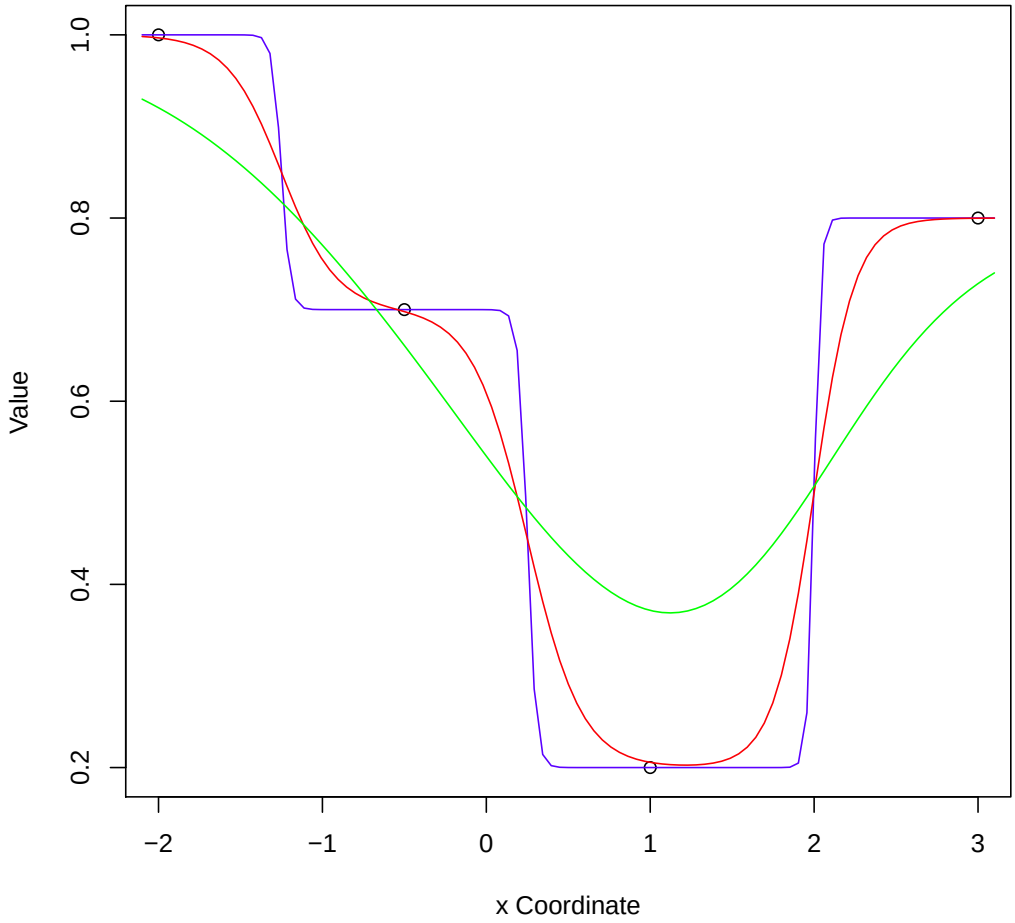


Figure 4.9.: This figure illustrates kernel density estimation on the one-dimensional data set used in 4.7. The different curves represent different settings of the bandwidth parameter: blue for $h = 0.2$, red for $h = 0.5$, green for $h = 1.0$. All curves are based on the Gaussian kernel. Note that the green curve displays rather aggressive smoothing and predicts values that differ considerably from the given input points.

while the bandwidth setting 1.0 predicts attribute values that are far away from the original input values. So, different settings of the bandwidth parameter define different degrees of fit, and the goal is to select the setting that yields the best degree of fit.

One of the methods to determine the bandwidth parameter is **least-squares cross-validation** (LSCV). The degree of fit between a kernel function and a data set is defined as the sum of squared residuals; a residual is the absolute difference between the original value and the predicted value at a given point. The goal is to minimize the sum of squared residuals, hence the name of **least squares**.

It is unfortunate to use the same data points for the definition of the kernel function and for its validation. To avoid this issue, one uses a technique called **cross-validation**. The original data set is partitioned, with one part used to define the kernel function, and the other part used for its validation. This partitioning is repeated several times, so that every data item at least serves once as a training example and once as a validation example.

With respect to dialectological maps, Rumpf et al. (2009) first determine an individual bandwidth value for each variant. They then take the weighted sum of all variant bandwidth values to obtain a single bandwidth value for each variable. The sum is weighted by the relative number of occurrences of each variable. For the time being, we depart from this approach by using a single bandwidth value determined from a small set of randomly selected linguistic variables. LSCV is relatively time-consuming, and to our knowledge, there is no scriptable tool that performs this estimation directly on *ArcGIS* data.[9]

4.6.4. Value normalization

Interpolation yields a distinct raster map for every variant. Hence, a linguistic phenomenon, or variable, is represented by a set of variant maps. If the interpolation is performed independently for each variant, the sum of all variant values may vary from cell to cell. This is not so much of a problem if the only task is to determine the most likely variant of a given variable at a given

9 Unfortunately, the tool used by Rumpf et al. (2009) could not be made available to us for copyright reasons.

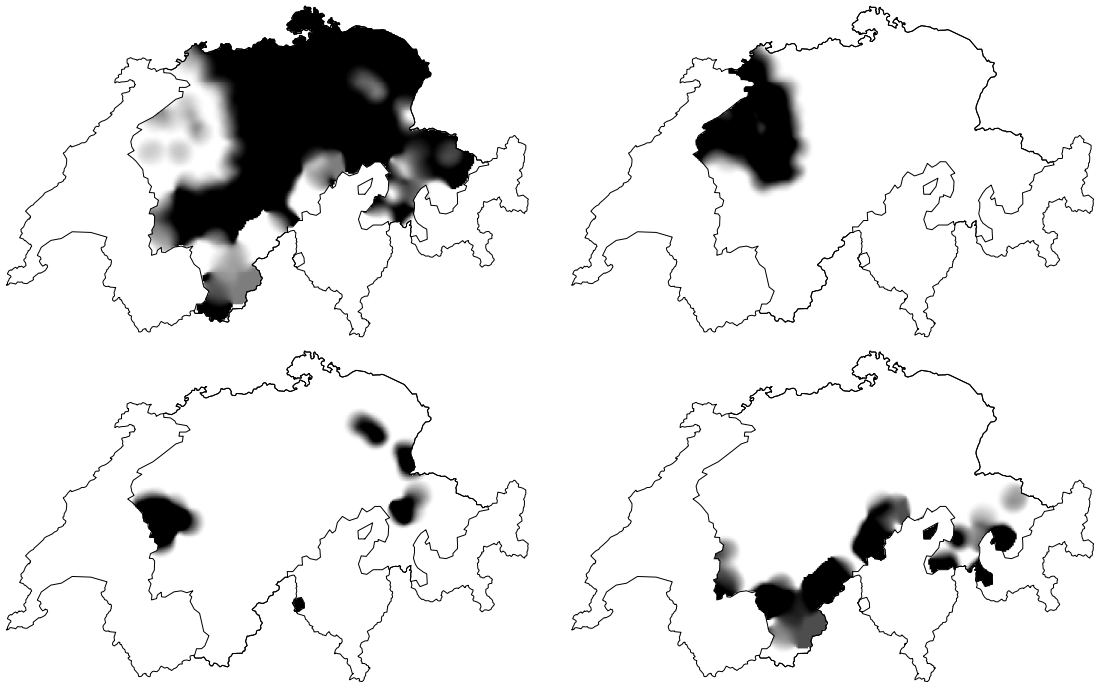


Figure 4.10.: Interpolated, normalized raster maps for the four main variants of the final-*nd* variable: *-nd* on the upper left, *-ng* on the upper right, *-nn* on the lower left, *-nt* on the lower right. Black areas represent high probabilities, white areas represent low probabilities.

geographic point. However, if maps from several variables are to be combined by addition or multiplication, they should all be weighted equally.

Value normalization pursues the goal of converting the variant values to probabilities. The normalized values can be interpreted as conditional probabilities $p(v \mid t)$, where v is the variant and t is any raster cell situated in German-speaking Switzerland (we call this set of cells *GSS*). As a result of normalization, the probabilities of all variants v describing a variable V sum up to 1 at any given point $t \in GSS$:

$$\underset{t \in GSS}{\forall} \sum_{v \in V} p(r \mid t) = 1.0$$

Figure 4.10 presents the interpolated, normalized raster maps for the final-*nd* variable discussed previously.

4.6.5. Map algebra

The cells of a normalized raster map contain probability values. Hence, a raster map can be interpreted as a two-dimensional matrix, on which any type of arithmetic operation can be performed. In particular, one can specify a **map algebra** which redefines the algebraic notations commonly applied on single numbers to operate on entire raster maps instead. For instance, multiplication of two maps is defined as cell-wise multiplication of the respective values. Addition of two maps is defined as cell-wise addition of the respective values.

In our probabilistic interpretation of map cells, multiplication corresponds to **intersection** or **logical conjunction**: multiplying maps *A* and *B* yields a map of the areas where both *A* **and** *B* are valid. Addition corresponds to **union** or **logical disjunction**: adding maps *A* and *B* yields a map of the areas where one of *A* **or** *B* is valid. If maps *A* and *B* represent variants of different linguistic variables, their point-wise addition may yield values that are higher than 1. In this case, we compute the point-wise **maximum** instead of addition.[10]

4.7. Web access

In the preceding sections, it has been shown how scanned SDS maps have been digitized (yielding electronic point maps) and then interpolated (yielding raster maps). We have made the resulting electronic maps available online, in interfaces based on *Google Maps*.[11] While these maps lack the detailed legends of the original SDS maps, the easy interactive access may be helpful not only to the scientific community, but also to a larger audience.

The point map interface is connected to a *Python* script that reads in the corresponding *ArcGIS*

10 At first sight, it might not make sense to compute the disjunction of two non-related variants. But imagine that a Standard German noun is rendered as a feminine in Eastern dialects and as a masculine in Western dialects. The Eastern variant will have a feminine plural ending, the Western variant will have a masculine plural ending. In the end, one will have to compare both forms, although the feminine and masculine plural endings come from independent variables.

11 The current URL address is http://latlntic.unige.ch/~scherrey/. This web access has been presented for the first time in Scherrer (2010).

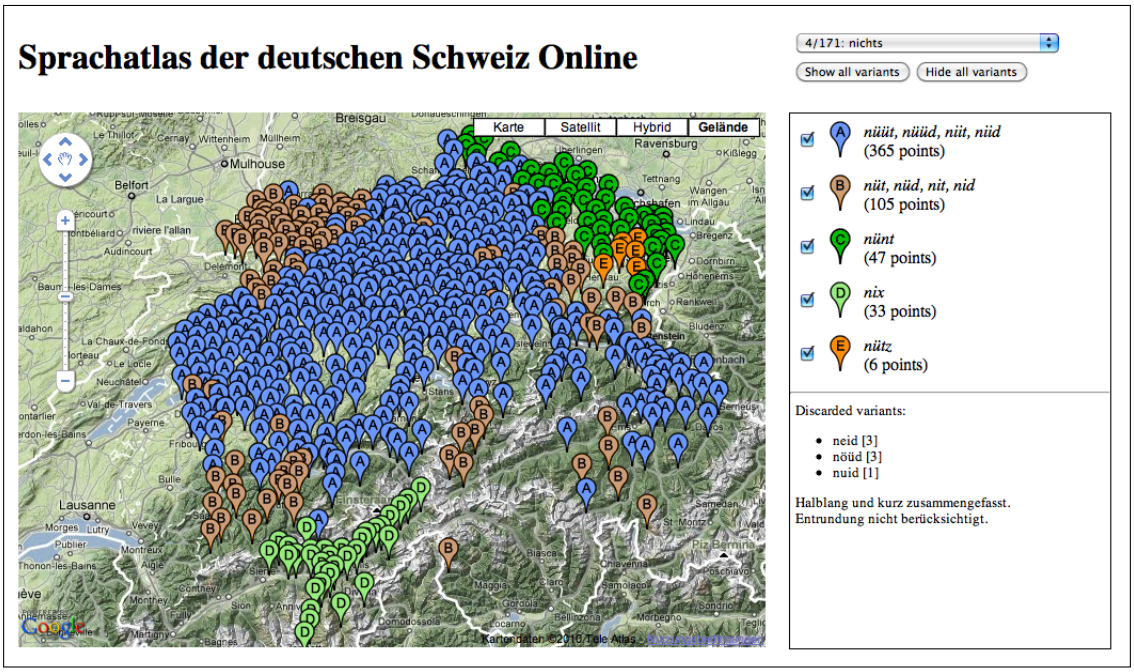


Figure 4.11.: Screenshot of the web interface displaying SDS point maps.

files and displays the dialectal variants in the form of markers on a *Google* map. Figure 4.11 shows a screenshot of this interface. The raster map interface allows the user to view the interpolated data in the form of semi-transparent images that are superposed on a *Google* map. Figure 4.12 shows a screenshot.

In principle, SADS data could be displayed in the same interfaces. However, we leave the choice of providing web access to the editors of the atlas.

4.8. Conclusion

In this chapter, we have defined some important concepts and processes of cartography and spatial analysis. For instance, we have explained what kind of map formats exist and which format we find most appropriate for our purposes. We have also reviewed different interpolation algorithms. Some cartographic choices are arbitrary – after all, cartography is sometimes

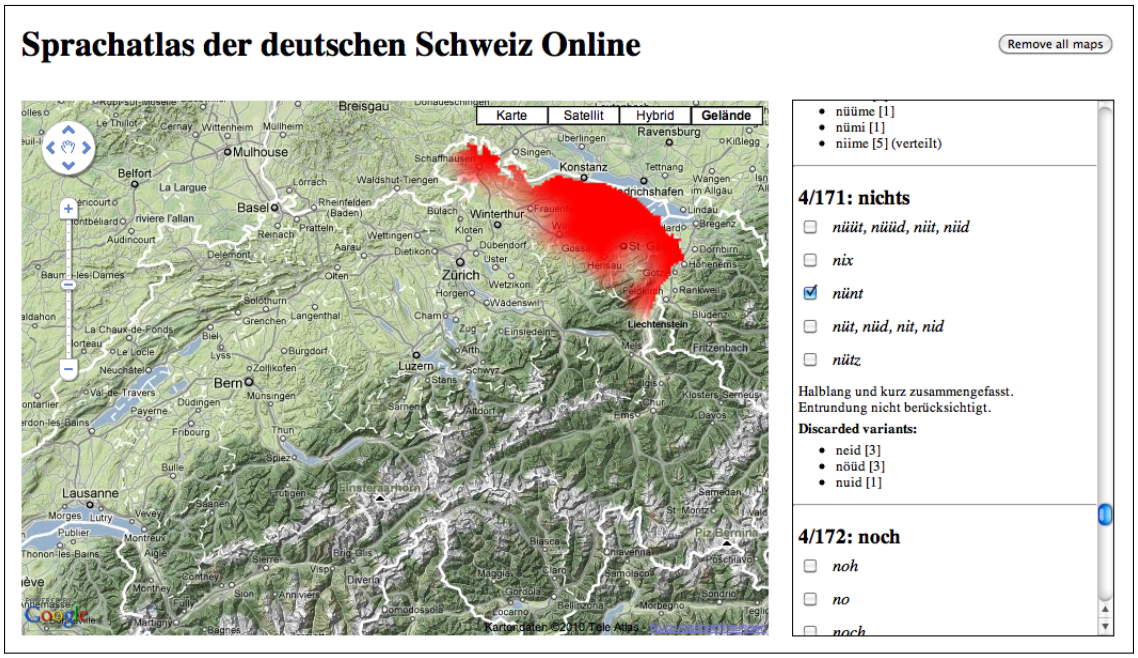


Figure 4.12.: Screenshot of the web interface displaying interpolated SDS surface maps.

considered an art rather than a science. While we cannot offer empirical justifications on every decision taken, we can at least explain the linguistic and computational motivations of our choices. This is what we have aimed for in this chapter.

5. Georeferenced transfer rules

5.1. Introduction

Once the linguistic phenomena to be treated have been selected according to the criteria specified in Chapter 3, the corresponding maps are scanned, converted and interpolated following the methods lined out in Chapter 4. This results in a set of fixed-grid raster maps whose cells contain probability values. These probabilistic maps represent the geographic distributions of the dialectological phenomena. They now need to be linked to transfer rules that transform Standard German utterances (words and sentence structures) into Swiss German utterances. This is the object of the present chapter. It consists of three parts. The first and second parts present the word-level transfer rules: these include phonetic, lexical and morphological transformations. The first part (Section 5.2) provides an implementation where the transfer rules are conceived as regular expressions and stored in a database. The second part (Section 5.3) presents a reimplementation using finite-state transducers. The third part (Section 5.4) deals with syntactic transformation rules.

Throughout this chapter, we use the terms **transfer rule** and **transformation rule** as synonyms. One rule represents one linguistic phenomenon (or variable) and may yield different outcomes for different dialect regions. In this sense, one rule is a set of rewrite actions that share the same left hand side. The different rewrite actions are called **rule variants**.[1]

The originality of our approach lies in the **georeferencing** of the rules and their variants. Each

1 In Scherrer and Rambow (2010), we have used a different terminology: we called each rewrite action a **rule**; the set of rules sharing the same left hand side was called **phenomenon**.

rule variant typically corresponds to an interpolated probability map. Hence, a **georeferenced transfer rule** is a transfer rule whose variants are marked with a map descriptor – the file system path leading to the raster map file. Using georeferenced rules allows to discard geographically incompatible rule variants immediately during the derivation process.

5.2. Database implementation of word-level transfer rules

Although the whole work presented in this thesis is framed as an instance of machine translation from Standard German towards Swiss German dialects, the word-level transfer can actually be viewed as a case of morphology generation. From this point of view, the word-level transfer rules presented here pursue the aim of Swiss German word generation, formulated as follows: *Given a Standard German root and a set of morphosyntactic features, generate all inflected forms that are valid in the different Swiss German dialects.*[2]

In contrast to other work in computational morphology generation,[3] this approach can be qualified as **cross-lingual** and **multi-dialectal**. It is cross-lingual in the sense that the language variety of the input root (Standard German) is different from the language variety of the output forms (Swiss German). It is multi-dialectal because it aims to generate all forms that occur in the different dialects of German-speaking Switzerland.

5.2.1. General system architecture

Standard German is characterized by rich and highly ambiguous inflection marking and by productive compounding processes. In contrast to other languages, German compound words look like single words, without any typographic indication (such as spaces or dashes) of the component boundary. Thus, a word form generator for Swiss German has to cope with these two phenomena. Since the word form generation system in its strict sense requires morphologically

2 This section is largely based on work presented in Scherrer (2011a).

3 For an overview of computational morphology systems, especially for Standard German, see Mahlow and Piotrowski (2009b).

analyzed and decomposed words, the translation pipeline consists of several preprocessing and postprocessing steps.

The entire process of translating a Standard German word into Swiss German is depicted in Figure 5.1. Boxes **(1)** and **(2)** reflect the preprocessing steps. Box **(3)** represents the word form generation system in its strict sense; it is presented in detail in the grey box on the right hand side of the figure. Box **(4)** is a postprocessing step that re-composes the separately translated word components into a single compound word.

The word form generator for Swiss German takes as input a Standard German simplex lemma and a set of morphosyntactic features that determine the inflected form to be generated. It consists of two generation paths, represented by boxes **(a)**-**(b)** and **(c)**-**(d)** respectively. If there is a lexical transformation rule that applies to the given lemma, it is applied **(a)**. Each lexical transformation rule may select a specific subset of phonetic transformation rules to be applied **(b)**. However, if no lexical rule applies to the lemma, the second path is chosen. The lemma is transformed into a Standard German root and annotated with lexical features like inflection class **(c)**, and phonetic transformation rules are applied to the Standard German root **(d)**.

The input of step **(e)**, independently of the path chosen upstream, consists of a dialectal root and a set of morphosyntactic and lexical features. In step **(e)**, these features are converted into inflectional affixes (suffixes and *umlaut* vowel alternation). Finally, some phonetic rules that specifically deal with the inflected dialectal forms are applied **(f)**.

In the following sections, each of these steps will be described in detail (throughout this description, we will refer back to Figure 5.1). Section 5.2.2 describes steps **(1)** and **(2)**. The lexical rules of **(a)** are presented in Sections 5.2.3 and 5.2.11, the annotation heuristics of **(c)** in Section 5.2.4. The phonetic rules used in steps **(b)**, **(d)** and **(f)** are dealt with in Sections 5.2.5 to 5.2.9. Sections 5.2.10 and 5.2.11 focus on the affixation rules of **(e)**.

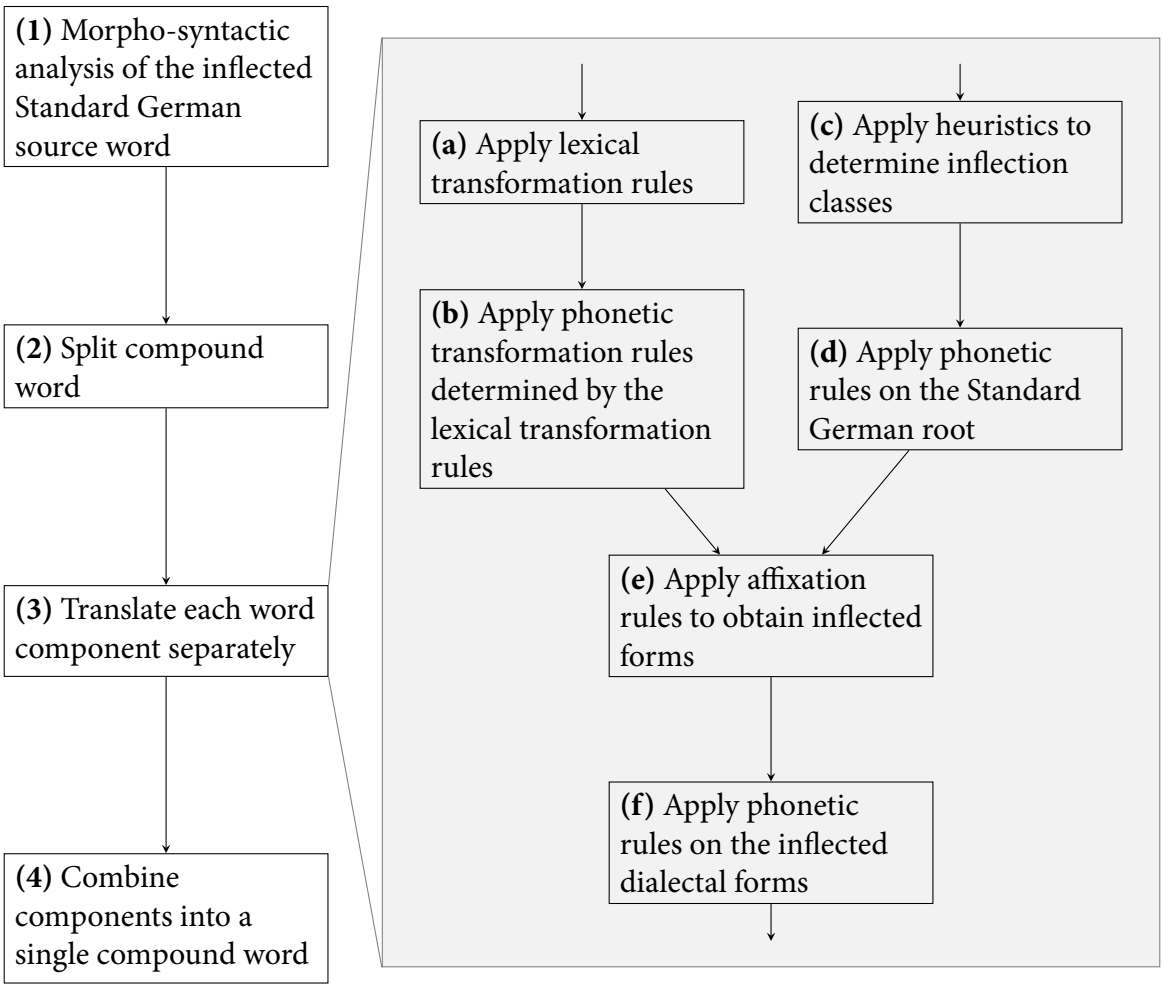


Figure 5.1.: General architecture of the dialect word generation system. The white boxes on the left hand side represent preprocessing and postprocessing steps.

5.2.2. Morpho-syntactic analysis and compound splitting

Since our system operates as a morphological generator, the input must be lemmatized and morpho-syntactically annotated with part-of-speech tags (e.g. common noun, auxiliary verb, personal pronoun) and morphological features (e.g. number, gender, person). Notably, we have decided to use the STTS tagset (Thielen et al. 1999), which has become a *de facto* standard for German part-of-speech tagging. Moreover, compound words must be decomposed, so that each component can be treated separately.

Compound word decomposition is indicated for two main reasons. First, phonetic rules should not operate across morpheme boundaries. For example, the phonetic rule that transforms *st* into *scht* should not be triggered in the word *Haustür* 'house door', because the *s* and the *t* belong to different components. Second, the lexical rule base and the word lists for annotation heuristics may not cover all relevant compound words. Decomposition allows to reduce the size of these resources and at the same time increase their efficiency. Compound noun decomposition has also been used in machine translation to increase lexical coverage (see Section 2.10).

We have explored three alternatives for morpho-syntactic analysis and compound splitting. The first alternative relies on the *Fips* parser for both subtasks. The second alternative uses *TreeTagger* for the morpho-syntactic analysis and *Morphisto* for compound splitting. The third alternative relies on manually annotated sentences such as those found in the TIGER treebank.

Fips

Fips (Wehrli 1997) is a multilingual parsing engine based on hand-crafted grammar rules and a manually populated lexicon. The German *Fips* lexicon contains about 43 000 lemmas.[4] Analyzing a sentence with *Fips* amounts to creating a constituent tree according to the grammar rules. During this process, morpho-syntactically ambiguous words are disambiguated. As a result, each word is associated with a unique part-of-speech tag and an unambiguous set of mor-

4 Figure as of 13.9.2012.

phological feature values.[5] *Fips* also includes a compound splitting module, which is triggered whenever a word form cannot be found in the lexicon.

Adapting the output of *Fips* for our purposes was not trivial, since *Fips* uses a proprietary tag set that is partially incompatible with the (more recent) STTS. Hence, STTS-compatible part-of-speech tags had to be generated on the basis of the internal *Fips* features. In some cases, the *Fips* features are not detailed enough to cover the STTS distinctions. For example, the Standard German word *am* can be used as a prepositional determiner with a noun phrase (1a), or as a superlative particle with adjectives (1b). These two uses are distinguished in STTS (tags APPRART for the former, PTKA for the latter), but not in *Fips*.

(1) a. Er arbeitet am schönsten Tag. 'He works on the nicest day.'
b. Er arbeitet am schönsten. 'He works most diligently.'

We also encountered issues when generating STTS-compatible morphological feature tags, since *Fips* does not reliably unify all features. For instance, the verb form *geht* 'goes' is morphologically ambiguous (third person singular, or second person plural), but *Fips* does not systematically keep a single reading. In this case, we used a heuristic to prefer the third person singular readings, which are much more frequent.

Technically, *Fips* only runs as an executable on Windows. Since the rest of our system is Unix-based, *Fips* is set up as a web service which communicates via HTTP.

TreeTagger

TreeTagger (Schmid 1995) is one of the most popular part-of-speech taggers and has achieved good performance on German data. However, the pre-trained German language model was insufficient for our purposes, since it only annotates part-of-speech tags, but none of the morphological features. Therefore, we trained an extended model on the basis of the TIGER treebank (S. Brants et al. 2002). The extended model contains 486 different tags (combining part-of-speech and morphological information), compared with 54 in the pre-trained model. It has been shown

5 We only consider the single most probable syntactic analysis.

that the performance of part-of-speech taggers depends largely on the size of the tag set (Hajic 2000), so we suppose (but have not empirically confirmed) that the extended model has a lower performance.

Compound splitting with Morphisto

Since the TreeTagger does not provide compounding information, we resort to an additional morphological analysis with *Morphisto* (Zielinski and Simon 2009). This tool takes an inflected Standard German word as input and displays all possible analyses. An analysis specifies (i) the components of a compound word, (ii) the lemmas of the components, (iii) the derivational affixes, (iv) the part-of-speech tags of the components and of the resulting word form, and (v) the morphological features. The following example shows the output for the compound noun *Kochtöpfe* 'cooking pots'.

(2) Koch<NN>Topf<+NN><Masc><Nom><Pl>
Koch<NN>Topf<+NN><Masc><Gen><Pl>
Koch<NN>Topf<+NN><Masc><Akk><Pl>
kochen<V>Topf<+NN><Masc><Nom><Pl>
kochen<V>Topf<+NN><Masc><Gen><Pl>
kochen<V>Topf<+NN><Masc><Akk><Pl>

Indeed, the output of *Morphisto* is unnecessarily ambiguous. All *Morphisto* analyses that are incompatible with the *TreeTagger* annotation are removed. For example, if *TreeTagger* has annotated *Kochtöpfe* with an accusative tag, the nominative and dative readings may be discarded.

Decomposition ambiguities are also handled by means of a heuristic. Following Koehn and Knight (2003), we rank the compound words according to the geometric mean of the frequencies of their components. Frequency information is extracted from the TIGER treebank (S. Brants et al. 2002). We keep the decomposition with the highest mean frequency; if the mean frequency is lower than the frequency of the entire word, it is not split at all. This heuristic eliminates many spurious decompositions, but unfortunately, being a heuristic, also eliminates some correct analyses. In example (2), the decomposition with the verb *kochen* 'to cook' should be preferred on semantic grounds. However, since the noun *Koch* 'cook' is more frequent than the verb *kochen*,

the former decomposition is retained.[6]

Morphisto has a serious drawback when used for noun decomposition: its output is a list of lemmas, from which it is not evident to recover the surface forms, since they may contain so-called linkage elements (*Fugenelemente*). These elements can be one of the affixes *-(e)s*, *-(e)ns* or *-(e)n*. They can also be negative, in the sense that a *-e* or *-en* appearing in the citation form is deleted. For instance, in example (2), the verbal suffix *-en* must be removed from *kochen* to yield *Kochtopf*. We have devised a specific script that retrieves the original word forms and annotates the components with a feature that specifies the linkage element used.

Summary

Preliminary tests have shown that the *TreeTagger-Morphisto* combination works better than *Fips*. This is mostly due to the use of frequency information in *TreeTagger* and *Morphisto*, or to the lack of its use in *Fips*, as exemplified above by the ambiguity of *geht*. As a result, most experiments rely on *TreeTagger* and *Morphisto*. However, *Fips* becomes interesting when additional syntactic information is required, which cannot be obtained with a simple part-of-speech tagger. For instance, the syntactic transformations presented in Section 5.4 rely on full parse trees.

The third alternative mentioned above relies on manually annotated data. For some experiments discussed in Chapter 6, we will directly use the sentences of the TIGER treebank with their morpho-syntactic tags. Since TIGER does not provide compounding annotation, we use *Morphisto* for compound splitting.

Concerning Figure 5.1, step **(1)** consists of annotating the input words with a part-of-speech tag and a set of morphological feature tags. This is done either by *TreeTagger*, by *Fips*, or by extracting the relevant information from the *TIGER* corpus. Step **(2)** consists of splitting compound words, so that each component can be treated separately in steps **(a)** to **(e)**. This is done by *Morphisto* or by *Fips*. When using *Fips*, the two steps are part of the parsing algorithm and are technically indistinguishable.

6 In practice, both decompositions yield the same dialect words, so that choosing the semantically wrong analysis does not hurt the dialect generation performance in this particular case.

5.2.3. Lexical rules

The main idea of our approach is to leverage the etymological proximity between Standard German and Swiss German and the resulting high proportion of cognate word pairs by deriving Swiss German words from Standard German stems with the help of phonetic rules (step **(c)**). However, this is not always feasible. Sometimes, dialects use stems with a different etymological origin. In other cases, the dialectal form is etymologically related with the Standard German counterpart, but its derivation is idiosyncratic and impossible to generalize by the means of phonetic rules. Therefore, we introduce lexical replacement rules, which are applied as step **(a)**. They are stored in two database tables, lexicalVariables and lexicalVariants.[7] Examples are given in Table 5.1.[8]

The lexicalVariables table specifies the lemma of a word and its part-of-speech tag, but it may also contain finer-grained morphological information (field feats). The latter is mainly used for irregular inflection patterns (see Section 5.2.11, p. 158).

The lexicalVariants table specifies a dialectal stem that completely replaces the Standard German lemma. It also allows to modify morphological features in the field addFeats. This functionality is used for example to change the gender feature when a masculine noun lexeme is replaced by a feminine noun lexeme. The field phonRulesAfter allows to specify a subset of phonetic rules to be applied after lexical substitution. This functionality is illustrated in the first example of Table 5.1, where two additional variants *niit* and *nit* exist. The distinction between *ü*-based and *i*-based variants is regular and already accounted for by the phonetic rule *ü-i*. Hence, the output of this phonetic rule is added to the corresponding variants (see also Section 5.2.7, p. 153).

Currently, the database contains 260 lexical variables and 559 lexical variants. Most of them are high-frequency adverbs, pronouns, prepositions, and irregular verbs (for the latter, see again Section 5.2.11, p. 158).

7 Recall the definition of *variables* and *variants* in dialectology, given in Section 1.2.2 on page 32.

8 This table, as well as the following ones, is to be read as follows: The header line describes the structure of the table. The entries below the header show examples.

lexicalVariables:

rulename	*lemma*	*tag*	*feats*	*datafolder*
nichts	*nichts*	PIS		4-171-nichts
immer	*immer*	ADV		6-026-immer

lexicalVariants:

rulename	*form*	*mapname*	*addFeats*	*phonRulesAfter*
nichts	*nüüt*	dp_nüüt		ü-i
nichts	*nüt*	dp_nüt		ü-i
nichts	*nünt*	dp_nünt		
nichts	*nütz*	dp_nütz		
nichts	*nix*	dp_nix		
immer	*immer*	dp_immer		
immer	*geng*	dp_geng		
immer	*all*	dp_all		

Table 5.1.: Detail of the database tables for lexical substitution. The first example shows the dialectal variants of Standard German *nichts* 'nothing', whose idiosyncratic behavior is difficult to capture with phonetic rules. The second example refers to the translation of Standard German *immer* 'always', where completely different lexemes are used throughout the Swiss German dialect area.

5.2.4. Inflection classes

As in Standard German, Swiss German nouns use different plural affixes according to their inflection class. Inflection class cannot be unambiguously determined neither from the phonological structure of the base form nor from gender (see Appendix A.4.3, page 344). Hence, Swiss German nouns have to be annotated with a specific inflection class feature. When using lexical rules, inflection class features are added directly by these rules, using the addFeats field. However, if lexical rules do not apply, these features are added in a separate module **(c)**, on the basis of heuristics.

While there is no one-to-one correspondence between Standard German and Swiss German inflection classes, the latter can be derived from the former in a relatively straightforward way. Thus, the idea is to use a Standard German lexicon annotated with noun inflection classes and convert them to their Swiss German counterparts. As a source, we use the German *Fips* lexicon (Wehrli 2007), which contains 22 365 noun lemmas[9] annotated with one of 22 Standard German noun classes. A simple script filters these entries (singular-only and plural-only nouns do not need inflection information) and converts the noun class information. The result is a file that associates Standard German noun lemmas with Swiss German inflection classes. Whenever a Standard German noun has to be translated, its inflection class is looked up in this file, and the corresponding inflection class feature is added.

In some cases, the Swiss German noun class does not correspond to the Standard German equivalent. For example, Swiss German dialects tend to use *umlaut* plural marking more often than Standard German (e.g. *Hunde — Hünd* 'dogs', *Pullis — Pülli* 'sweaters'). Such particularities can be handled by manually modifying the inflection class file.

5.2.5. Regular expressions

The phonetic transformation rules used in steps **(b)**, **(d)** and **(f)** are implemented with regular expressions. This section provides a short introduction to this formalism.

9 The relevant lexicon entries were extracted on 12.4.2012.

Regular expressions are a means of defining patterns of strings. String pattern matching is an important problem in many areas of computing, for example in information retrieval, automated text editing and source code compilation (Aho 1990, p. 257). Pattern matching is defined as follows: given a pattern (defined in the form of a regular expression) and a string, return *True* if the pattern is found in the string, *False* otherwise.

The use of regular expressions has been popularized and standardized with the UNIX operating system, for instance with the command line tools *grep*, *(s)ed* and *(f)lex* (Hopcroft, Motwani, and Ullman 2006, pp. 109-114). These tools use a standardized notation which has been integrated in high-level programming languages like *Perl*. Table 5.2 shows the most important symbols used in regular expressions. Table 5.3 contains some commented examples.

A regular expression describes what is called a **regular language** in formal language theory. However, the syntax of regular expressions has been extended over time for practical purposes, so that the expressive power of *Perl*-compatible regular expressions exceeds the one of purely regular languages.

For example, regular expressions can also be used for **pattern replacement**: given a search pattern, a replacement pattern and a string, it substitutes the substring spanned by the search pattern with the replacement pattern. Thus, pattern replacement transforms one string into another. The replacement patterns may contain back referencing expressions corresponding to the groups defined in the search pattern. Crucially, this allows the swapping of two elements, which is not possible with purely regular languages (see examples in Table 5.4).

In the transfer rules presented below, the pattern replacement mechanism is used in the following way: each search pattern consists of a left context, a group whose content is to be replaced, and a right context. The group is delimited by parentheses.

Symbol	*Meaning*
a	the character *a*
|	either the character on the left or the character on the right of the bar
.	one arbitrary character
[]	one of the characters listed in the brackets
[^]	one character *not* listed in the brackets
()	to group a sequence of characters
?	the preceding expression is optional (i.e. repeat 0 or 1 times)
*	repeat the preceding expression 0 or more times
+	repeat the preceding expression 1 or more times
$	matches the end of the string
^	matches the beginning of the string

Table 5.2.: The most important symbols used to define regular expressions.

Regular expression	*Explanation*
a..le	accepts strings containing *apple, asyle, agkle, …*
b?all	accepts strings containing *ball* or *all*
[bf]all	accepts strings containing *ball* or *fall*
b|(fall)	accepts strings containing *b* or *fall*
*po.*t*	accepts strings containing *pot, port, pocket, …*
^[aeiou]	accepts strings starting with a vowel
^[aeiou]+$	accepts strings consisting only of vowels (at least one)

Table 5.3.: Some examples of regular expressions.

Match expression	*Replacement expression*	*Explanation*
John	*Jack*	replaces *John* by *Jack*
o+	*o*	condenses all sequences of *o* into a single *o*
([bf])all	*\1in*	replaces *ball* by *bin* and *fall* by *fin*
(.)(.)	*\2\1*	swaps two single characters

Table 5.4.: Examples of pattern replacement with regular expressions. *\1* and *\2* are back references to the groups defined with parentheses in the match expression.

phoneticVariables:

RULENAME	*REGEX*	*PRIORITY*	*DATAFOLDER*
2-120-nd	*[aeiouäöü](nd)$*	101	2-120

phoneticVariants:

RULENAME	*REPLACE*	*MAPNAME*
2-120-nd	*ng*	dp_ng
2-120-nd	*nn*	dp_nn
2-120-nd	*nt*	dp_nt
2-120-nd	*nd*	dp_nd

Table 5.5.: Detail of the database tables for phonetic transformations. The example shows the rule for final *nd*, already discussed in the previous chapter. The regular expression used to detect this pattern consists of the group to be matched *(nd)*, preceded by a vowel *[aeiouäöü]* and followed by the end of the word marker *$*.

5.2.6. Phonetic transformations

Most Swiss German words are cognates of Standard German words. Hence, regular phonetic[10] transformations suffice to derive many Swiss German word roots from their Standard German counterparts. Phonetic transformation rules are stored in two database tables, phoneticVariables and phoneticVariants.

Each phonetic transformation variable is characterized by a name (rulename), a search pattern in the form of a regular expression (regex), an integer determining the rule order (priority), and a file system path in which the corresponding maps are to be found (datafolder). Each variable has one or more variants, which are linked by the rulename attribute. Variants are distinguished by the replacement pattern (replace) and the file name of the corresponding map (mapname). Example entries are shown in Table 5.5.

The order of the rules is important, as they are applied in cascade: each rule operates on the result of the application of the previous rule. The rule set has been ordered manually. In practice,

10 As explained in Appendix A.2, we use **phonetic transformation** as a generic term for phonetic as well as phonological transformations.

no strict ordering has to be imposed, as some rules are independent of others because they operate in different phonetic contexts.

Currently, 135 phonetic variables are implemented. They correspond to 314 variants.

5.2.7. Different application contexts of phonetic transformations

The algorithm provided in Figure 5.1 mentions three contexts of use of phonetic transformation rules:

- The application of a lexical rule may trigger a manually specified subset of phonetic rules **(b)**.
- A predefined set of phonetic rules is applied on the unmodified Standard German root, if no lexical rule exists for that root **(d)**.
- A different set of phonetic rules is applied on the inflected Swiss German form **(f)**.

This rather complicated architecture can be justified on linguistic grounds. Here, we explain the relevant criteria with some examples:

- The Standard German lemma *nichts* 'nothing' yields, among others, the dialectal variants *nüüt, nüt, niit, nit* (see Table 5.1). The distribution between *ü*-variants and *i*-variants is completely regular and already accounted for in a phonetic rule. Thus, the most economic description of the phenomenon consists of the one presented in Table 5.1: a lexical rule generating *nüüt, nüt*, and the subsequent application of the phonetic rule *ü-i* to obtain *niit, nit*. On the other hand, the Standard German lemma *sonst* 'otherwise' yields the seven variants *susch, süsch, sus, süs, sundsch, schüsch, suschter*. The *ü*-variants are already dialectal and should not be transformed further to *i*. In other words, the phonetic rule *ü-i* must be blocked here. Thus, it makes sense to specify the phonetic rules in **(b)** case by case.
- The derivative suffix *-ung* is pronounced *-ig, -ing, -ung* depending on the dialect. In order to avoid overgeneration, the corresponding phonetic rule is triggered only if *ung* occurs

in word-final position. It must therefore be applied before adding the plural suffix *-e*. This means that some phonetic rules must be applied in step **(d)** and cannot be applied in step **(f)**.

- Some Wallis dialects have an infinitive ending in *-u*. Moreover, in a subset of these dialects, *u* is transformed to *ü*. However, the two rules never apply in cascade ($Inf \rightarrow u \rightarrow ü$). This means that some phonetic rules (such as the *u-ü* rule) must apply in step **(b)**, before the affixation rules of step **(c)**.

- Different dialects use different *schwa* qualities, written as *e, ä, a*, in word stems as well as in inflectional affixes. The most economic solution therefore consists of applying a single phonetic rule to all occurrences of *schwa*, after generating inflectional affixes. This rule must be located in step **(f)**, not in step **(d)**.

The number range in the priority field of the phoneticVariables table determines whether a phonetic rule is applied in step **(d)** or in step **(f)**.

Section (6) explores a different approach to ordering the different rule types.

5.2.8. Phonetic or graphemic rules?

Section 5.2.6 was called "phonetic transformations". However, what was presented there rather corresponds to graphemic rules. In other words, the transfer rules operate on orthographic strings. Implicitly, we assumed that Standard German orthography was transparent enough to approximate the phonetic structure of words. Here, we would like to discuss the legitimacy of this assumption.

In Section 3.2, the Standard German spelling conventions have been presented, and six areas of grapheme-phoneme mismatch have been identified. But only one specific kind of mismatch may have a negative impact on the definition of "phonetic" rules with regular expressions, namely, if one grapheme corresponds to two different phonemes. Notably, allophonic variation (pronunciation differences determined by the phonetic context) is not an issue as long as the

phonetic context can be modelled in the transfer rule. As a result, only two of the six mismatch areas are really problematic.

First, Standard German orthography fails to distinguish open and closed vowels. This is not a problem for Standard German, as openness is entirely determined by vowel length. In Swiss German, the situation is more complicated, but our decision not to use diacritics (as motivated in Section 3.2) saves us from bothering further about openness distinctions.

Second, vowel length is spelled inconsistently. As a result, rules that operate only on long (or on short) vowels are difficult to write. It has been suggested to use phonetic transcriptions as input instead of Standard German orthographic strings. This would considerably simplify the vowel length distinction, but it would also introduce two new issues:

- The Dieth spelling conventions for Swiss German systematically use letter doubling to represent long vowels. However, many dialect writers often use spelling conventions closer to the standard language. For instance, if a Standard German word contains the sequence *ah*, the dialect form will contain *ah* rather than *aa* to indicate vowel length. Likewise, Standard German *aa* should not end up as *ah* in Swiss German. By using phonetic transcriptions, this information would be lost.

- The Standard German grapheme *ie* not only stands for lengthening, but also signals its origin as a Middle High German diphthong that is preserved in Swiss German (see Appendix A.2.1, page 296). Thus, even if *ie* is pronounced the same as *i* or *ih* in Standard German, the Swiss German equivalents are different. Collapsing the three spellings under a single phonetic transcription therefore amounts to throwing away crucial etymological information.

To sum up, one of the six mentioned phenomena, vowel length, cannot be handled adequately with orthographic representations. However, its most straightforward alternative, phonetic transcriptions, introduces new problems. Therefore, we decided to stick to orthographic representations despite their drawbacks.

Another issue is the status of intermediate representations. Recall that the phonetic rules are applied in cascade. So, after the first rule application, some parts of the word form already show

their Swiss German appearance, while other parts are still in Standard German. This ambiguity is problematic for subsequent rules. For example, if a rule is triggered at the presence of *u*, it does not define if this *u* is thought to be a Standard German *u* or a dialectal *u* resulting from a previous transformation. In practice, this problem can be avoided by correctly ordering the rules, but there are some edge cases where this distinction would be crucial. In Section 5.3.4 (page 178), we outline a possible solution to this issue by distinguishing (lower-cased) Standard German graphemes from (upper-cased) Swiss German graphemes.

5.2.9. Lexical restrictions of phonetic transformations

Dialect evolution sometimes yields unpredictable results. Some phonetic rules apply to certain words but not to others, without there being a clearly identifiable cause. For example, the stem vowel in the words *gar* 'even', *Gras* 'grass', *sparen* 'to save', *Arbeit* 'work' and *Axt* 'axe' is changed to *ä* in some Northeastern dialects (see Figure 3.4 on page 98). It is impossible to define a phonetic context that might trigger these and only these transformations. Therefore, we chose to use a "whitelist" which enumerates all the lemmas that undergo this transformation.

In other cases, the cause of a specific evolution is known, but it is difficult to detect for practical reasons. For example, the two Middle High German vowels *û* and *ou* have fallen together in Modern Standard German *au*, but have remained distinct in Swiss German (*uu* and *au/ou*, respectively, see Appendix A.2.1). In a model based on Standard German input, it is thus impossible to predict the correct Swiss German form. No phonetic cue tells us that Standard German *Haus* 'house' should become *Huus*, but that Standard German *Baum* 'tree' should remain *Baum*. Again, we use a whitelist to enumerate the lemmas in either class.

Currently, the whitelist contains a total of 13 000 lemmas associated to 39 rules. It has been manually compiled on the basis of the *Derewo* lemma list (Institut für Deutsche Sprache, Programmbereich Korpuslinguistik 2007). For other rules, a blacklist (words that are excluded from the rule) was more practical. It contains 450 lemmas associated to 3 rules. Whitelists and blacklists only apply to phonetic rules in step **(d)**. They are stored as distinct tables in the database.

5.2.10. Affix generation

Step **(e)** deals with the generation of inflectional affixes on the basis of morphosyntactic and lexical features.

The tables affixVariables and affixVariants define the inflectional affixes for regular noun, verb, adjective and pronoun inflection. While most rules deal with simple suffixation, more complex affixation types are also supported. The simplest one is to add a suffix. In this case, the suffix column contains the suffix, and the regexFind and specialInfix columns remain empty. The second possibility allows to remove some material before adding the suffix: the material to be removed is specified by the regular expression in regexFind. If using regular expressions is not practical or not powerful enough, there is a third possibility: one can specify a particular affixation function (written in *Python*) in the field specialInfix. This functionality is used for example to add *umlaut*.

Table 5.6 shows examples of noun plural formation. Rule *n0-uml* adds umlaut to the stem vowel independently of the dialect chosen. Rules *n1-e* and *n1-er* add suffixes depending on the phonetic environment specified in regexFind. Rule *n1-ene* illustrates the use of dialect-dependent suffixes, with map information given in the fields datafolder and mapname. All rules depend on inflection class information as specified by the NCl tags (see Section 5.2.4). The special symbol ~ is used to handle *sandhi* (p. 311). It is substituted by *n* when the following word starts with a vowel, and dropped otherwise.

Currently, the database contains 82 variables and 165 variants. It covers the inflectional paradigms of adjectives, nouns, regular verbs, determiners, and preposition-determiner combinations.

For irregular inflection paradigms, it is difficult to draw a clear line between lexical transformations and affixation patterns. The following section describes an instance of this problem, the so-called short verbs.

affixVariables:

RULENAME	*TAGS*	*MORPHO*	*DATAFOLDER*
n0-uml	NN	Pl,NCl_uml,NCl_uml_er,NCl_uml_e	
n1-e	NN	Pl,NCl_e,NCl_uml_e	
n1-er	NN	Pl,NCl_uml_er	
n1-ene	NN	Pl,NCl_ene	3-187-ene

affixVariants:

RULENAME	*REGEXFIND*	*SUFFIX*	*SPECIALINFIX*	*MAPNAME*
n0-uml			umlaut	
n1-e	*([^i])e?$*	*\1e~*		
n1-er	*(er)?$*	*er*		
n1-ene	*(i)$*	*ine*		dp_ine
n1-ene	*(i)$*	*ene*		dp_ene
n1-ene	*(i)$*	*eni*		dp_eni

Table 5.6.: Detail of the affix generation tables, showing selected rules for noun plurals.

5.2.11. Suppletive morphology: short verbs

Short verbs (see Appendix (187), p. 360) are characterized by short, irregular forms and rather obscure morpheme boundaries between stem and affix. As an additional difficulty, according to SDS data, there are Northwestern dialects in which short verbs are inflected like regular verbs. Such cases of suppletive morphology can be handled with the lexical substitution tables presented in Section 5.2.3. The field feats allows to restrict the lexical substitution rules to certain inflected forms; the field addFeats allows to add specific inflection class features to apply the correct affixation rules.

The example displayed in Table 5.7 shows the different plural forms of the verb *gehen* 'to go'. The first variant generates a short stem *gö* and adds the feature KurzV, which will trigger an affixation rule common to all short verbs, yielding *göö, göi, gönd, gön* etc. The second and third variants generate long stems and add the feature RegV, which will trigger regular verb affixation rules, yielding *gange* or *gönge*.

lexicalVariables:

rulename	*lemma*	*tag*	*feats*	*datafolder*
gehen-pl	*gehen*	VVFIN	Pl,Pres,Ind	3-058-gehen-pl

lexicalVariants:

rulename	*form*	*mapname*	*addFeats*	*phonRulesAfter*
gehen-pl	*gö*	dp_gö	*KurzV*	kurzV-pl-vokal ö-e
gehen-pl	*gang*	dp_gang	*RegV*	
gehen-pl	*göng*	dp_göng	*RegV*	

Table 5.7.: Extract of the lexical substitution tables, showing the relevant entries for generating plural forms for the short verb *gehen*.

5.2.12. Rule combination

As mentioned previously, the word-level rules are applied in cascade. In other words, the output of the first rule serves as the input of the second rule, and so on. All rule variants are associated with a probabilistic maps defining their occurrence area. (Some rules only have a single variant which is valid in all of German-speaking Switzerland. In this case, the map shows constant probability of 1.0.)

More formally, we define a **derivation** as the path leading from a Standard German word to a dialect word to which no more rules can be applied. Derivations may contain rules from the different steps **(a)** to **(e)**; all rules have equal weights.

The notation $w_0 \overset{*}{\rightarrow} w_n$ represents an iterative derivation leading from a Standard German word w_0 to a dialectal word form w_n by the application of n transfer rule variants of the type $w_i \rightarrow w_{i+1}$. For each point of German-speaking Switzerland,[11] the probability of a derivation corresponds to the joint probability of the rule variants it consists of. Hence, the probability map of a derivation is defined as the **pointwise product** (**Hadamard product**) of all variant

11 We call the set of cells in the raster map covering German-speaking Switzerland *GSS*.

maps it consists of:

$$\underset{t \in GSS}{\forall} p(w_0 \overset{*}{\rightarrow} w_n \mid t) = \prod_{k=0}^{n-1} p(w_i \rightarrow w_{i+1} \mid t)$$

Note that in dialectological transition zones, there may be several valid outcomes for a given w_0 and a given t.

This formula is rather sensitive to the length of the derivation. All weights being equal, a short derivation will be preferred to a longer one. For example, a derivation consisting of 2 rules, each weighted at 0.7 at a given point, will yield a final probability of $0.7 \times 0.7 = 0.49$. In contrast, a derivation consisting of 4 rules weighted at 0.7 each will yield a final probability of 0.24. This behaviour can be changed by slightly modifying the formula. Instead of computing the product of the map values, one can compute their geometric mean:

$$\underset{t \in GSS}{\forall} p(w_0 \overset{*}{\rightarrow} w_n \mid t) = \sqrt[n]{\prod_{k=0}^{n-1} p(w_i \rightarrow w_{i+1} \mid t)}$$

The universal quantifier indicates that the probability computation is done for the whole of German-speaking Switzerland. The result of that computation is a probabilistic map which specifies the validity area of the derivation. However, one may be interested in the probability value of a derivation at a specific point $t_j \in GSS$. In this case, the computation can be simplified by using numeric multiplication rather than Hadamard multiplication in a matrix.

5.2.13. A full example

In this section, we would like to illustrate the complete word generation process with a walk-through example, starting with the Standard German word form *aufsuchen* 'to visit, to prospect'. The different steps correspond to the numbering scheme of Section 5.2.1.

Step (1) The morphological analysis of the word form *aufsuchen* yields the following readings:

- auf<PREF>suchen<+V><1><Pl><Pres><Konj>

- auf<PREF>suchen<+V><1><Pl><Pres><Ind>
- auf<PREF>suchen<+V><3><Pl><Pres><Konj>
- auf<PREF>suchen<+V><3><Pl><Pres><Ind>
- auf<PREF>suchen<+V><Inf>

To simplify the presentation, let us choose the third person plural indicative reading.

Step (2) On the basis of the above morphological analysis, the compound splitting algorithm yields two components:

- auf PTKVZ
- suchen VVFIN 3.Pl.Pres.Ind

Step (3) The two components *auf* and *suchen* are now treated separately. We start with *auf*.

Step (a) There is a lexical transformation rule for *auf* that yields four forms. Each form is valid in the region specified by the respective map:[12]

- uf PTKVZ 1-106/dp_u
- uf PTKVZ 1-106/dp_uu
- uuf PTKVZ 1-106/dp_uu
- uif PTKVZ 1-106/dp_ui

Step (b) The first three variants of step **(a)** are annotated with the phonetic rule u-ü. Its application yields seven variants:

- uf PTKVZ 1-106/dp_u; 1-152-entrundung/dp_u
- üf PTKVZ 1-106/dp_u; 1-152-entrundung/dp_ü

12 This lexical rule accounts for the fact that the long/short vowel distinction in the general monophthongization rule does not hold for prepositions and particles: *auf* is pronounced as a short *uf* even in most regions where *Haus* 'house' would be pronounced *Huus*. In the absence of more detailed map data, we simulate this behaviour on the basis of the *Haus* maps in the given lexical rule: if the map shows a short *u*, the outcome is necessarily short; if the map shows a long *uu*, the short and the long variant are generated without further geographic distinction.

- uf PTKVZ 1-106/dp_uu; 1-152-entrundung/dp_u
- üf PTKVZ 1-106/dp_uu; 1-152-entrundung/dp_ü
- uuf PTKVZ 1-106/dp_uu; 1-152-entrundung/dp_u
- üüf PTKVZ 1-106/dp_uu; 1-152-entrundung/dp_ü
- uif PTKVZ 1-106/dp_ui

The second variant (*üf*) is discarded because the two underlying maps do not overlap in any point of German-speaking Switzerland.

Step (e) Verb particles are not inflected, so no affixation rules apply.

Step (f) No phonetic rules apply to the six forms obtained above.

Step (3) The third step is now repeated for the component *suchen*. No lexical transformation rules apply in this case, so the path **(c)-(d)** is taken.

Step (c) The lemma *suchen* is shortened to the root[13] *such* and annotated with the feature RegV, which means that regular verb inflection is used (this is the default setting – all irregular verbs are handled with lexical rules).

Step (d) Three phonetic rules can be applied to *such*. The first rule, u-ue, deterministically transforms *u* to *ue* in all words whose lemma figures in the associated whitelist; *suchen* is one of them. As a deterministic transformation valid in the entire Swiss German area, it is not associated to a map:

- suech [suchen] VVFIN 3.Pres.Ind.Pl.RegV

The second rule is u-ü, already seen above. It yields two variants:

- suech [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u
- süech [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_ü

13 We define the root of a word as identical to its citation form, except for verbs, where specific infinitive affixes are stripped off.

The third phonetic rule, schwaQualDi, adjusts the schwa quality in a diphthong. It consists of three variants, *e*, *ä* and *a*. Combined with the two variants obtained above, this yields six variants:

- suech [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_e
- suäch [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_ä
- suach [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_a
- süech [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_ü; 3-001-schwa-eäa/dp_e
- süäch [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_ü; 3-001-schwa-eäa/dp_ä
- süach [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_ü; 3-001-schwa-eäa/dp_a

The *süach* variant is immediately discarded because the two relevant maps do not overlap.

Step (e) Affixation rules are applied to all five roots obtained by the phonetic rules. A single affixation rule, v-3-flex|3pl-reg, is compatible with the morphosyntactic features of the given input. It yields five potential suffixes, depending on the target dialect: *-e~*, *-end*, *-und*, *-id* and *-ed*. These five suffixes are combined with the five roots obtained above, yielding 25 forms. However, four of them are geographically incompatible and are immediately filtered out. Some of the resulting 21 forms are given below:

- sueched [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_e; 3-033/dp_ed
- sueche~ [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_e; 3-033/dp_e
- suachend [suchen] VVFIN 3.Pres.Ind.Pl.RegV 1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_a; 3-033/dp_end
- suächid [suchen] VVFIN 3.Pres.Ind.Pl.RegV

1-152-entrundung/dp_u; 3-001-schwa-eäa/dp_ä; 3-033/dp_id

- süächund [suchen] VVFIN 3.Pres.Ind.Pl.RegV
  1-152-entrundung/dp_ü; 3-001-schwa-eäa/dp_ä; 3-033/dp_und
- …

Step (f) A single phonetic rule, hasliN, applies to the forms with a *e~* ending. It transforms the *~* into a *n* in some regions. This yields 27 different dialectal forms.

Step (4) We now have five dialectal forms for *auf* and 27 dialectal forms for *suchen*. These forms are combined pairwise, and geographically incompatible derivations are discarded. At the end of the generation process, 95 dialect forms are obtained.

This high number of derivations is surprising, even in the light of the highly variable Swiss German dialect landscape. For example, the derivation *üfsuechid* occurs in a small, largely unpopulated area near Muotathal (SZ), and its maximum probability value in this area is only 0.63. In the original SDS maps, no inquiry point features this map combination. Thus, it results from the specific parameters chosen in the map interpolation process. While the goal of map interpolation was precisely to predict new feature combinations at previously unseen places, such spurious derivations should be avoided. For instance, weighting the derivations by population density could represent an effective pruning technique.

In the preceding derivation, we have generated all word forms that are valid in any point of German-speaking Switzerland. However, one often only needs to translate words into one specific dialect. By indicating the coordinates of the target dialect beforehand, derivations that are not valid at these coordinates can be eliminated immediately.

Let us illustrate this feature with the Chur (GR) dialect. For the *auf* component, the derivation *uf* is the only valid one. For the *suchen* component, the following four derivations are valid (the rightmost number represents the probability):

- suachend [suchen] VVFIN 3.Pres.Ind.Pl.RegV 0.772961
- suächend [suchen] VVFIN 3.Pres.Ind.Pl.RegV 0.137922

- suache~ [suchen] VVFIN 3.Pres.Ind.Pl.RegV 0.046941
- suechend [suchen] VVFIN 3.Pres.Ind.Pl.RegV 0.031865

Re-composing yields the following dialect forms: *ufsuachend, ufsuächend, ufsuache~, ufsuechend*.

5.2.14. Limitations of the database approach

The approach presented here relies on a database with different tables for different types of transformation rules. The rules are applied in cascade with the help of *Python* scripts. In particular, this architecture explicitly models the relation between linguistic variables and linguistic variants according to the principles of relational databases. While this approach allows for easy debugging of the rule base, it may be insufficient and impractical in certain respects. Four potential issues are worth mentioning:

- All scripts, as well as the database design, have been written with Swiss German word translation in mind. They are difficult to reuse for another language setting or another task. Moreover, our system cannot be easily compared with related work (even on such simple aspects as the number of rules) because of implementation differences.

- The regular expression notation presented above and used in our *Python* implementation is very general and rather inconvenient for certain language-specific patterns. This complicates debugging and raises issues of maintainability. For example, the regular expression matching intervocalic *nd* has to be written as follows: ([aeiouäöü])(nd)([aeiouäöü]). In other systems, one may simply define a character class Vowel and use it in the regular expression: (Vowel)(nd)(Vowel).

- The rule application mecanism is rather slow, because a lot of database queries are required to translate a single word, and because all probabilistic maps are consulted at run-time in order to prune geographically incompatible derivations immediately. Moreover, all regular expressions are applied independently of each other, requiring a lot of internal data conversion.

- The transformation rules are unidirectional: they require Standard German input and yield Swiss German output. This is due to the procedural nature of our approach and to the specific capabilities of the regular expression engine. However, it would be a great advantage if the same rules could be used in both directions. Crucially, this would allow us to annotate existing dialect data, for example for part-of-speech tagging or improved dialect identification.[14]

For these reasons, we have investigated a different approach that relies on an existing finite-state toolkit. This approach is detailed in the next section.

5.3. Finite-state implementation of word-level transfer rules

This section discusses an alternative to the implementation presented above. It is largely based on finite-state automata and finite-state transducers. We start by giving a theoretical overview of the relevant concepts and methods (Section 5.3.1). We then discuss the specific requirements for multi-dialectal coverage (5.3.2), and present the toolkit that best suits these needs (5.3.3). In 5.3.4, we present and motivate some conceptual differences in the derivation algorithm with respect to the database implementation. Some of them are directly related to the capabilities and limitations of finite-state methods, while other differences – such as the use of Middle High German data to correctly select the phonetic rules (Section 5.3.5) – are technology-independent. Finally, Section 5.3.6 gives an overview of this implementation and discusses some problems that we have encountered.

5.3.1. Finite-state methods

The theoretical basis of finite-state methods has been presented in several reference works (Wehrli 1997; Beesley and Karttunen 2003; Hopcroft, Motwani, and Ullman 2006). The following presentation is largely based on this material.

14 It is shown in Section 6.5 that dialect identification is possible with the database system, but it requires some complications that may deteriorate its performance.

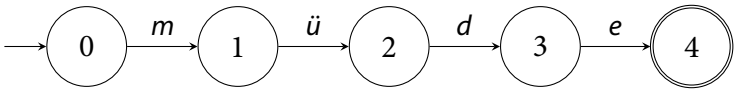


Figure 5.2.: Example of a finite-state automaton representing the Standard German word *müde* 'tired'.

Finite-state automata

In formal language theory, an **automaton** is defined as a formal specification of a set of strings (called a **language**). It consists of a set of **states** and a set of **transitions** between states. The states are graphically represented as vertices and the transitions as edges of a directed graph. If the set of states is finite, the automaton is called **finite-state automaton**. Each transition is associated with one symbol; these symbols define the strings accepted by the automaton. Among the set of states, one is distinguished as the **initial state**, and one or more states are distinguished as the **final states**. Figure 5.2 shows the graphical representation of a finite-state automaton. By convention, the initial state of an automaton is signalled with an incoming arrow, while final states are drawn with a double circle.

Formally, a finite-state automaton is a tuple

$$A = \langle Q, \Sigma, q_0, F, \delta \rangle$$

where Q is the set of states and Σ represents the alphabet, i.e. the set of allowed symbols. Both sets are finite. q_0 is the initial state (thus, $q_0 \in Q$) and F is the set of final states (thus, $F \subseteq Q$). δ is the transition function defined as follows:

$$\delta : Q \times \Sigma \times Q \rightarrow \{0, 1\}$$

For each tuple consisting of a source state, a symbol, and a target state, δ defines if such a transition exists in the graph ($\delta = 1$) or not ($\delta = 0$).

A **path** in A is a sequence of transitions $\langle d_1, \ldots, d_n \rangle$ such that $\prod_{i=1}^{n} d_i > 0$ and for each i ($1 \leq i < n$), the source state of d_{i+1} is identical to the target state of d_i. The **label** of a path is the concatenation of the symbols of the transitions defined by the path. A is said to **accept** a

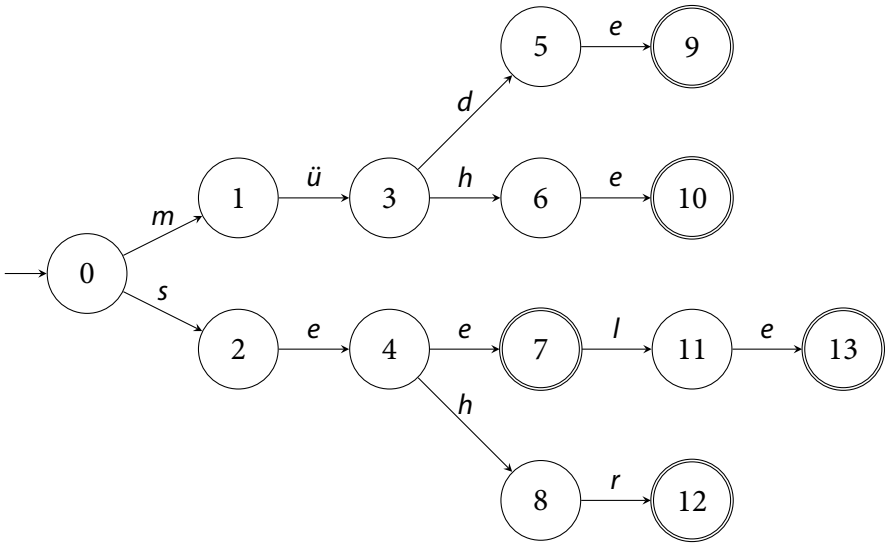


Figure 5.3.: Example of a (case-insensitive) trie for the Standard German words *müde* 'tired', *Mühe* 'trouble', *See* 'lake', *Seele* 'soul', *sehr* 'very'.

string $x \in \Sigma^*$ if there exists a path with label x from the initial state q_0 to a final state $q_f \in F$.

An automaton can be used in recognition or generation mode. In **recognition** mode, it takes as input a string, and returns a binary value: 1 if it accepts the string, 0 if it does not. In **generation** mode, the automaton returns the labels of all possible paths.

Finite-state automata are commonly used to represent words. Each letter of a given word is represented by a transition from one state to another (see Figure 5.2). The length of the word determines the number of transitions and states. Several words can be combined in the same automaton: all words start in the same initial state, but use different paths later on. If such an automaton is minimized to eliminate unnecessary states, it corresponds to a **trie** (see Figure 5.3).

All formal languages that can be defined by finite-state automata are regular languages. Thus, finite-state automata are equivalent to regular expressions in terms of their generative capacity (not taking into account the extensions made to *Perl*-compatible regular expressions).

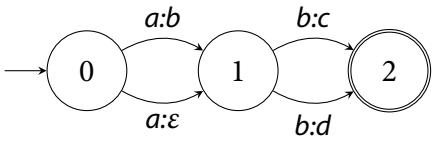


Figure 5.4.: Graphical representation of a non-deterministic transducer. The input symbol is written on the left, the output symbol on the right of the colon. This transducer transforms the sequence *ab* into the sequences *bc* (two upper transitions), *bd* (upper, then lower transition), *c* (lower, then upper transition), and *d* (two lower transitions).

Finite-state transducers

A **finite-state transducer** is an extension of a finite-state automaton: instead of just accepting or rejecting given strings, a transducer can transform a given input string into a different output string. Therefore, each transition defines two symbols – one symbol to be consumed on the input string and one symbol to be written on the output string. The set of symbols contains the special symbol ε, the **empty symbol**. Used on the input side, it does not consume a letter on the input string, allowing the input string to be shorter than the output string. Used on the output side, it does not produce a symbol on the output string, allowing the input string to be longer than the output string. Depending on the particular structure of the graph, some input strings may not be transduced at all, and some input strings may yield several possible output strings. A transducer with the latter property is **non-deterministic** (see Figure 5.4 for an example).

Compared to the definition of finite-state automata, the formal definition of finite-state transducers requires two distinct symbol sets and a modified transition function:

$$T = \langle Q, \Sigma, \Gamma, q_0, F, \delta \rangle$$

where Q is the set of states and Σ and Γ represent the input and output alphabets respectively. All sets are finite. q_0 is the initial state and F is the set of final states. The transition function δ is defined as follows:

$$\delta : Q \times (\Sigma \cup \{\varepsilon\}) \times (\Gamma \cup \{\varepsilon\}) \times Q \to \{0, 1\}$$

A path in a transducer is defined in the same way as a path in an automaton. The **input label** of a

path is the concatenation of the input symbols of the transitions defined by the path; the **output label** of a path is defined analogically. A transducer T is said to **transduce** a string $x \in \Sigma^*$ into a string $y \in \Gamma^*$ (notation: $x[T]y$) if there exists a path from the initial state q_0 to the final state $q_f \in F$ whose input label is x and whose output label is y.

Transducers are mainly used in generation mode: the transducer takes an input string and returns all possible output strings defined by its transitions. Transducers can also be used in reverse generation mode, where a string of the output side is given and a string of the input side is returned as result. In recognition mode, a transducer either accepts or rejects a given pair of strings.

In the following, we rely on finite-state transducers to implement the transfer rules. The input side of the transducers will represent Standard German strings, the output side will represent dialectal strings. In our experiments, the input and output alphabets will be identical, given our decision to use orthographic representations on both sides.

Composing finite-state machines

Transducers may be merged so that they apply in cascade, one after another. This operation is known as **composition**. Composition between two transducers is defined as follows: Given a transducer T on the alphabets Σ and Γ and a transducer R on the alphabets Γ and Δ, there exists a transducer $T \circ R$ on the alphabets Σ and Δ such that $x[T \circ R]z$ if and only if there exists a string $y \in \Gamma^*$ such that $x[T]y$ and $y[R]z$. In other words, composition allows to take the output of T as the input of R.

An automaton can be viewed as a special case of a transducer that only contains identity transitions (i.e., the input symbol is always identical to the output symbol). Accordingly, composition can also be applied if T or R is an automaton. However, if both T and R are automata, this operation is commonly called **intersection**.

We have mentioned above that transducers and automata can be used in recognition or generation mode. In fact, this distinction rather reflects two ways of using composition. In generation mode, a string a, encoded in the automaton A, and a transducer T are used to generate pairs

of target strings. The composition $A \circ T$ yields a transducer that contains all accepted string pairs $\langle a, b_i \rangle$. The input side of this transducer is not interesting: it always contains the string a. We can convert it to a finite-state automaton containing only the output strings b_i by using a **projection** operation.

In recognition mode, the transducer is merely used to determine the validity of a string pair $\langle a, b \rangle$. Again, suppose that they are encoded in two distinct finite-state automata A and B. The double composition $A \circ T \circ B$ yields a transducer that either contains the string pair $\langle a, b \rangle$ (accepted), or the empty set (rejected).

The possibility of using composition is a crucial advantage of the finite-state approach with respect to the regular-expression approach presented in the preceding section. Just as every transfer rule is implemented as a regular expression in the former approach, every transfer rule is implemented as a finite-state transducer here. However, regular expressions cannot be composed (at least not in the standard *Python* implementation), whereas finite-state transducers can. For example, all phonetic rule transducers are composed into a single transducer that describes the entire phonetic transformation process.

Weighted transducers

In a non-deterministic transducer like the one presented in Figure 5.4, all paths are considered equally likely. This may not reflect a given linguistic reality. This shortcoming can be resolved by using **weighted transducers**, where numerical values are associated with all transitions. By aggregating these transition weights, each input-output label pair is assigned a weight. These weights allow for ranking the output candidates of a given input string.

One might find weighted transducers useful for dialect word generation as pursued here: the weight could represent the probability of the transformation at a given geographical point. Indeed, this is how we proceeded in similar earlier work that was based on a single dialect (Scherrer 2007a; Scherrer 2007b). However, the multi-dialect approach pursued here would require building a unique combination of transducers for every dialect. With a target resolution of 1km² (as posited in Section 4.6.1), about 25 000 transducers would be required to cover

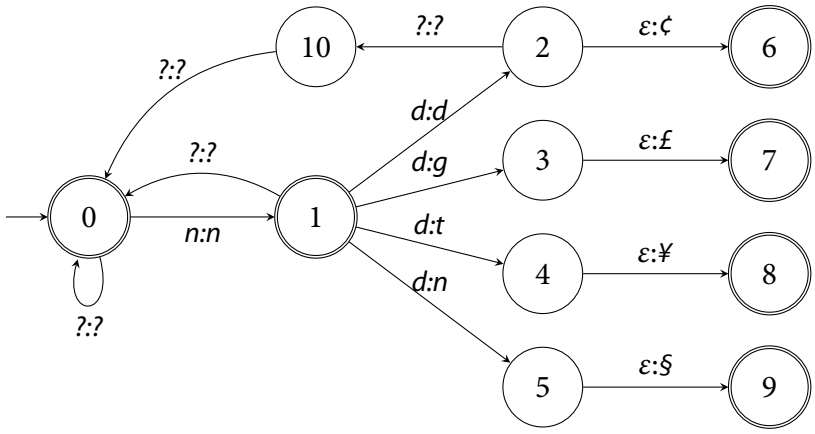


Figure 5.5.: Transducer for the final *nd* rule with added special symbols (left), and corresponding symbol dictionary (right). The transitions marked with *?:?* are default transitions; they match all symbols except those for which a distinct transition is available. These default transitions allow the transducer to accept without transformation all words that do not end in *nd*.

the Swiss German dialect area – a rather impractical solution. We therefore propose to use unweighted transducers that generate all possible dialectal variants. The elimination of unwanted dialectal variants with the help of probability maps is done in a second step, outside of the transducer. Details of this approach are discussed below.

5.3.2. Georeferencing with flag diacritics

Existing finite-state toolkits are primarily designed for creating morphological analyzers of standardized language varieties. Hence, they do not contain specific facilities for multidialectal coverage with georeferencing. As specified in the introductory paragraph of this chapter, georeferencing consists of adding the file system path of a probabilistic map to each rule variant. In the database implementation, a specific field in the database containing this information is accessed at run-time. Finite-state toolkits do not allow this type of interaction.

One solution to this issue would be to add a special symbol to each variant inside the transducer, and to rely on an external dictionary that relates these special symbols to maps at the end of

the derivation process. Figure 5.5 illustrates this approach with the rule concerning Standard German final *nd*, already used as an example in the preceding chapters. In isolation, introducing such symbols could do the trick, but it is impractical if several rules are applied in cascade. With rules applying at different positions of the word, special symbols will end up scattered all over the string. For example, if the following rule applies to word-final plosives, it will miss the relevant contexts because the special symbols ¢ and ¥ are in the way between the plosives (*d* and *t*) and the word end.

Another idea consists of adding a distinct data structure that keeps track over the transitions used in a path. At each transition, the corresponding map data would be added to that data structure. Such systems are known as **pushdown automata** or **augmented transition networks** (ATNs), but they describe context-free languages, a set of formal languages that are more complex than regular languages. These technologies are used less frequently in recent approaches to Natural Language Processing, and the number of available toolkits is restricted.

The concept of **flag diacritics** (Beesley 1998) elegantly combines the two ideas presented above. Flag diacritics are special symbols introduced into finite-state transducers. But in contrast to the symbols shown in Figure 5.5, the generation and recognition application routines treat these symbols like ε-transitions on both sides, and store the information carried by them in a separate data structure:

> When analysis and generation routines recognize and use Flag Diacritics at runtime, they are essentially cheating, adding a bit of memory to the finite-state paradigm. […] with Flag Diacritics, the application routines can "remember", during the course of application, useful bits of information in the form of feature-value settings, and the application routines can then use that information to constrain which subsequent paths are followed. (Beesley and Karttunen 2003, p. 341)

To sum up, flag diacritics are special symbols used in a finite-state transducer, which are interpreted as if they were stack operations of an ATN (Beesley 1998, p. 123). There are different operations related to them: flags can be set, unset, reset, tested, or unified. Thus, flag diacritics provide a full-fledged feature unification framework that is powerful enough to describe long-distance relationships such as those occurring in the morphology of Arabic.

Our use of flag diacritics is more modest: they just introduce information about the maps that need to be consulted in the derivation. We do not make use of the unification features.

5.3.3. The *xfst* toolkit

Several finite-state toolkits have recently been made available to the scientific community. One may cite *xfst* (Beesley and Karttunen 2003), *sfst* (Schmid 2005), *OpenFst* (Allauzen et al. 2007), and *foma* (Hulden 2009). Among these, *OpenFst* is the only toolkit that supports weighted transducers; *xfst* is the only toolkit that fully supports flag diacritics. Additionally, *hfst* (Lindén, Silfverberg, and Pirinen 2009) is a meta-toolkit for finite-state applications that can be connected to the three backends *sfst*, *foma* and *OpenFst*. All programs except *xfst* are open source. Because of its support for flag diacritics, we have chosen to use *xfst* for our work. Moreover, Beesley and Karttunen (2003) provide a good introduction to using *xfst* for linguistic engineering.

xfst provides a special syntax for defining context-dependent rewrite rules. Rules written in this formalism are then compiled into finite-state transducers. This syntax is much more convenient for phonetic and morphological transfer rules than the general-purpose syntax of regular expressions. The following example is a simplified version of the final *nd* rule:

(3) **define** ndFinal [d -> [d | g | t | n] || n _ .#.];

The keyword define is used to define a new finite-state transducer with the name ndFinal. The definition itself is enclosed in square brackets. It consists of two parts, the rewrite action on the left of the double bar, and the context definition on the right of the double bar. The rewrite action is non-deterministic: a d is substituted by any of the four variants d, g, t, n. The context definition tells us that the symbol to be matched (_, in the example d) must be preceded by n and followed by the end of the word (.#.).

This version of the rule does not specify the dialect regions in which the four variants are valid. As explained above, flag diacritics are used to this end. The expanded version of the rule looks as follows (split over several lines for better readability):

(4) **define** ndFinal [

```
    d -> [ d "@U.2-120.dp_nd@" | g "@U.2-120.dp_ng@" |
           t "@U.2-120.dp_nt@" | n "@U.2-120.dp_nn@" ]
|| n _ .#. ];
```

The expressions enclosed by the @ sign are flag diacritics and consist of three parts, separated by dots. The first part consists of one upper-case letter and determines the type of action to be executed. Throughout our work, we use the default *update* action (U) that sets a new key-value pair if the key does not exist yet, or updates the value of an existing key. The second part determines the key to be read or written, and the third part the value associated to that key. We have decided to associate the linguistic variables with the keys and the variants with the values, as illustrated in the example above. The flag diacritics are enclosed in quotation marks to treat them as single symbols. Thus, each flag diacritic corresponds to a single transition in the resulting transducer.

When composing transducers that contain flag diacritics in *xfst*, two usage patterns have to be distinguished. By default, the composition algorithm treats flag diacritics as ordinary symbols. In this case, they may occur as conditions in transformation rules, allowing feature unification. In our setting, this behavior is not desirable, and we would prefer existing flag diacritics to be invisible to the composition algorithm. This can be achieved by activating the flag-is-epsilon switch in *xfst*.

5.3.4. General architecture

Besides technical changes due to the use of the *xfst* toolkit, we have also slightly modified the general algorithm of the word generation process, with the aim of making it at the same time simpler and linguistically more accurate. Figure 5.6 shows the different steps of the process; it is to be compared with Figure 5.1 (page 142) representing the database implementation.[15] In the following paragraphs, we describe and motivate the differences between the two systems.

Two new symbols are introduced in Figure 5.6: (i) black circles represent the compilation of text-based input into a word automaton; (ii) the white circle represents the conversion from

15 In order to keep the graphs as similar as possible, there is no step **(b)** in Figure 5.6.

finite-state machines into text-based output. In other words, all steps between the black and white circles are handled completely by *xfst*. Each of them corresponds to the composition of two finite-state machines.

Derivational affixes

The morphological analyzer *Morphisto* used in step **(1)** not only splits compound words, but also detects derivational affixes. While derivation was not handled specifically in the database system, a specific marker *&* is now added between the base and the derivational affix:

(5) Wohn&ungen 'apartments'

Thanks to the presence of this marker, phonetic rules dealing with derivational affixes can be defined more precisely. For example, the three final letters of *jung* 'young' are no longer considered a derivational affix. The same behavior is simulated when using *Fips* for compound splitting.

Pre-annotation of phoneme classes

In Section 5.2.9 (page 156), we have argued that some phonetic transformations cannot be restricted by their phonetic context alone, and that their range of application should be defined by manually compiled word lists. In the database system, a separate data table containing these word lists is looked up every time a phonetic rule is applied.

The finite-state toolkit cannot interact with a database at runtime. Therefore, lexical restrictions must be incorporated in the word automaton before its composition with the rule transducers. We use special symbols to this end. For example, the special symbols *§au1a§*, *§au1b§* and *§au2§* are used to distinguish the different types of Standard German *au*-diphthongs (see Appendix A.2.1, pages 292ff.):

(6) a. Haus → *h §au1a§ s* → *h u u s*
b. Baum → *b §au2§ m* → *b o u m, b a u m*

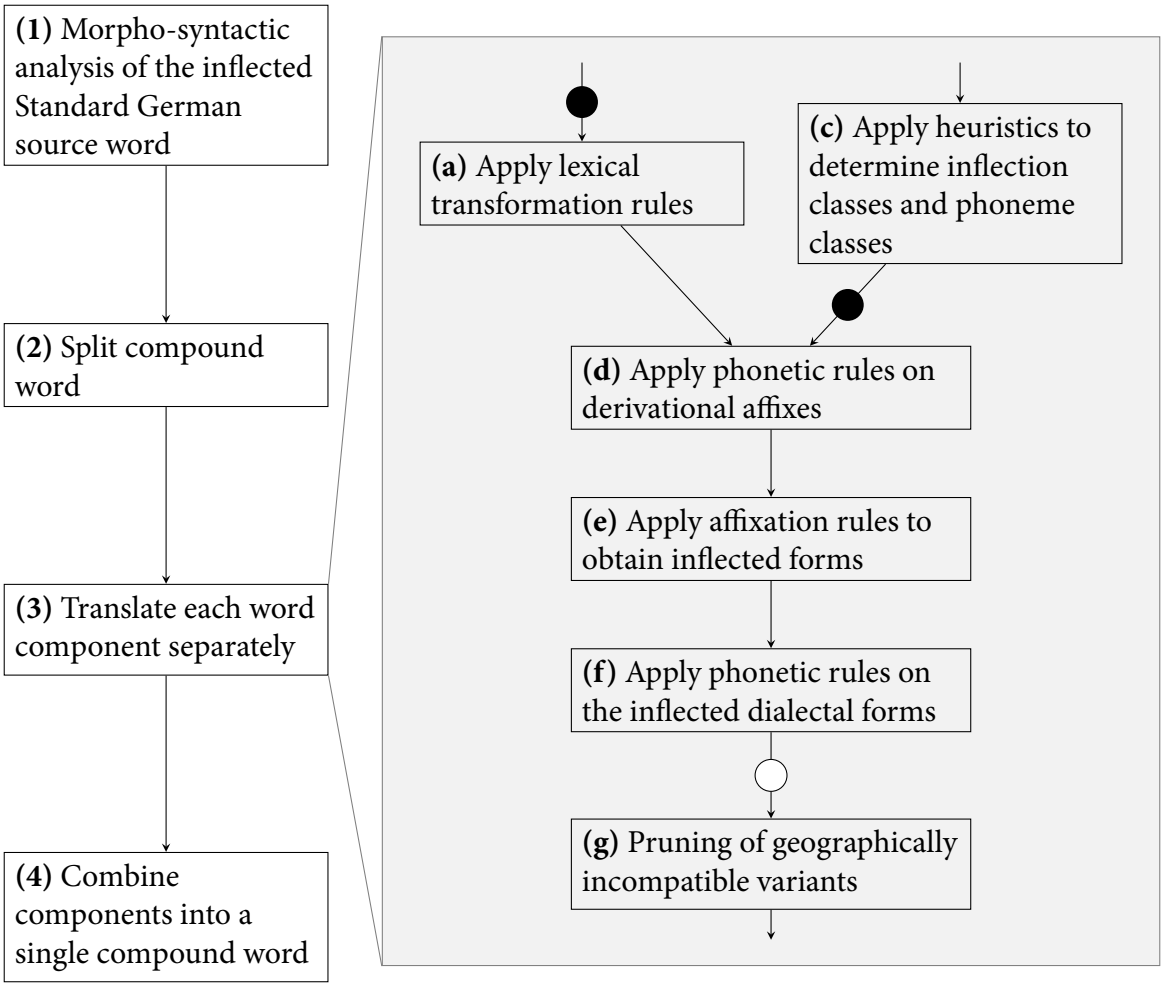


Figure 5.6.: General architecture of the dialect word generation system. The white boxes on the left hand side represent preprocessing and postprocessing steps. Steps between the black and white circles are handled by *xfst*.

xfst treats the special symbols as single characters. They are defined as vowels, so that prevocalic or postvocalic transformation rules continue to apply before or after such a special symbol.

This annotation is performed in step **(c)**. Two types of word lists are used: the manually compiled lists from *Derewo* already used in the database system, and automatically compiled lists from a Middle High German dictionary. The latter approach is described in Section 5.3.5.

Simplified application of phonetic rules

In the database system, three subsets of phonetic rules are applied at different stages of the word generation process. This complicated architecture was justified in Section 5.2.7 (page 153). Moreover, in Section 5.2.8, we have raised the issue of intermediate representations: in the course of a derivation, some graphemes already have adopted their dialectal value, while others are awaiting their transformation by a subsequent rule. Indeed, in the database system, Standard German and dialectal graphemes are not distinguished formally. In consequence, some phonetic rules have to be blocked at some stages of the derivation, which led to the complicated architecture of the database system.

Here, we propose a simple yet effective solution: lower case letters are used for graphemes that may be transformed further, while upper case letters are used for graphemes that have attained their final dialectal value and should not be modified by the following rules. At the end of the derivation process, all characters are normalized to lower case.

With this distinction, nearly all phonetic rules can be applied in step **(f)**, after applying lexical and affixation rules. This change affects two of the four examples cited in Section 5.2.7:

- The lexical rule related to *nichts* 'nothing' yields *nüt, nüüt* (with lower case *ü*), allowing the application of the *ü-i* rule and resulting in the forms *nit, niit*. In contrast, the lexical rule related to *sonst* 'otherwise' yields *sÜsch, schÜsch, …*. The *ü-i* rule is blocked by the upper case *Ü*, which prevents the generation of the unfelicitous forms *sisch, schisch, …*.

- The Wallis infinitive suffix *u* is realized as *U*. Consequently, even if the phonetic rule *u-ü* is executed after the affixation rules, it will not interfere with the infinitive suffix.

The only phonetic rules that should be applied before the affix rules, i.e., in step **(d)**, are those which deal with derivational affixes (cf. the *-ung* example on page 153). This changes the nature of the distinction between steps **(d)** and **(f)**. In the database system, phonetic rules were located by default in step **(d)**, and only a small set of rules requiring fully inflected forms was located in step **(f)**. In the finite-state system, the default location of phonetic rules is step **(f)**. Only derivational rules (which are – maybe unduly – considered a special type of phonetic rule) are located in **(d)**. In consequence, the distinction between steps **(d)** and **(f)** receives a much clearer linguistic motivation.

Pruning at the end of the derivation process

In the walk-through example of Section 5.2.13, we have insisted on the fact that geographically incompatible rule combinations are immediately discarded. This means that the word generation script must have access to the probabilistic maps at each stage of the derivation. This is no problem in the database system, but in the finite-state system, the entire derivation process is handled by *xfst*, which does not contain the necessary routines to load probabilistic maps. Instead, the maps are accessed only in step **(g)**, once the results of the transducer composition are converted into a text-based format. At that moment, derivations obtaining zero probability are discarded. In terms of performance, this does not have an adverse effect: the efficiency gains obtained by the use of composition within an optimized finite-state toolkit outweigh the efficiency loss due to delayed pruning.

5.3.5. Determining lexical restrictions from Middle High German data

In Section 5.2.9, we have argued that the Standard German orthographic representation is underspecified, and that manually compiled word lists should be used to restrain the application of some phonetic rules. One source of underspecification was related to language change, because Middle High German distinctions fell together in Modern High German, but were maintained in Swiss German. For this type of underspecification, it has been suggested[16] to

16 Michael Piotrowski and Thomas Kappeler, p.c.

infer the word list from an existing Middle High German – Modern High German dictionary instead of manually creating such a list on the basis of a dialect speaker's knowledge.

The work of Gerhard Köbler[17] provides a freely accessible electronic Modern High German – Middle High German dictionary with 54 000 entries. For example, it contains the following word pairs:

(7) a. Haus – hūs
b. Baum – boum

As can be seen, the ambiguous Modern High German diphthong *au* can be disambiguated by looking at its Middle High German counterpart. The following graphemes could be disambiguated in this way:

(8) a. au – ū / ou
b. äu, eu – iu, ū / öu, ou
c. ei – ī / ei
d. u, uh – uo / u
e. ü, üh – üe / ü
f. ie – ie / i
g. a, ah – ā / a
h. ähe – æje
i. ühe – üeje

Of course, disambiguation only works if the word pairs in the lexicon are cognates. We ensured this with the following filtering technique. Two words are considered cognates if (i) they appear on the same line (i.e. are translations of each other), (ii) the character preceding the ambiguous grapheme matches, and (iii) the character following the ambiguous grapheme matches. Matching characters are defined on the basis of the relevant orthographic conventions, so that for example a Modern High German *ß* matches with a Middle High German *z*.

Other Swiss German phenomena could not be predicted on the basis of Middle High German

17 http://www.koeblergerhard.de/germanistischewoerterbuecher/mittelhochdeutscheswoerterbuch/mhdwbhin.html, accessed on 30.8.2011.

	STEP	NUMBER OF RULES	SIZE ON DISK
Lexical rules	**(a)**	523	3.3 MB
Derivation rules	**(d)**	33	27 KB
Affixation rules (nouns)	**(e)**	50	211 KB
Affixation rules (verbs)	**(e)**	59	28 MB
Affixation rules (other)	**(e)**	80	4.0 MB
Phonetic rules	**(f)**	167	249 KB

Table 5.8.: Transducers used in the finite-state system.

data. For these, we continue to use the manually annotated *Derewo* word lists.

5.3.6. Conclusion

In this section, we have presented an implementation of the word-level transfer rules that is based on *xfst*, consisting of six transducers used in four steps. Table 5.8 shows some statistics. Due to their complexity and size, the transducers containing the affixation rules have been split into three files, which are applied one after another. The number of rules corresponds to the number of -> operations (according to the *xfst* syntax) in the rule definition files; this manner of counting is somewhat arbitrary, but the phenomena treated in these transducers are the same as in the database system.

We have encountered some problems with this finite-state approach. The first problem is technical. The most efficient way of integrating *xfst* transducers in *Python* scripts is the use of the library *xfsm*. It allows to create, load and compose finite-state machines inside *Python* scripts. However, this library does not provide an access to the flag-is-epsilon switch (Section 5.3.3). As a result, some rules are not applied correctly because flag diacritics from previous rules appear in the input. This problem can be avoided by calling the *xfst* executable from inside *Python* through a UNIX pipe, which does allow setting flag-is-epsilon. With this technique, the executable as well as the transducers have to be loaded into memory separately for each word, resulting in a drastic slowdown of the translation process. For illustration, Table 5.9 shows the times and CPU loads required for the transformation of 100 Standard German sentences, averaged over

System	*Time*	*CPU Load*
Database	0h 4min	27%
XFSM library	0h 12min	33%
XFST pipe	1h 40min	90%

Table 5.9.: Speed comparison of the different systems.

three runs. The database system is fastest, closely followed by the *xfsm* library[18]. As expected, the *xfst* pipe is much more resource-hungry.

The second problem is related to translating backwards, i.e. from Swiss German into Standard German. Indeed, one of the advantages of finite-state transducers is the possibility to use them in both directions, and this possibility was also one of our main motivations for this reimplementation. However, it turns out that the transfer rules are so ambiguous that backwards translation is simply untractable. For example, composing the dialect word *hei* '(they) have' with the phonetic transducer **(f)** and the verb-affix transducer **(e)** takes several days. Further work will show if slightly reformulated rules may alleviate this issue.

In many regards, the finite-state system is simpler and more elegant than the one using a database of regular expressions. First, the *xfst* toolkit is highly optimized, and the characteristics of finite-state machines are well-known. Second, the implementation details have been reviewed and have led to a simpler and more generic architecture. Third, the integration of additional resources (derivation splitting, Middle High German word lists) leads to a linguistically more accurate model. In this sense, it supersedes the database system. We chose to keep the description of the latter mainly for pedagogical and historical reasons. In fact, most of the experiments reported in Chapter 6 have been carried out with the database system, and we would like to present the results in a way that is consistent with the published version of the respective papers. We also believe that, through the separation of variables and variants, the architecture of the database system more closely follows the specific requirements of multi-dialectal word generation.

18 As mentioned above, the results produced by the *xfsm* library are not always correct. The performance figures are provided for comparison only.

Despite their differences, the two systems have turned out to be mostly equivalent in terms of generation accuracy, and would also be comparable in terms of speed if the *xfsm* library could be used correctly. Their performance will be compared in detail in Chapter 6.

5.4. Syntactic transfer rules

Besides the word-level differences, there are also syntactic differences between Standard German and Swiss German (see Appendix A.4). Just as the word-level transformations rely on morphologically and lexically annotated Standard German input, the syntactic transformations require syntactically annotated input text. And just as the word-level rules are triggered by a specific phonetic, lexical or morphological pattern in the source word, the syntactic rules fire at the presence of a predefined syntactic pattern. Again, the syntactic rules are georeferenced, i.e., linked to probability maps. The maps for syntactic phenomena are extracted from the SADS (Section 3.4).

In the following, we first present the annotation format that we have decided to use. Then, we illustrate the implementation of the syntactic transfer rules with an example. The coverage of these rule set has been evaluated experimentally and will be reported later on (Section 6.6).

5.4.1. The CoNLL-X annotation format

For the CoNLL-X shared task on multilingual dependency parsing, a common format has been proposed for syntactic annotation (Buchholz and Marsi 2006). This format has been used in various tasks and with various tools since then. We have chosen it again for our work.

The CoNLL-X format is a compact, easy to read, plain text format. It represents each word form on a separate line, together with part-of-speech, lemma and morphological information as well as with dependency links. The words are numbered, so that dependency links refer to their heads by word number. Thus, the CoNLL-X format represents syntactic relations with dependency links rather than constituent structures. This leads to (formally) simpler tree

ID	*FORM*	*LEMMA*	*CPOSTAG*	*POSTAG*	*FEATS*	*HEAD*	*DEPREL*
1	für	für	APPR	PREP	–	4	PP
2	eine	eine	ART	ART	Acc.Sg.Fem	3	DET
3	Statistik	Statistik	NN	N	Acc.Sg.Fem	1	PN
4	reicht	reichen	VVFIN	V	3.Sg.Pres.Ind	0	ROOT
5	das	das	PDS	PRO	Nom.Sg.Neut	4	SUBJ
6	nicht	nicht	PTKNEG	PTKNEG	–	4	ADV
7	.	.	$.	$.	–	0	ROOT

Figure 5.7.: Example of a CoNLL-X style annotated sentence. Each word (FORM) is numbered (ID), lemmatized (LEMMA), annotated with two levels of part-of-speech tags (CPOSTAG and POSTAG), annotated with morphological information (FEATS) and with dependency relations. HEAD indicates the ID of the head word, and DEPREL indicates the type of dependency relation. For example, the word at position 1 (*für*) depends on the word at position 4 (*reicht*) by a PP relation.

representations since no non-terminal categories have to be posited. This allows for easier conversion of the output of existing constituent-structure parsers based on various linguistic theories (see Section 5.4.4).

The CoNLL-X specification does not determine the part-of-speech tag set and the morphological features required in the format. Moreover, the attachment choices for certain linguistic phenomena are not universally agreed on. These decisions are left open to the treebank or parser developers. All of our work relies on a reference implementation of the CoNLL-X specification for German; it consists of a converted version of the TIGER treebank (S. Brants et al. 2002), which was created for Kübler (2008).[19]

Figure 5.7 shows an example of an annotated sentence from TIGER. For illustration purposes, the dependency links of the same sentence are reproduced graphically in Figure 5.8.

As mentioned above, we assume that the Standard German input text is syntactically annotated in the CoNLL-X format. This annotation is obtained either by selecting manually annotated Standard German sentences from the reference treebank, or by using an existing Standard

19 Thanks to Yannick Versley for making this version available to us.

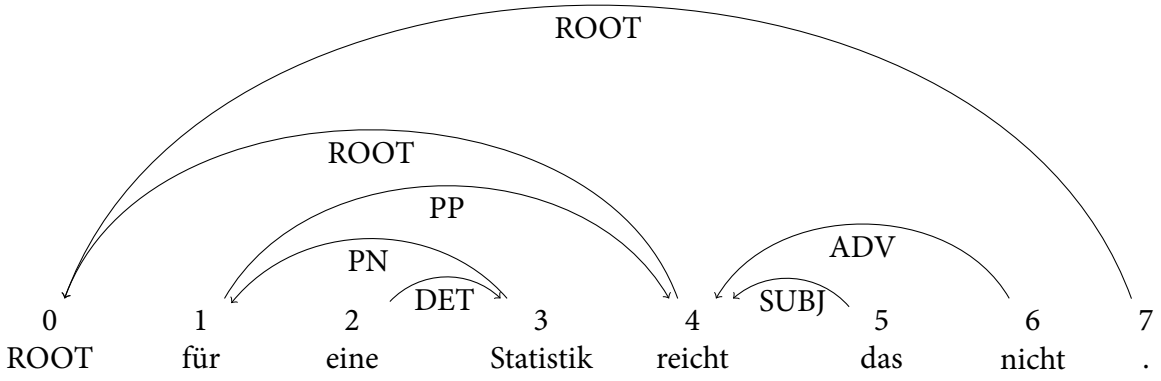


Figure 5.8.: Graphical representation of the dependency structure of the sentence displayed in Figure 5.7.

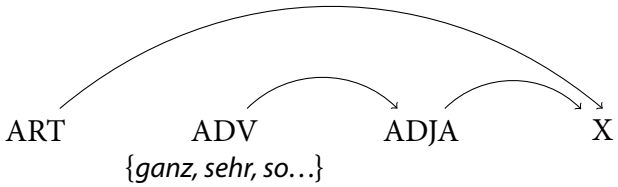
German parser whose output is consistent with the reference. While we adopt the former option for testing purposes, the transfer rules must be aimed to be sufficiently generic so that they apply correctly with the latter option.

5.4.2. Rule implementation

We have manually created transformation rules for a dozen of syntactic and morphosyntactic phenomena (the selected phenomena are listed in Table 3.2 on page 110). These rules (i) detect a specific syntactic pattern in a sentence and (ii) modify the position, content and/or dependency link of the word nodes in that pattern. They are implemented in a *Python* script.

As an example, let us describe the transformation rule for article doubling (see Appendix A.4.2, p. 343). This rule detects the following syntactic pattern:

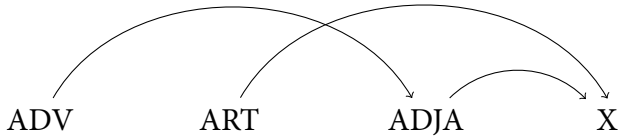
(9)



X symbolizes any type of node that possesses an article and an adjective as dependents. In practice, *X* usually is a noun.

The rule then produces three valid Swiss German patterns – as said above, the transformation rules may yield different output structures for different dialects. One of the three variants is identical to the Standard German structure produced in (9). In a second variant, the positions of the article and the adverb are exchanged without modifying the dependency links:

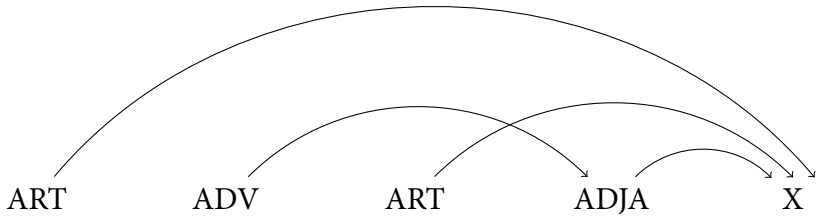
(10)



This transformation yields non-projective dependencies (i.e. crossing arcs), which are problematic for some parsing algorithms. But since the original TIGER annotation already contains non-projective dependencies, there is no additional complexity involved in the resulting Swiss German structures.

The third variant contains two occurrences of the determiner, before and after the intensity adverb. We chose to make both occurrences dependents of the same head node:

(11)



As mentioned previously, the SADS data tell us which of the three variants is accepted in which of the 384 inquiry points. Depending on the data, more than one variant may be accepted at a

given inquiry point.[20]

The syntactic transformation rules are contained in a single script file. It stores the CoNLL-X formatted sentences as objects of the TSentence class. Each transformation rule is represented as a method attached to this class. All transformation methods follow the same structure:

- A hash table links all rule variants with the corresponding SADS map files.
- The original sentence structure is copied as many times as there are rule variants.
- Each sentence structure is modified according to the transformations required by the respective variant, and annotated with the path of the map file.
- The original sentence structure is discarded, and the list of modified sentence structures is returned.

One might argue that storing the result of each variant in a separate data structure is not very efficient, and that it would be more adequate to create a single trellis in which the nodes and arcs are marked with the rule variant(s) that generated them. However, the implementation of a trellis is more difficult, and given the relatively small amount of transfer rules, we simply did not find this worth the hassle.

The syntactic transformation script contains 21 rules. Note that there is no perfect correspondence between SADS maps and rules. On the one hand, some rules – such as the replacement of preterit tense – are valid in all regions of German-speaking Switzerland and do not require maps. On the other hand, complex phenomena – such as the paraphrase of Standard German genitive case – are handled by several rules that apply to different semantic and syntactic contexts (e.g. human vs. non-human attributes). The order of the rules is chosen manually; this is a relatively straightforward task, because many rules operate in mutually exclusive syntactic contexts and do not interfere with each other.

20 As it happens, the variants of the article doubling phenomenon do not seem to define precisely delimited dialect areas. Rather, all variants seem to be accepted – albeit to varying degrees – in all regions (Steiner 2006).

5.4.3. An example

As with the word-level rules, all syntactic transformation rules are applied in cascade. The following example shows an original Standard German sentence (12a) along with three prototypical dialectal variants, obtained by the cascaded application of our transformation rules. The Mörschwil dialect (SG) shows genitive replacement[21] and relative pronoun replacement[22] (12b). The Central Swiss dialect of Sempach (LU) additionally shows prepositional dative marking[23] (12c), while the Guttannen dialect (BEO) shows an instance of verb raising[24] (12d). All segments subject to transformation are underlined. These three variants only show the syntactic transformations and do not include word-level transformations. For illustration, the last example (12e) includes word-level translations and corresponds thus to the "real" dialect spoken in Mörschwil.

(12) a. **Original:** Einen besonderen Stellenwert verdient dabei die alarmierende Zahl junger Menschen, die der PDS ihre Stimme gegeben haben.
'Special importance should be paid to the alarming number of young people who have given their vote to the PDS.'

b. **Mörschwil:** Einen besonderen Stellenwert verdient dabei die alarmierende Zahl von jungen Menschen, wo der PDS ihre Stimme gegeben haben.

c. **Sempach:** Einen besonderen Stellenwert verdient dabei die alarmierende Zahl von jungen Menschen, wo *i* der PDS ihre Stimme gegeben haben.

d. **Guttannen:** Einen besonderen Stellenwert verdient dabei die alarmierende Zahl von jungen Menschen, wo der PDS ihre Stimme haben gegeben.

e. **Mörschwil ("real"):** En bsondere Stellewert vedient debii di alarmierend Zahl vo junge Mensche, wo de PDS iri Stimm ggee hend.

21 See Appendix A.4.3, page 349.

22 See Appendix A.4.2, page 337.

23 See Appendix A.4.2, page 329.

24 See Appendix A.4.6, page 371.

5.4.4. Converting *Fips* output to the *CoNLL-X* format

The syntactic transformation rules are supposed to work with any Standard German text that is analyzed in the CoNLL-X format and consistent with the TIGER treebank annotation guidelines. Adapting the output of an existing parser to these requirements will allow us to parse any Standard German sentence and translate it into syntactically correct Swiss German.

Here, we illustrate this adaptation task with *Fips* (see also Section 5.2.2). *Fips* closely follows the theoretical framework of Government and Binding. Thus, it creates constituent structures, which must be converted to dependency graphs according to the CoNLL-X specifications. The conversion of *Fips* output to the CoNLL-X format amounts to the following steps:

- Make sure that the number of tokens per sentence is identical. For example, *Fips* analyzes compound nouns as two tokens, while TIGER considers them as one token since they are not separated by spaces.

- Normalize the word form and the lemma. This mainly concerns compounds, but also numbers or pronouns where the base form is not universally agreed on.

- Generate STTS-compatible tags for part-of-speech and morphological features (see Section 5.2.2).

- Determine the head node and the corresponding dependency relation. This is the core part of the constituent-to-dependency conversion. In most cases, the head nodes can be retrieved by going up to the next lexically realized mother node, starting at the trace of the node if it has one. However, the linguistic interpretation of some phenomena differs between TIGER and *Fips*:

  Noun phrases *Fips* follows the DP-hypothesis, according to which the head of a noun phrase is the determiner, not the noun (hence the name of DP, or determiner phrase). In TIGER, the determiners are dependent of the nouns.

  Coordinations In *Fips*, the coordination particle is the root of a subtree consisting of the conjoined phrases. In TIGER, the first conjunct is the head of the coordination

particle, which in turn is the head of the second conjunct (see also Figure 6.7 on page 236).

Subordinate clauses The dependency chaining between the matrix clause verb, the subordination conjunction and the embedded clause verb differs as well.

In these cases, we first create a sentence table with the *Fips*-type dependencies. Once the table is filled for the entire sentence, the abovementioned patterns are detected, and the dependency links are reordered.

5.4.5. Conclusion

In this section, we have shown how Standard German sentences are syntactically transformed into Swiss German. The syntactic transfer rules require fully parsed sentences according to the CoNLL-X format specification. It might be argued that this requirement is too costly given the close relationship between Standard German and Swiss German, and that simpler solutions based on surface pattern matching should be sufficient. This is certainly true for those rules that cover small phrases with a relatively fixed internal structure. The article doubling rule presented in Section 5.4.3 probably falls in this category. In contrast, rules such as verb (projection) raising or preterit tense replacement trigger relatively complicated reordering mechanisms.

If recent results in statistical machine translation can be used as evidence, surface string methods are insufficient in such cases. Word reordering based on surface word strings is commonly used in SMT, and has proven sufficient for local reorderings such as the change from pre-nominal to post-nominal adjective placement. However, it is largely unable to handle structural differences over larger spans of text, such as moving English verbs in subordinate clauses to the German sentence-final position (Koehn 2010, 142ff). This is one of the main reasons why the performance of English-German SMT systems still lags behind those of English-French or English-Spanish systems, which require fewer long-distance reorderings (for an overview of recent state-of-the art results on these language pairs, see for example Callison-Burch, Koehn, Monz, and Zaidan 2011, Tables 44-46).

5.5. Data structures

To conclude this chapter, we present the data structures that are used during the sentence transformation process. These structures aim at representing words and sentences in a way that is compatible with the CoNLL-X format and that supports the integration of word-level and syntactic transformation rules into a single system. They work invariably with the database architecture (depicted in Figure 5.1, page 142) and with the finite-state architecture (depicted in Figure 5.6, page 177).

Figure 5.9 contains a class diagram of the data structures; it also lists the most important attributes and methods of each class.

Word objects represent simplex word forms throughout the word-level transformation process. Each **Word** objects has different attributes for morphosyntactic annotations, but also some attributes related to the transformation process: *trace* keeps track of all applied rules, *prob* contains the probability of the derivation if a single coordinate pair has been selected, and *map* contains all probabilities if no coordinates have been specified.

WordSet is a list that contains items of type **Word**. During the transformation process, the rules may generate several alternative forms; each form is stored in a distinct **Word** object and added to the **WordSet**.

WordList is a list that contains items of type **WordSet**. For simplex words, it contains a single **WordSet** item, but for compound nouns, a distinct **WordSet** item is created for every component and added to the **WordList**. This additional layer allows each component to be transformed independently.

The **Node** and **Sentence** classes are used for the syntactic transformations. A **Node** contains the actual linguistic material in form of a **WordList** object, and adds syntactic information like word number, number of the head word, and the dependency relation between itself and its head word.

Sentence is a list that contains items of type **Node**; it represents the syntactic structure of a

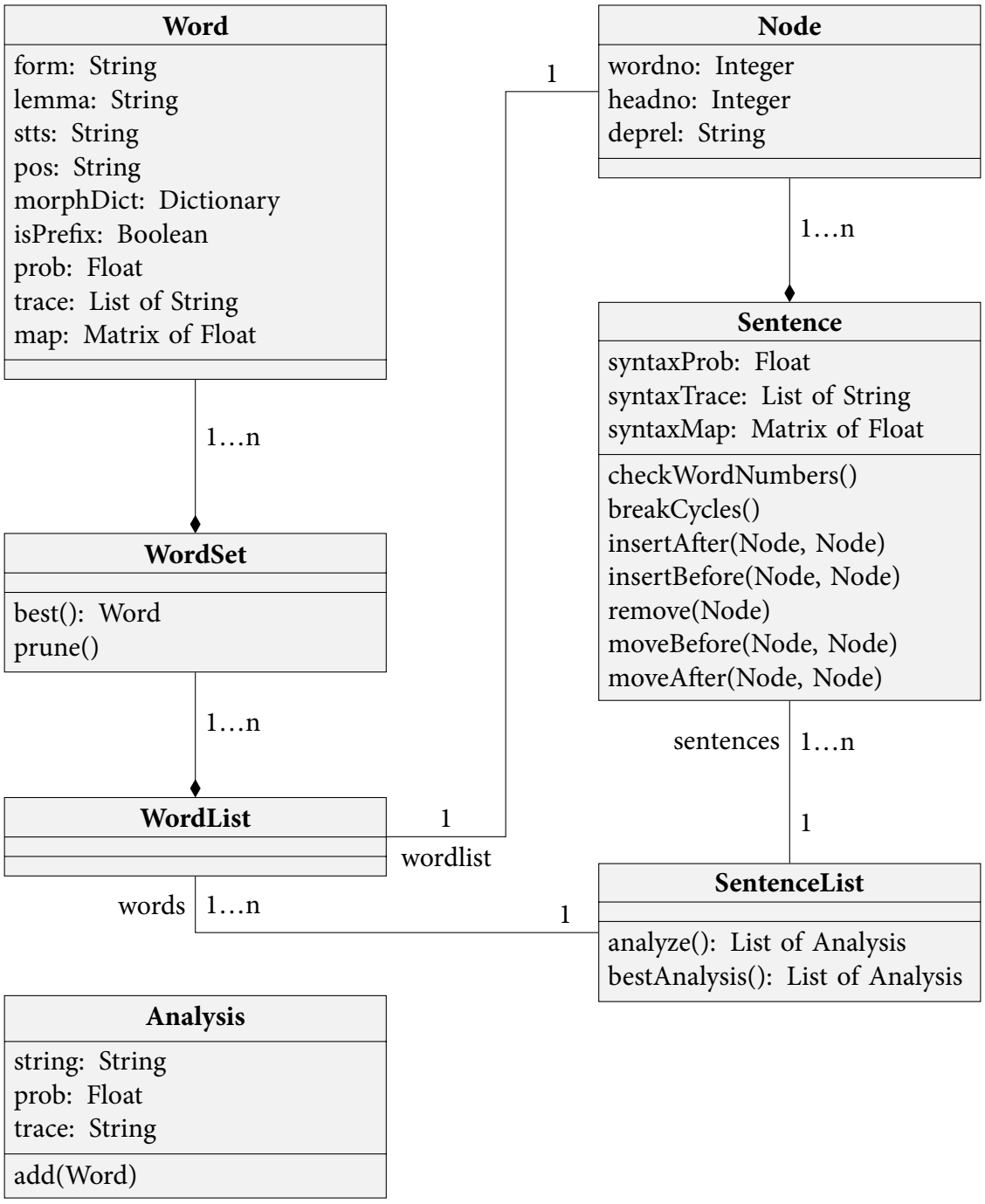


Figure 5.9.: Diagram of the data structures and their attributes and methods.

sentence. During the transformation process, the syntactic rules may generate several word orders, each of which is represented by a distinct **Sentence** object containing a set of distinct **Node** objects. However, the associated **WordList** objects are kept in place. **Sentence** objects also keep track of the applied syntactic rules and of the resulting probabilities, by means of the attributes *syntaxProb*, *syntaxTrace* and *syntaxMap*. Furthermore, the **Sentence** class defines some methods that are used by the syntactic transformations to insert, remove and move **Node** objects. The *checkWordNumbers()* and *breakCycles()* methods ensure that the syntactic structures are well formed.

The **SentenceList** class encompasses everything that results from the transformation of a Standard German source sentence: the *sentences* attribute contains all syntactic configurations obtained by the application of the syntactic transformation rules, while the *words* attribute contains all word forms obtained by the application of the word-level transformation rules. Because of the multiple levels of ambiguity, it is no easy task to retrieve the different derivations with their probabilities. The methods *analyze()* and *bestAnalysis()* traverse the entire data structure and create simple **Analysis** objects with the respective text string and probability. *analyze()* yields all analyses, while *bestAnalysis()* only yields the analysis/analyses with the highest probability.

The **Analysis** class represents one derivation from the first to the last word as a simple string. Its probability and rule identificators (*trace*) are stored as well.

All classes have special print routines to display their content. For instance, the print routine of the **Sentence** class displays a sentence in the CoNLL-X format.

5.6. Summary

In this chapter, we have introduced the concept of georeferenced transfer rule, and presented different implementations of it. We have first focused on the word-level transformations, which we dealt with in two alternative systems. Then, we have shown how syntactic patterns can be transformed. Finally, we have outlined the data structures underlying the different transformation processes.

Like the web-accessible digitized maps (see Section 4.7), the word-level rules can be used interactively on the Internet.[25] Several types of applications are supported:

Word-by-word translation Here, the user provides a Standard German text and the coordinates of the target dialect. The program calls *TreeTagger* or *Fips* to automatically annotate the text, and then translates each word independently into the specified dialect. All variants that are valid in that dialect are displayed. Depending on the length of the text, the execution time of this script can be rather long.

Single word translation In the previous application, *TreeTagger* or *Fips* automatically chooses an analysis for each source word. Sometimes, this choice is wrong. This interface allows the user to correct these choices: (s)he provides one Standard German word, *TreeTagger* yields a list of possible morpho-syntactic analyses, the user selects one of them, and the translation script uses this information to translate the word into the specified dialect.

Morphology generation In this web application, the user indicates a Standard German lemma and selects a set of morphosyntactic tags from a list. The translation program uses this information to generate the inflected word forms of five predefined Swiss German dialects. This is a web interface of the experiment presented in Section 6.3.

Dialect word lookup This and the following application use Swiss German dialect input. Since our systems do not support backwards translation, this behavior is simulated by compiling a bilingual multidialectal dictionary using Standard German to Swiss German translation, and then retrieving dialect words in this dictionary (see Chapter 6 for further details). This word lookup tool expects a word from any dialect; if it is found in the dictionary, its area of occurrence is computed and displayed on a map.

Dialect identification This application extends the previous one by accepting entire sentences of dialect text as input. The maps of all words are combined to create a single occurrence map for the sentence. This is a simplified web interface of the experiments reported in Section 6.5.

25 All web applications are accessible on http://latlntic.unige.ch/~scherrey/.

These web applications are ideal for a casual use of the dialect word translation system, but they do not provide a reliable quantitative assessment of the translation quality. This is the object of the next chapter, where several aspects of the translation system will be evaluated numerically on the basis of a multi-dialectal corpus.

6. Experiments

6.1. Introduction

This chapter gathers several experiments that have been conducted over the last years to assess the different components of the dialect translation system. We start by presenting the corpus used in most of these experiments (Section 6.2). In Section 6.3, the coverage of the word-level transfer rules is evaluated on the basis of word lists from five dialects. In Section 6.4, we use the same word-level transfer rules to translate entire sentences; this task offers a more reliable picture of the effective translation performance. In Section 6.5, we illustrate how the geographical information contained in the word-level rules can be used to identify the dialectal origin of a text. Section 6.6 presents the development process of the syntactic rules. Finally, we repeat the experiments of Section 6.4 after integrating the syntactic rules in the translation system (in Section 6.7).

6.2. The *Wikipedia* corpus

In order to evaluate our translation system, its output must be compared with data that is known (or at least assumed) to be correct – with a **gold standard**. Ideally, such data should consist of words or sentences in Standard German and their equivalents in multiple Swiss German dialects. Moreover, it should be indicated in which dialect(s) the texts are written. To our knowledge, the only source that fulfills all these requirements is the *Gespräch am Neujahrstag*, a parallel text of about 20 sentences first transcribed in 24 dialects (Phonogrammarchiv der Universität Zürich

Figure 6.1.: Screenshot of a Swiss German *Wikipedia* article. The small box on the upper right contains the annotation *Dialäkt: Bärndütsch* 'Bern dialect'.

1952) and later extended to 47 additional dialects (Hotzenköcherle and Brunner 1972-1976). The text is designed to reflect a large number of dialectal phenomena contained in the SDS. Due to its small size, its particular content, and its lack of availability in electronic format, we refrained from using this text. Instead, we have collected data that partially fits our needs and have completed them with further annotations.

Our text corpus is extracted from the Alemannic *Wikipedia*[1]. In contrast to most other sources of dialect data, *Wikipedia* texts are produced in written form; they are not transcriptions of speech data, and do not contain incomplete sentences, hesitations and false starts. Using such clean data is especially important if it is intended to be syntactically analyzed. On the other hand, noisier data could give us clearer indications about the robustness of the model.

The Alemannic *Wikipedia* allows authors to write articles in any dialect. Some of the articles were tagged with dialect categories directly by the *Wikipedia* authors (see Figure 6.1). Eight

1 http://als.wikipedia.org; the corpus was extracted on 18.3.2010.

Ziitskale	3	Zeitskalen
blibe	3	geblieben bleiben
Papscht	3	Papst
Gegesatz	3	Gegensatz
Sunne	3	Sonne
Di	3	Die
Im	3	Im
dene	3	diesen
wichtig	3	wichtig wichtige
Römer	3	Römer
la	3	lassen

Figure 6.2.: Excerpt of the translated word list for Zürich dialect, with dialect words on the left, frequency counts in the center, and Standard German translations on the right.

Di drei räschtleche Mitglider hei bschlosse, aus Drüergschpann witerzmache.
Die drei restlichen Mitglieder beschlossen, als Dreiergespann weiterzumachen.
'The three remaining members decided to continue as a threesome.'

Figure 6.3.: An original Bern dialect sentence (above) and its Standard German translation (below).

dialect categories contained more than 10 articles. We selected five of them for our corpus: Basel (BA), Bern (BE), Eastern Switzerland (OS), Wallis (WS), and Zürich (ZH). We randomly extracted 100 sentences from different articles for a development set, and 100 sentences for a test set. For the dialect identification experiment (Section 6.5), we added data from a sixth dialect, *Seislertütsch* (FR). Data for this dialect was scarce, so that we used 50 sentences per set.

We then manually translated these texts back to Standard German. Two types of translations were made. First, the text was split into a word list, and each word of the list was translated separately. If a word was ambiguous, all its possible translations were retained. These word lists are used in Section 6.3 (an example is shown in Figure 6.2). Second, the sentences were translated completely and adapted to Standard German syntax and morphology. These translations are used in Section 6.4 (see Figure 6.3 for an example).

It might have been more straightforward to choose an existing Standard German corpus and have it translated into several dialects by native speakers. However, dialect text created by translation would likely have been influenced by Standard German spelling and syntax. Our choice of using Swiss German originals and translating them back to Standard German hopefully diminishes this risk.[2]

6.3. Coverage of the word-level transfer rules

In this section, we report the results of an experiment intended to determine the coverage of the word-level transformation rules.[3] In this experiment, we start by deriving a morphological dictionary that relates Standard German word forms to dialectal word forms. This step relies on the word-level transfer rules and is described in detail in Section 6.3.1. We then count the proportion of words in a given dialect text that are present in the morphological dictionary. Sections 6.3.2, 6.3.3 and 6.3.4 present and discuss the results of several variants of the experiment.

6.3.1. Creation of the morphological dictionary

The dialect words contained in the test corpus are analyzed using a **morphological dictionary**. Generally, the entries of a morphological dictionary consist of an inflected word form, its lemma and the corresponding morpho-syntactic features. We extend this definition to introduce a **cross-lingual multi-dialectal morphological dictionary** (cf. Section 5.2), whose entries contain (i) a dialect word form, (ii) its Standard German lemma, (iii) its morpho-syntactic features, (iv) the map showing in which regions the dialect form is valid. Such a dictionary can be derived with the help of two resources: a Standard German word list, and the word-level transfer rules.

2 Some of the Swiss German *Wikipedia* entries are more or less literal translations from the Standard German *Wikipedia* entries, so these influences cannot be entirely excluded. Such texts could be discarded by a more thorough data selection, but this would also decrease the overall corpus size.

3 This section draws on Scherrer (2011a).

The lemmas (ii) and the morpho-syntactic features (iii) are extracted from a large-coverage Standard German word list, consisting of the leaf nodes of the TIGER treebank (S. Brants et al. 2002). These nodes are already lemmatized, tagged and morphologically disambiguated. We discarded *hapax legomena*, i.e., words with one single occurrence in the TIGER treebank, as well as forms that contained the genitive case or preterit tense attribute, since the corresponding grammatical categories do not exist in Swiss German. The resulting list contains about 36 000 Standard German entries.

The dialect forms are derived by applying the word-level transfer rules to each item of the Standard German word list (consisting of a lemma and a set of features). The rule application is not restrained by the indication of target coordinates, so that all variants valid in any Swiss dialect region are generated. Each derived dialectal word is associated with a map that specifies its regional distribution.[4] Named entities and words tagged as "foreign material" were not transformed. Three morphological dictionaries have been generated: one with the database implementation of the word-level rules (see Section 6.3.2), and two with the finite-state implementation (see Section 6.3.3). The derived dictionary contains 440 000 entries if generated with the database system, and about 420 000 entries when generated with the finite-state system.

At test time, each word appearing in the multi-dialect corpus is looked up in one of these dictionaries, and traced back to the Standard German word from which it was derived. The underlying idea is that the coverage of the morphological dictionaries is a good proxy for the coverage of the transfer rules themselves.

Such morphological dictionaries are useful for different tasks. In one experiment[5], one of them was used to semi-automatically normalize the Archimob dialect corpus (see Section 7.3.5). The goal was to annotate each dialect word with its Standard German counterpart in order to provide better searchability of the corpus. The morphological dictionary has also served as a complementary resource for training a part-of-speech tagger for Swiss German (Aepli and Hollenstein 2012).

4 Technically, we do not store the probability map, but the sequence of rule variants involved in the derivation. The probability map is restored from this rule sequence at test time.

5 Anne Göhring, University of Zurich, p.c.

6.3.2. Evaluation

When trying to look up a dialect word in the morphological dictionary, five cases may arise:

1. The dialect word is not available in the dictionary because the Standard German word it should have been derived from was not contained in the TIGER word list.

2. The dialect word is not available in the dictionary, although the Standard German word it derives from is contained in the TIGER word list.

3. The dialect word is available in the dictionary, but it has been derived from an incorrect Standard German word.

4. The dialect word has been correctly derived from a Standard German word, but it is predicted to be valid in a dialect area that does not overlap with the real origin of the word. For example, our transfer rules may predict that a derivation occurs in region A, while the corpus shows us an occurrence of this word in region B (where $A \cap B = \emptyset$).

5. The dialect word has been correctly derived from a Standard German word and its area prediction matches the real origin of the word.

Throughout this experiment, we distinguish between **word types** and **word tokens**: a word form occurring three times in the corpus counts as one word type, but as three word tokens. Our model shows consistently higher token coverage than type coverage. This means that on average, frequently occurring words are better covered than rarely occurring words.

We first obtained coverage figures without looking at the maps associated with the derived dialect words (i.e., counting cases 4 and 5 as correct). In this scenario, when a Basel dialect word is analyzed, we also accept derivations that are only valid in the region of Bern. The results are presented on the left of Table 6.1. Except for the notoriously difficult Wallis dialect, the figures are fairly consistent across dialects: about 40% of word types and about 60% of word tokens are analyzed correctly.[6]

6 These figures are recall values. We did not compute precision: since several forms may be valid in a dialect, low precision is not necessarily a bad thing.

	Without geo-filtering					With geo-filtering				
	BA	BE	OS	WS	ZH	BA	BE	OS	WS	ZH
Types	42%	40%	41%	25%	45%	37%	29%	27%	17%	40%
Tokens	62%	57%	60%	44%	65%	57%	47%	44%	30%	58%

Table 6.1.: Percentages of correctly analyzed dialect words, obtained with the database system.

The second scenario involved **geographical filtering**. This means that only analyses that obtained a minimal probability of 10% in the most representative city of the respective dialect area are considered correct (in other words, we count case 5 as correct, but not case 4). As representative cities, we chose the city of Basel for BA, the city of Bern for BE, St. Gallen for OS, Brig for WS, and Zürich for ZH. Results are given on the right of Table 6.1. With respect to the first scenario, there is only a slight performance drop (about 5% for types as well as tokens) for Basel and Zürich dialects, but the three other dialects show performance drops ranging from 8% to 16%. The latter regions are larger and show more internal dialect variation than the former. In addition, the *Wikipedia* authors of the three latter regions probably use a dialect that diverges from the reference city dialect chosen for our evaluation.

The approach explained above, based on a Swiss German morphological dictionary derived from a Standard German word list, has a serious drawback: it only recognizes dialect words whose Standard German counterparts occur in the word list. As a result, many compound nouns and proper nouns are not recognized, even if the transformation rules would derive them correctly (case 1 above). Thanks to the Standard German translations available in the *Wikipedia* corpus, we can quantify the impact of this case: the upper bound of our system lies at about 70% of word types and about 80% of word tokens.

6.3.3. Comparison of the word-level implementations

The results mentioned in Table 6.1 are obtained with a morphological dictionary created with the database implementation. We derived two other dictionaries using the finite-state implementation: the first one uses the *Derewo* word lists like the database system (as explained in Section 5.2.9), while the second one uses word lists automatically inferred from Middle High

	Without geo-filtering					*With geo-filtering*				
	BA	***BE***	***OS***	***WS***	***ZH***	***BA***	***BE***	***OS***	***WS***	***ZH***
Types	44%	42%	42%	25%	48%	39%	29%	26%	16%	42%
Tokens	62%	59%	61%	43%	65%	57%	44%	39%	28%	56%

Table 6.2.: Percentages of correctly analyzed dialect words, obtained with the finite-state system and *Derewo* word lists.

	Without geo-filtering					*With geo-filtering*				
	BA	***BE***	***OS***	***WS***	***ZH***	***BA***	***BE***	***OS***	***WS***	***ZH***
Types	44%	42%	41%	25%	47%	39%	29%	26%	16%	42%
Tokens	62%	59%	60%	43%	65%	56%	45%	38%	28%	56%

Table 6.3.: Percentages of correctly analyzed dialect words, obtained with the finite-state system and Middle High German word lists.

German data (see Section 5.3.5). In this way, the impact of the implementation and of the word list type can be assessed independently. The results of these two experiments are shown in Tables 6.2 and 6.3 respectively. The performance of all three systems is similar. The finite-state systems perform slightly better without geo-filtering, but slightly worse with geo-filtering. Using Middle High German word lists again marginally deteriorates the results. It should be noted that the finite-state implementation is significantly faster: generating a morphological dictionary takes about 1 day, compared to about 3 days with the database system (both tasks could be sped up by using parallelization).

The reason for the decrease in geo-filtered recall values lies in a single linguistic phenomenon, namely the vowel opening rules *u* → *o* and *i* → *e*. Our system predicts open vowel values (*o, e*) for BE and OS dialects, while the texts contain closed values (*u, i*) by and large. The closed values exist in other dialects and are generated, which means that the problem only concerns geo-filtered results. In contrast to the database implementation, the finite-state approach applies this phonetic rule also to some high-frequency words that have been previously generated by lexical rules, such as *und* → *ond* 'and' or *isch* → *esch* 'is', thereby exacerbating the number of errors especially in token counts. We return to this issue in Section 6.4.4.

The performance decrease related to the use of Middle High German word lists is due to their more restricted coverage, and to their lack of precision. For example, the Standard German word *Tat* 'deed' corresponds to Middle High German *tāt*, which is transformed to (Northern) Swiss German *tot*, which dialectally is not accurate.

Other types of errors are common to all three systems. These are discussed below.

6.3.4. Error analysis

Several types of errors have been identified. First, some errors are due to different **spelling choices**, which is expected given the lack of binding spelling rules for Swiss German dialects. For example, our system generated *bestaat* 'consists' for Zürich dialect, while the *Wikipedia* corpus contained *bestaht*. Both variants are pronounced identically; the former conforms to the Dieth spelling guidelines, while the latter is closer to Standard German spelling. We found that the *Wikipedia* authors prefer a spelling closer to Standard German especially for long, complex words. Further examples of this issue are reproduced below (Standard German spelling on the left of the bar, *Wikipedia* spelling on the left of the slash, generated dictionary entry on the right of the slash):

(1) a. hat – hät / häd (ZH) 'has'
b. Eidgenossenschaft – Eidgenosseschaft / Aidgenosseschaft (ZH) 'confederation'
c. Personen – Persone / Pärsone (BE) 'persons'
d. verschiedenen – verschiedene / verschidene (ZH) 'different'

Other errors arise from **missing rule coverage**. For instance, Standard German *Kirche* 'church' is phonetically transformed to *Chirche*, while a lexical transformation should be used to obtain *Chile* in Zürich dialect. Likewise, some specific inflectional affixes for Wallis dialect have not been implemented, which partially explains the lower overall scores for this dialect.

Another type of error is due to **diachronic change**. Standard German *zeigt* 'shows' yields the three forms *zäägt, zaagt, zeigt* in Eastern Swiss German. While all of these forms have been used in that region in the 1950s (at the time of the SDS inquiries), they have become marginal today.

The most frequently used version now (at least in the reference city, St. Gallen) is *zaigt*, which is indeed what we find in the *Wikipedia* texts.

Finally, errors may arise from the **imprecise dialect annotation** in the *Wikipedia* corpus. An article annotated as written in Basel dialect may be written either in the dialect of the city of Basel or in one of the Baselland dialects. Since we consistently generated the dialect of a reference city, there may be mismatches. For example, the BA corpus contains *hai* for Standard German *haben* 'have', where our system predicts *hän* or *häi*. This difference probably arises from the author's use of a different dialect than the reference city dialect of Basel.

To sum up, the error analysis reveals some issues that have been identified since the beginning of our work. Accommodating all patterns of spelling variation while keeping dialectal precision is a challenge. The obsolescence of the SDS material is also a hindrance that we cannot avoid as long as there is no recent dialect survey that achieves the systematicity and completeness of the SDS. Finally, the limited number of lexical rules results in coverage gaps; unfortunately, adding lexical rules is a costly task, and the relevant map information may not always be available.

In conclusion, the results show that fair lexical and morphological coverage is obtained with both implementations of the transfer rules. However, the fact that geo-filtering diminishes the recall values quite drastically means that some of the georeferenced rules hurt more than they help. Reviewing the effects of each rule separately might help us identify the most blatant mismatches. An additional bottleneck is the relatively small size of the Standard German word list. This can be remedied by using a larger resource. Preliminary experiments have been conducted with the *Text+Berg digital* corpus (Volk et al. 2010). Another possibility is to forgo using morphological dictionaries altogether; this approach is pursued in the next section.

6.4. Translation of entire sentences

The evaluation task reported here pursues two goals. First, we put forward a metric that allows to measure the similarity between a text and its translation in another language or dialect. We claim that linguistic similarity is a good estimator of the ease of translating from one variety to the other. In the second part, we apply this measure to compare automatically generated dialect

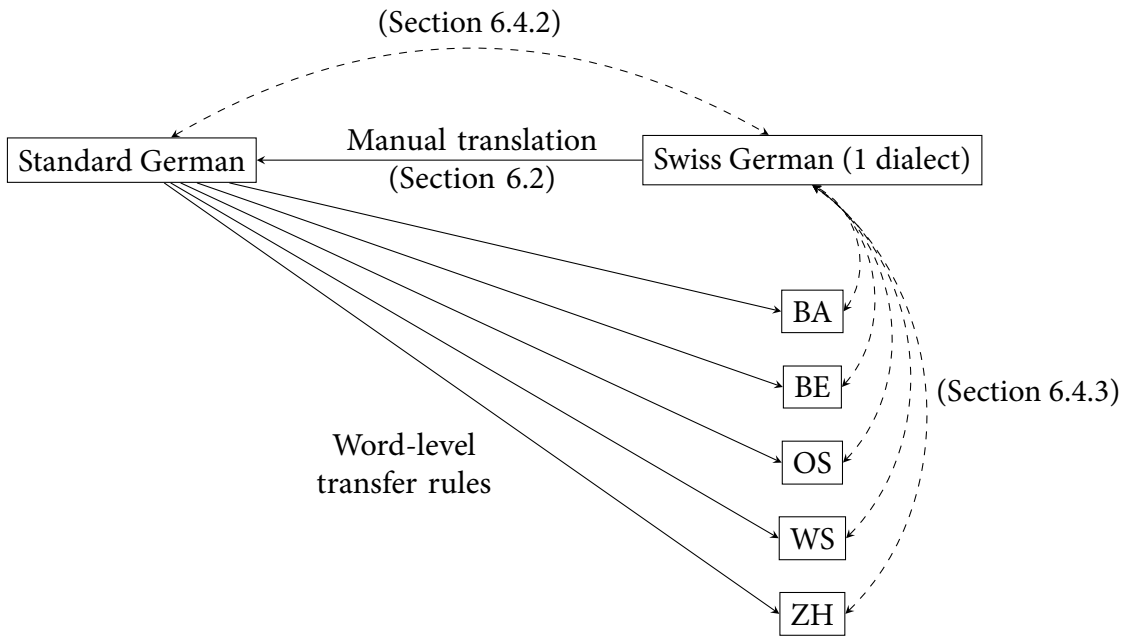


Figure 6.4.: Flowchart of the sentence translation experiment. Plain lines represent translations of texts made either manually or by the system. Dashed lines represent comparisons of two texts for evaluation purposes.

texts with original dialect texts. Throughout this section, we use the *Wikipedia* corpus with its sentence-by-sentence translations.

Figure 6.4 contains a schematic representation of the experiments reported here. The two boxes above represent the data of the *Wikipedia* corpus: Swiss German sentences in one of five dialects and their Standard German translations. All dashed lines represent comparisons carried out by computing similarity values; the respective similarity measure is presented in Section 6.4.1. Section 6.4.2 deals with the dashed line between the two upper boxes. Its goal is to assess the difficulty of translating into the different dialects. In Section 6.4.3, we use the word-level transfer rules to translate the Standard German texts into all five dialects (boxes below), and compare them with the original Swiss German text (dashed lines on the right).

6.4.1. Measures of linguistic distance and similarity

In a recent strand of dialectometric research, dialect differences are determined on the basis of the graphemic or phonetic differences between words from different dialects (see Chapters 2.5, 2.6, 7.3.5). One of the most common metrics used in this context is **Levenshtein distance** or **string edit distance** (Levenshtein 1966; Nerbonne, Heeringa, and Kleiweg 1999). Levenshtein distance is defined as the smallest number of insertion, deletion and substitution operations required to transform one string into another.

(2)

b	i	i	s	c	h	p	i		u
b	i		s	c	h	p	i	i	l
0	0	1	0	0	0	0	0	1	1

Example (2) shows two words and the associated operation costs. Each insertion, deletion and substitution operation has a cost of 1, while the identity operation (keeping the same character in both strings) amounts to a cost of 0. In the example, there is one deletion operation ($i \rightarrow \varepsilon$), one insertion operation ($\varepsilon \rightarrow i$) and one substitution operation ($u \rightarrow l$), so that the Levenshtein distance between *biischpiu* and *bischpiil* is 3.

Instead of measuring the distance between two words, one may also measure their similarity. A well-known similarity measure is **Longest Common Subsequence** (LCS). It corresponds to Levenshtein distance with inverted weights: 1 for the identity operation, 0 for the insertion, deletion and substitution operations. The string pair of example (2) has an LCS value of 7.

It has been suggested to normalize these values so that they range between 0 and 1, independently of the length of the word. The underlying idea is that a distance of 3 for two long words does not mean the same as a distance of 3 for two very short words. Among other proposals, Heeringa, Kleiweg, et al. (2006) suggest normalizing Levenshtein distance by the length of the alignment. In (2), the length of the alignment is 10. Normalized Levenshtein distance is $\frac{3}{10} = 0.3$. Likewise, **Longest Common Subsequence Ratio** (LCSR) corresponds to LCS normalized by the length of the longer of the two strings; the above example yields an LCSR value of $\frac{7}{9} = 0.78$.[7]

7 In many cases, the length of the alignment is identical to the length of the longer of the two strings. However, (2) shows a case where this equivalence does not hold: both strings are of length 9, but the alignment length is 10.

BA	*BE*	*OS*	*WS*	*ZH*
80.60%	82.77%	80.68%	77.61%	82.53%

Table 6.4.: LCSR similarity values between the dialect texts and their Standard German translations. An LCSR value of 100% means complete identity.

These distance and similarity measures have mainly been used to compare data from different dialects, such as those constituted by dialectological surveys. An example of this use is reported in Section 7.3.5. Here, we use these measures to compare the Swiss German dialects with Standard German (Section 6.4.2), and to compare automatically generated Swiss German texts with Swiss German texts written by native speakers (Section 6.4.3).

6.4.2. Measuring language distance and translation difficulty

The numerical comparison of the Swiss German dialect texts with their (manually created) Standard German translations allows us to assess the difficulties to be expected for our translation system. In other words, we claim that the more similar a dialect is to Standard German, the easier it is to translate into.

We computed the LCSR value for each dialect subcorpus of the *Wikipedia* set and its Standard German translation. Table 6.4 shows that the similarity values range from 77% to 83%, depending on the dialect. Rather surprisingly, the BE and ZH dialect texts seem to be closer to Standard German than BA and OS. The WS dialect is clearly more distant, which is in line with common knowledge about Swiss German dialects.

These LCSR values may be compared with those of unrelated language pairs. For this purpose, we used the test set of the WMT'12 translation task for the language pairs English-(Standard) German and English-French.[8] These parallel corpora consist of 3000 sentence pairs from different news sources; translations were provided by professional translators (Callison-Burch,

8 http://www.statmt.org/wmt12/, accessed 31.5.2012.

Koehn, Monz, Post, et al. 2012).[9] We obtained an LCSR value of 42.95% for English-German, and 47.03% for English-French. These results are more than 30% lower than any value reported in Table 6.4. This discrepancy confirms the hypothesis underlying the entire thesis that translating between very closely related language varieties is fundamentally different from translating between the major standardized languages of Europe. In this sense, it justifies our use of fundamentally different translation techniques such as phonetic transfer rules.

It must be noted that LCSR is very sensitive to word reordering. For example, (3) and (4) both obtain LCSR values of $\frac{3}{6}$, although the former string pair is intuitively more similar than the latter.

(3)

a	b	c	d	e	f			
			d	e	f	a	b	c
0	0	0	1	1	1	0	0	0

(4)

a	b	c	d	e	f			
			d	e	f	g	h	i
0	0	0	1	1	1	0	0	0

This is an issue if the syntactic structures of the two texts differ heavily, as is the case with Standard German sentences containing preterit tense. In such sentences, the Standard German preterit verb is aligned with the Swiss German auxiliary verb, and not with the more similar Swiss German past participle, because the latter is likely to occur in another syntactic position.[10] More generally, LCSR may not be well suited as a similarity measure for the comparison of language pairs with highly different word orders. Moreover, chance matches resulting from a shared alphabet may yield unexpectedly high LCSR values even for completely unrelated languages.

9 A quick look at the data casts doubts about the quality of the translations – doubts that have been largely acknowledged over the course of the previous WMT competitions. We still deem this data good enough to figure as a comparison benchmark.

10 The *Wikipedia* texts contain between 41% and 64% sentences with preterit verbs. However, the percentage of preterit sentences does not seem to be correlated with the LCSR values.

6.4.3. Measuring translation performance and localization precision

In machine translation, the performance of a system is usually measured by comparing its output with reference translations made by professional translators. In the last few years, a couple of metrics have become de-facto standards: BLEU (Papineni et al. 2002), TER (Snover et al. 2006), or METEOR (Lavie and Agarwal 2007). While these metrics have showed fair correlations with human judgements, they are not well suited for our dialect translation approach. Namely, they consider words as atomic entities and only distinguish between word identity and word difference. In consequence, as soon as the proposed translation deviates by a single character from the reference, the whole word is considered wrong. This behavior is not so much of a problem in standard machine translation frameworks, where word pairs are stored explicitly in a lexicon (or phrase table, as this resource is commonly called in Statistical Machine Translation). In contrast, our model translates words by converting Standard German graphemes one by one. A sensible evaluation metric should therefore weight a single conversion error less severely than several conversion errors occurring in the same word. LCSR is one potential evaluation metric for such systems. It has first been proposed by Tiedemann (2009) to evaluate the accuracy of machine translation output in the context of letter-based translation for closely related languages. We follow his idea in this section.

As illustrated in Figure 6.4, each Standard German text of the *Wikipedia* corpus is translated into all five dialects, using our translation system.[11] We then compare these translations with the original dialect text, written in one of the five dialects. We put forward two hypotheses:

1. Any of the five generated dialect texts is closer to the original dialect text than the Standard German text used as input to the translation system. The underlying reasoning is that the five Swiss German dialects all belong to the Alemannic subgroup, whereas Standard German does not.

2. The translation generated in the same dialect as the source should be more similar than the four other translations. This hypothesis provides a justification for the use of georeferenced rules that apply in restricted geographical areas.

11 As in the previous experiment, we use the coordinates of five reference cities to determine the exact target dialect.

Generated	*Original*				
	BA	***BE***	***OS***	***WS***	***ZH***
Standard	80.60%	**82.77%**	80.68%	**77.61%**	82.53%
BA	**80.94%**	81.29%	79.81%	75.88%	81.33%
BE	76.82%	79.91%	75.81%	72.50%	77.97%
OS	77.40%	79.57%	78.13%	72.78%	78.97%
WS	76.57%	79.17%	75.04%	75.82%	78.16%
ZH	80.71%	**82.51%**	**81.26%**	**76.54%**	**83.30%**

Table 6.5.: LCSR values for each *Wikipedia* subcorpus (one per column). The first line reproduces the figures of Table 6.4 and is included to test hypothesis 1. The following lines report the figures obtained by generating all five dialects.

Table 6.5 presents the results of this experiment. Hypothesis 1 must be rejected: for BE and WS dialects, the Standard German text is closer to the original than any of the five generated dialect texts. For the other dialects, the Standard German text lies in second or third position, in the middle of the five generated dialect texts. Likewise, hypothesis 2 must be rejected for three out of five dialects: the automatically generated ZH text is closer to the BE original than the automatically generated BE text (79.91% vs. 82.51%); the same holds for OS (78.13% vs. 81.26%) and WS (75.82% vs. 76.54%). Ideally, we would have obtained the highest values on the diagonal.

These results are rather disappointing. They show that the rules either transform too much or transform wrongly, so that in many cases, we would have been better off by not translating the Standard German text at all. They also show that the georeferencing is not precise enough, so that generating Zürich dialect nearly always is the safest choice. In the following section, we would like to comment on these results.

6.4.4. Discussion

In contrast to the experiment of Section 6.3, this experiment does not rely on a morphological dictionary built beforehand. This means that there are no coverage restrictions resulting from the size of the Standard German word list. Instead, the Standard German texts have been

morpho-syntactically analyzed by *TreeTagger*, whose accuracy is not perfect (as mentioned in Section 5.2.2). As a result, the performance of the translation model is diminished in comparison with the experiment of Section 6.3, where manually annotated tags have been used. A thorough evaluation of the Swiss German output would therefore ask for a thorough evaluation of the tagger.

Most of the good performance of ZH dialect generation, and of the bad performance of BE, OS and WS generation can be attributed to a few rules that overgenerate. Tables 6.6, 6.7 and 6.8 show some examples.

When generating BE dialect (Table 6.6), the system applies *l*-vocalization ($l \rightarrow u$), vowel opening ($i \rightarrow e$, $u \rightarrow o$) and *nd*-to-*ng*-transformation, determined according to the SDS data for the reference city Bern. However, these characteristics are not systematically found in the *Wikipedia* data, either because the authors chose not to reflect them in their spelling, or because they use a dialect that does not feature these properties. Moreover, our lexical rules generate typical BE forms like *u* or *geng*, whereas the authors visibly prefer using the more standard-like forms *und* and *emmer*. All these transformations are not valid in ZH dialect, which means that generated ZH dialect words end up being closer to the BE original from *Wikipedia* than generated BE words.

The situation is similar for Eastern Swiss dialect (Table 6.7). The SDS, and thus our phonetic rules, state that this dialect uses vowel opening, yielding *met* instead of *mit*, and *bsonders* instead of *bsunders*. Likewise, the verbal plural ending *id* does not appear in the *Wikipedia* data. Again, the words generated for ZH dialect do not show these regional particularities and end up being closer to the *Wikipedia* words.

In the case of WS dialect, the situation is slightly different (Table 6.8). On average, this dialect requires more transformations than the dialects on the North of the Alps. In most words, some transformations are applied successfully (*u*-suffix, *nd*-to-*nn*-transformation, unrounding $\ddot{u} \rightarrow i$). However, some of the Wallis-specific rules have been tested less thoroughly than others, leading to higher amounts of overgeneration and undergeneration (e.g., $u \rightarrow \ddot{u}$ in *verbünnu*, lack of unrounding of *oi*). In the end, the original words from *Wikipedia* lie somewhere in the middle between the words generated for WS dialect and the words generated for ZH dialect.

Wikipedia BE	*Generated BE*	*Generated ZH*	*Translation*
gschtaltet	gschtautet	gschtaltet	'shaped'
tiefländer	töiflänger	tüüfländer	'lowlands'
süde	söde	süde	'South'
faltezüg	fautezög	faltezüg	'fold train'
und	u	und	'and'
emmer	geng	immer	'always'

Table 6.6.: Words from the Bern subcorpus of *Wikipedia* articles. The system output for Zürich dialect (third column) often is closer to the original (first column) than the system output for Bern dialect (second column).

Wikipedia OS	*Generated OS*	*Generated ZH*	*Translation*
mit	met	mit	'with'
finded	findid	finded	'find'
bsundrigs	bsonders	bsunders	'particular'
johrhundert	johrhondert	jahrhundert	'century'

Table 6.7.: Words from the Eastern Swiss subcorpus of *Wikipedia* articles. The system output for Zürich dialect (third column) often is closer to the original (first column) than the system output for Eastern Swiss dialect (second column).

Wikipedia WS	*Generated WS*	*Generated ZH*	*Translation*
öi	oi	au	'also'
öffnu	effnu	öffne	'open'
värbunnu	verbünnu	verbunde	'connected'
grind	grint	gründ	'reasons'
diräktor	derektor	direktor	'director'

Table 6.8.: Words from the Wallis subcorpus of *Wikipedia* articles. The original (first column) often lies in between the system output for Wallis dialect (second column) and the system output for Zürich dialect (third column).

As mentioned above, a large part of the bad performance of OS and BE can be explained with the vowel opening rule. In order to quantify the impact of this specific rule, we have disabled it and repeated the experiment. This change raised the OS-OS value from 78.13% to 81.23%, and the BE-BE value from 79.91% to 82.56%. These figures are significantly higher, but still lie slightly below the best performing dialect (Standard German for BE, ZH dialect for OS). This experiment shows that a single phonetic rule can have a drastic impact on the overall results. However, its simple removal is not an adequate solution in our opinion. Vowel opening does exist in many Swiss German dialects and is consistently reflected in spelling in some of them (Aargau, Luzern). Even in the *Wikipedia* BE and OS texts, there are indeed mentions of the open vowel qualities.

6.4.5. Conclusion

In this section, we have presented a measure to assess linguistic distance (normalized Levenshtein distance), and a related measure to assess linguistic similarity (longest common subsequence ratio). We have shown that the latter can be used to quantify the expected difficulty of generating a specific dialect on the basis of Standard German source material. We have also used LCSR to compare automatically generated dialect texts with the original *Wikipedia* dialect texts. We have put forward two claims that a successful multi-dialectal sentence generation system should fulfill. Unfortunately, both claims could only be partially satisfied. The error analysis showed that a small number of rules generate output that is incompatible with the *Wikipedia* texts, but not necessarily incompatible with the real pronunciations found in that dialect. Future work should address this shortcoming by measuring the (positive or negative) impact of each rule separately, on different types of texts, for different dialects.

This experiment relies solely on the word-level transfer rules. Whenever the Standard German input text contains words that are not realizable in Swiss German, such as preterit or genitive forms, the Standard German base form is simply copied into the Swiss German output without further transformation. This leads to rather unnatural dialect translations. For example, the sentence in Figure 6.5 features a preterit verb *war* and a genitive noun phrase *der Grafen der alten Homburg*. The BA translation contains the radical of the verb *sei* instead of the perfect paraphrase *isch gsi*, as well as the uninflected and untransformed words of the genitive noun phrase. The

Original BA	Im Mittelalter ischs Dorf im Bsitz vo de Grafe vo dr alte Homburg gsi.
Standard German	Im Mittelalter war das Dorf im Besitz der Grafen der alten Homburg.
Generated BA	Im Mittelalter sei s Dorf im Bsitz der Grafen der alt Homburg.
English Translation	'In the Middle Ages, the village was a property of the Counts of Old Homburg.'

Figure 6.5.: A BA sentence from *Wikipedia* with its Standard German translation, and with the BA translation generated from the latter.

inclusion of the syntactic transfer rules – together with a full syntactic analysis to replace the error-prone shallow tagging – is expected to yield gains in terms of accuracy and fluency of the resulting dialect texts. The corresponding experiments are presented in Sections 6.6 and 6.7.

6.5. Shibboleth-based dialect identification

In Section 6.3, we have discussed the creation of a multi-dialectal Swiss German lexicon with the help of the word-level transfer rules. In this lexicon, every Swiss German word is associated with a probabilistic map that determines the dialectal area in which the word form occurs.

If one is given a set of words written in some unknown Swiss German dialect, one can retrieve the words in the lexicon and overlay the corresponding maps. The resulting map depicts the dialectal area in which the given words are most likely to occur. This is, in a nutshell, the goal of dialect identification, as presented in this section. Most of this work has been published as Scherrer and Rambow (2010).

First, we introduce the problem in a more detailed way and review some related work. We then present our dialect identification model, and finally show the performance of this approach in comparison with a simpler baseline model.

6.5.1. Introduction and related work

Dialect identification (dialect ID) can be viewed as an instance of language identification (language ID) where the different languages are very closely related. Written language ID has been a popular research object in the last few decades, and relatively simple algorithms have proven to be very successful. The central question of **language identification** is the following: given a segment of text, which one of a predefined set of languages is this segment written in? Language identification is thus a classification problem.

Various language identification methods have been proposed in the last three decades. Hughes et al. (2006) and Řehůřek and Kolkus (2009) provide recent overviews of different approaches. One of the simplest and most popular approaches is based on character n-gram sequences (Cavnar and Trenkle 1994). For each language, a character n-gram language model is learned, and test segments are scored by all available language models and labeled with the best scoring language model. Related approaches involve more sophisticated learning techniques (e.g., feature-based models such as support vector machines). A completely different approach relies on the identification of entire high-frequency words in the test segment (Ingle 1980). Other models have proposed to use morpho-syntactic information.

Dialect identification comes in two flavors: spoken dialect ID and written dialect ID. These two tasks are rather different. Spoken dialect ID relies on speech recognition techniques which may not cope well with dialectal diversity. However, the acoustic signal is also available as input. On the other hand, written dialect ID has to deal with non-standardized spellings that may occult real dialectal differences. Moreover, some phonetic distinctions cannot be expressed in orthographic writing systems and limit the input cues in comparison with spoken dialect ID. As an example of spoken dialect ID, Biadsy, Hirschberg, and Habash (2009) classify speech samples from four Arabic dialects plus Modern Standard Arabic (see Section 2.8).

The *Chochichästli-Orakel*[12] presents an original approach to the identification of Swiss German dialects. By specifying the pronunciation of ten predefined words, the web site creates a probability map that shows the likelihood of these pronunciations in the Swiss German dialect area. The probability maps of these ten words are extracted from the SDS.

12 http://dialects.from.ch

Let us now define our dialect ID method. It is based on the concept of **shibboleth**:

> One of the most venerable dialect observations, and perhaps the most fatal one, is recorded in the Old Testament, when the Gileadites were battling the Ephraimites along the Jordan. Whenever the Gileadites captured a fugitive, they asked him if he was an Ephraimite. If he said no, they would then ask him to name an ear of corn, which the Gileadites called a *shibboleth*. According to the scriptural account (Judges XII, 6), "He said *sibboleth*, for he could not frame to pronounce it right. Then they took him and slew him."
>
> The word *shibboleth* has entered English and many other languages with the meaning 'test word' or more generally 'a distinguishing trait'. (Chambers and Trudgill 1998, p. 13)

We argue that almost every Swiss German word is a potential shibboleth in that it allows to distinguish two dialects from each other. Then, the more words we combine, the more precise the delimitation of the dialectal area will become. If given a text of an unknown dialect, we collect all words of the text and try to determine their dialectal area. So, in contrast to Ingle (1980), we do not only use function words, but the entire vocabulary for dialect identification, in the hope that this increases accuracy by keeping the test segments relatively short. And in contrast to the *Chochichästli-Orakel*, we do not limit ourselves to a small set of lexical items which most likely do not occur in a random dialect text.

To sum up, what is presented here corresponds to written dialect ID, applied to the Swiss German dialect area. The proposed model follows a bag-of-words approach, assuming that every dialectal word form is linked to a probability map defining its area of occurrence. By combining the cues of all words of a sentence, it should be possible to obtain a fairly reliable geographic localization of that sentence.

6.5.2. The model

The dialect ID system consists of a Swiss German lexicon that associates word forms with their geographical extension (already discussed in Section 6.3), and of a testing procedure that splits a sentence[13] into words, looks up their geographical extensions in the lexicon, and combines the word-level maps into a sentence-level map (see following subsections).

Word lookup and dialect identification

The goal of dialect identification is to compute a probability map for a text segment of unknown origin. As a preprocessing step, the segment is tokenized, punctuation markers are removed and all words are converted to lower case. The identification process as such can be broken down in three levels:

1. The probability map of a text segment depends on the probability maps of the words contained in the segment.

2. The probability map of a word depends on the probability maps of the derivations that yield the word.

3. The probability map of a derivation depends on the probability maps of the rules it consists of.

In practice, every word of a given text segment is looked up in the lexicon. The lookup yields *m* derivations from *m* different Standard German words.[14] Recall that the dialect lexicon already contains the probability maps of the derivations, as computed by the application of the transfer rules (see Section 5.2.12). If the lookup does not succeed (either because its Standard German

13 The model does not require the material to be syntactically well-formed. Although we use complete sentences to test the system, any sequence of words is accepted.

14 Theoretically, two derivations can originate at the same Standard German word and yield the same Swiss German word, but nevertheless use different rules. Our system handles such cases as well, but we are not aware of such cases occurring with the current rule base.

equivalent did not appear in the TIGER treebank, or because the rule base lacked a relevant rule), the word is skipped. Let us thus explain the first two levels in more detail, in reverse order.

Computing the probability map for a word

A dialectal word form may originate in different Standard German words. For example, the three following derivations all lead to the same dialectal form:[15]

(5) a. sind [VAFIN] $\overset{*}{\rightarrow}$ si (valid only in Western dialects)
b. sein [PPOSAT] $\overset{*}{\rightarrow}$ si (in Western and Central dialects)
c. sie [PPER] $\overset{*}{\rightarrow}$ si (in the majority of dialects)

Our system does not take the syntactic context into account and therefore cannot determine which derivation is the correct one. We approximate by choosing the most probable one in each geographic location. The probability map of a Swiss German word w is thus defined as the **pointwise maximum**[16] of all derivations leading to w, starting with different Standard German words $w_0^{(j)}$:

$$\underset{t \in GSS}{\forall} \; p(w \mid t) = \max_j p(w_0^{(j)} \overset{*}{\rightarrow} w \mid t)$$

This formula does not take into account the relative frequency of the different derivations of a word. This may lead to unintuitive results. Consider the two derivations:

(6) a. heute [ADV] $\overset{*}{\rightarrow}$ hit (valid only in Southwestern dialects)
b. Hit [NN] $\overset{*}{\rightarrow}$ hit (valid in all dialects)

The occurrence of the adverb *hit* 'today' in a dialect text is a good indicator for Southwestern dialects, but it is completely masked by the potential presence of the borrowing *Hit* 'hit' in

15 As motivated in Section 3.2, we do not distinguish between open and closed vowels and collapse *i* and *y* into *i*.

16 Note that these derivations are alternatives and not joint events. This is thus not a joint probability.
As in preceding chapters, *GSS* is defined as the set of all cells that exist in the fixed-grid raster map of Standard German Switzerland.

all dialects.[17] We can avoid this by weighting the derivations by the **word frequency** of w_0: the adverb *heute* is more frequent than the noun *Hit* and is thus given more weight in the identification task. This weighting can be justified on dialectological grounds: frequently used words tend to show higher interdialectal variation than rare words.

Another assumption in the above formula is that each derivation has the same **discriminative potential**. Again, this is not true: a derivation that is valid in only 10% of the Swiss German dialect area is much more informative than a derivation that is valid in 95% of the dialect area. Therefore, we propose to weight each derivation by the proportional size of its validity area. The discriminative potential of a derivation d is defined as follows:[18]

$$DP(d) = 1 - \frac{\sum_{t \in GSS} p(d \mid t)}{|GSS|}$$

Our definition of the discriminative potential reminds Goebl's Weighted Identity Value (*Gewichtender Identitätswert*, GIW; Goebl 1984). In its formulation, Goebl computes a so-called co-identity value *KOI* (*Koidentität*) between two measurement points j, k with respect to a linguistic feature i:

$$KOI_{jk}(i) = 1 - \frac{f(TAX_{jk}(i)) - 1}{n_i \cdot w}$$

where $f(TAX_{jk}(i))$ corresponds to the absolute frequency of the linguistic feature i, n_i represents the number of (non-null) measurement points in the data set, and w represents an empirically determined weight.

Our formulation differs from Goebl's in three respects. First, we use probability instead of absolute frequencies, since this is what our map data provides. Second, we set $w = 1$. Third, we do not subtract 1 in the numerator of the formula. Goebl introduced this correction to avoid an edge case: if $f(TAX_{jk_i}) = n_i$, the *KOI* value is 0, which is undesirable. In our case, this

17 Of course, this particular example could be avoided by distinguishing upper and lower case. However, other similar derivation pairs do exist.

18 d is a notational abreviation for $w_0 \overset{*}{\rightarrow} w_n$.

would mean that a single derivation valid in all over German-speaking Switzerland is discarded without Goebl's correction, but kept with it. We do not mind discarding such derivations since they do not improve dialect identification in any way.

The results reported below will show the relative impact of these two weighting techniques and of the combination of both with respect to the unweighted map computation.

Computing the probability map for a segment

The probability of a text segment s can be defined as the joint probability of all words w contained in the segment. Here, we compute the **pointwise product** of all word maps. We performed some smoothing in order to prevent erroneous word derivations (mainly due to non-implemented lexical exceptions) from completely zeroing out the probabilities. We assumed a minimum word probability of $\varphi = 0.1$ for all words in all cells:

$$\underset{t \in GSS}{\forall} \; p(s \mid t) = \prod_{w \in s} \max\left(\varphi, p(w \mid t)\right)$$

6.5.3. Experiments and results

Data

We use the *Wikipedia* data set of Section 6.2 to evaluate our model. It contains a development set consisting of 550 sentences for 6 dialects, and a test set of equal size. The development set was also used to train the baseline model discussed below.

We created a subset of the test set that was balanced according to the population size of the six dialect regions (statistical details about the dialect regions can be found in Table 6.9). The resulting subset contained 291 sentences. The idea behind weighting the test corpus size by the population of the respective dialect region is that texts from dialects with many speakers are more likely to occur than texts from rare dialects. Roughly, this weighting can be viewed as a

Wikipedia name	*Abbreviation*	*Population*	*Surface*
Baseldytsch	BA	8%	1%
Bärndütsch	BE	17%	13%
Seislertütsch	FR	2%	1%
Ostschwizertütsch	OS	14%	8%
Wallisertiitsch	WS	2%	7%
Züritüütsch	ZH	22%	4%

Table 6.9.: The six dialect regions selected for our tests, with their *Wikipedia* name and the abbreviation used henceforth. We also report the percentage of the German-speaking population living in the regions, and the percentage of the surface of the region relative to the entire country.

prior, the probability of the text being constant:

$$p(dialect \mid text) = p(text \mid dialect) * p(dialect)$$

In order to test the robustness of our model, we collected a second set of texts from various web sites other than *Wikipedia*. The dialects of these texts could be identified through metadata,[19] but we nevertheless checked this information for plausibility. The *Web* data set contains 144 sentences (again distributed proportionally to population size) and is thus roughly half the size of the *Wikipedia* test set.

The *Wikipedia* data contains an average of 17.8 words per sentence, while the *Web* data shows 14.9 words per sentence on average.

Baseline: N-gram model

To compare our dialect ID model, we created a baseline system that uses a character-n-gram approach. This approach is fairly common for language ID and has also been successfully applied

19 We mainly chose websites of local sports and music clubs, whose localization allowed to determine the dialect of their content.

Dialect	Wikipedia			Web		
	Prec.	Recall	F-meas.	Prec.	Recall	F-meas.
BA	34%	61%	44%	27%	61%	37%
BE	78%	51%	61%	51%	47%	49%
FR	28%	71%	40%	10%	33%	15%
OS	63%	64%	64%	50%	38%	43%
WS	58%	100%	74%	14%	33%	20%
ZH	77%	62%	69%	77%	41%	53%
Average			62%			46%

Table 6.10.: Performances of the 5-gram model on *Wikipedia* test data (left) and *Web* test data (right). The columns refer to precision, recall and F-measure respectively. The average is weighted by the relative population sizes of the dialect regions.

to dialect ID (Biadsy, Hirschberg, and Habash 2009). However, it requires a certain amount of training data that may not be available for specific dialects, and it is uncertain how it performs with a set of dialects that are very similar to each other.

We trained 2-gram to 6-gram models for each dialect with the SRILM toolkit (Stolcke 2002), using the *Wikipedia* development corpus. We scored each sentence of the *Wikipedia* test set with each dialect model. The predicted dialect was the one which obtained the lowest perplexity. In other words, we assume that all test sentences are written in one of the six dialects. The 5-gram model obtained the best overall performance, and results on the *Wikipedia* test set were surprisingly good (see Table 6.10, leftmost columns).[20] Note that in practice, 100% accuracy is not always achievable; a sentence may not contain a sufficient localization potential to assign it unambiguously to one dialect.

However, we suspect that these results are due to overfitting. It turns out that the number of Swiss German *Wikipedia* authors is very low (typically, one or two active writers per dialect), and that every author uses distinctive spelling conventions and writes about specific subjects. For instance, most ZH articles are about Swiss politicians, while many OS articles happen to deal with religion and mysticism. Our hypothesis is thus that the n-gram model learns to recognize a

20 In the result tables, we omit decimal places as all values are based on 100 or less data points. We did not perform statistical significance tests on our data.

specific author and/or topic rather than a dialect. This hypothesis is confirmed on the *Web* data set: the performance drops by 15 percentage points or more (same table, rightmost columns). The drops are similar for $n = [2..6]$. Recall that in all our evaluations, the average F-measures for the different dialects are weighted according to the relative population sizes of the dialect regions because the size of the test corpus is proportional to population size.

We acknowledge that a training corpus of only 100 sentences per dialect provides limited insight into the performance of the n-gram approach. We were able to double the training corpus size with additional *Wikipedia* sentences. With this extended corpus, the 4-gram model performed better than the 5-gram model. It yielded a weighted average F-measure of 79% on *Wikipedia* test data, but only 43% on *Web* data. The additional increase on *Wikipedia* data (+17% absolute with respect to the small training set), together with the decrease on *Web* data (−3% absolute) confirms our hypothesis of overfitting. An ideal training corpus should thus contain data from several sources per dialect.

To sum up, n-gram models can yield good performance even with similar dialects, but require large amounts of training data from different sources to achieve robust results. For many small-scale dialects, such data may not be available.

Our model

The n-gram system presented above has no geographic knowledge whatsoever; it just consists of six distinct language models that could be located anywhere. In contrast, our model yields probability maps of German-speaking Switzerland. In order to evaluate its performance, we thus had to determine the geographic localization of the six dialect regions defined by the *Wikipedia* authors (see Table 6.9). We defined the regions according to the respective canton boundaries and to the German-French language border in the case of bilingual cantons (BE, FR, WS), resulting in the map reproduced in Figure 6.6.

The predicted dialect region of a sentence s is defined as the region in which the most probable

Figure 6.6.: The localization of the six dialect regions used in our study.

point has a higher value than the most probable point in any other region:

$$Region(s) = \arg\max_{Region} \left(\max_{t \in Region} p(s \mid t) \right)$$

Experiments were carried out for the four combinations of the two derivation-weighting techniques presented in Section 6.5.2 and for the two test sets (*Wikipedia* and *Web*). All experiments refer to the database implementation of the word-level transfer rules.[21]

Results are displayed in Tables 6.11 to 6.14. The FR segments obtained particularly low scores. Most of them were indeed misclassified as BE, which reflects the geographic and linguistic proximity of these two regions.

The tables show that frequency weighting helps on both corpora; the discriminative potential only slightly improves performance on the web corpus. Crucially, the two techniques are additive,

21 As in the previous experiments, the finite-state implementation yields largely comparable results. In order not to overload this presentation, we do not discuss them here.

Dialect	Wikipedia			Web		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BA	41%	19%	26%	80%	22%	35%
BE	42%	62%	50%	48%	76%	59%
FR	0%	0%	0%	17%	33%	22%
OS	36%	41%	38%	45%	41%	43%
WS	3%	14%	5%	8%	33%	13%
ZH	65%	33%	44%	62%	37%	46%
Weighted Average			40%			46%

Table 6.11.: Performances of the word-based model using unweighted derivation maps.

Dialect	Wikipedia			Web		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BA	50%	33%	40%	57%	22%	32%
BE	47%	60%	53%	60%	79%	68%
FR	0%	0%	0%	0%	0%	0%
OS	29%	31%	30%	46%	50%	48%
WS	11%	29%	15%	17%	33%	22%
ZH	60%	47%	53%	65%	53%	58%
Weighted Average			44%			**53%**

Table 6.12.: Performances of the word-based model using derivation maps weighted by word frequency.

Dialect	Wikipedia			Web		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BA	34%	31%	32%	38%	17%	23%
BE	46%	47%	47%	54%	76%	63%
FR	11%	14%	13%	20%	33%	25%
OS	34%	50%	40%	53%	59%	56%
WS	5%	14%	7%	0%	0%	0%
ZH	47%	27%	34%	75%	43%	55%
Weighted Average			37%			51%

Table 6.13.: Performances of the word-based model using derivation maps weighted by their discriminative potential.

Dialect	Wikipedia			Web		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BA	46%	28%	35%	33%	11%	17%
BE	47%	62%	54%	58%	84%	69%
FR	0%	0%	0%	20%	33%	25%
OS	35%	31%	33%	47%	47%	47%
WS	8%	29%	13%	14%	33%	20%
ZH	63%	53%	58%	66%	51%	58%
Weighted Average			**46%**			52%

Table 6.14.: Performances using derivation maps weighted by word frequency and discriminative potential.

so in combination, they yield the best overall results. In comparison with the baseline model, there is a performance drop of about 16 percent absolute on *Wikipedia* data. In contrast, our model is very robust and outperforms the baseline model on the *Web* test set by about 7 percent absolute.

These results seem to confirm what we suggested above: that the n-gram model overfitted on the small *Wikipedia* training corpus. Nevertheless, it is still surprising that our model has a lower performance on *Wikipedia* than on *Web* data. The reasons for this discrepancy are probably the same that have been shown to cause errors in Section 6.3.4. First, *Web* writers use a spelling that is more phonetic and more closely corresponds to the Dieth spelling used in the transformation rules. *Wikipedia* authors, in contrast, often translate existing Standard German articles and fail to completely adapt the spelling to Swiss German. Second, *Wikipedia* articles use a proportionally larger amount of proper nouns and low-frequency words which cannot be found in the lexicon and which therefore reduce the localization potential of a sentence.

However, one should note that the word-based dialect ID model is not limited to the six dialect regions used for evaluation here. It can be used with any size and number of dialect regions of German-speaking Switzerland. This contrasts with the n-gram model which has to be trained specifically on every dialect region. Using the Swiss German *Wikipedia* as a data source, only two additional dialect regions present an equivalent amount of text.

Variations

In the previous section, we have defined the predicted dialect region as the one in which the most probable point (**maximum**) has a higher probability than the most probable point of any other region. The results suggest that this metric penalizes small regions (BA, FR, ZH). In these cases, it is likely that the most probable point is slightly outside the region, but that the largest part of the probability mass is still inside the correct region. Therefore, we tested another approach: we defined the predicted dialect region as the one in which the **average** probability is

Dialect	Wikipedia			Web		
	Max.	Avg.	Comb.	Max.	Avg.	Comb.
BA	35%	32%	32%	17%	43%	43%
BE	54%	39%	54%	69%	54%	69%
FR	0%	7%	7 %	25%	11%	11%
OS	33%	23%	33%	47%	49%	47%
WS	13%	13%	13%	20%	31%	20%
ZH	58%	60%	60%	58%	68%	68%
W. Avg.	46%	40%	**47%**	52%	55%	**58%**

Table 6.15.: Comparison of different evaluation metrics. All values refer to F-measures obtained with frequency and discriminative potential-weighted derivation maps. **Max** refers to the metric used in Table 6.14, **Avg** refers to the average metric, **Comb** to the combined metric. The underlined values in the **Avg** and **Max** columns represent those used for the **Comb** metric.

higher than the average probability in any other region:

$$Region(s) = \arg\max_{Region} \left(\frac{\sum_{t \in Region} p(s \mid t)}{|Region|} \right)$$

This metric effectively boosts the performance on the smaller regions, but comes at a cost for larger regions (Table 6.15). We also combined the two metrics with a simple heuristics: the **average** metric is used for regions that cover less than 5% of the Swiss territory (BA, FR, ZH according to Table 6.9, page 223), whereas the **maximum** metric is used for regions that cover more than 5% of the Swiss territory (BE, OS, WS). This **combined** metric further improves the performance of our system while relying on an objective measure of region surface.

We believe that region surface as such is not so crucial for the metrics discussed above, but rather serves as a proxy for linguistic heterogeneity. Geographically large regions like BE tend to have internal dialect variation, and averaging over all dialects in the region leads to low figures. In contrast, small regions show a quite homogeneous dialect landscape that may extend to adjacent regions. In this case, the probability peak is less relevant than the average probability in the entire region. It would be interesting to develop more fine-grained measures of linguistic

heterogeneity in order to test these claims, for example by using the geostatistical methods developed in the field of dialectometry (see Chapter 7).

6.5.4. Discussion

In our experiments, the word-based dialect identification model skipped about one third of all words (34% on the *Wikipedia* test set, 39% on the *Web* test set) because they could not be found in the lexicon.[22] While our model does not require complete lexical coverage, this figure shows that the system can be improved. First, the rule base can be extended to better account for lexical exceptions, orthographic variation and irregular morphology. Second, a mixed approach could combine the benefits of the word-based model with the n-gram model. This would require a larger, more heterogeneous set of training material for the latter in order to avoid overfitting. Additional training data could be extracted from the web and automatically annotated with the current model in a semi-supervised approach.

In the evaluation presented above, the task consisted of identifying the dialect of single sentences. However, one often has access to longer text segments, which makes our evaluation setup harder a realistic dialect identification task would be, all the more since a single sentence may not always contain enough discriminative material to assign it to a unique dialect region. Testing our dialect identification system on the paragraph or document level should thus provide more realistic results.

6.6. Correctness of the syntactic transfer rules

The preceding experiments were aimed at evaluating the performance of the manually-built rule base on independently collected dialect data. The goal of this experiment is more modest: it rather represents a principled method of creating and debugging the syntactic rule base. Three main characteristics should be pointed out. First, we use syntactically annotated Standard

22 These figures are consistent with the observations made in Section 6.3.2.

German texts for development and testing. No dialectal raw data is used, the correctness of the output is merely annotated by a native speaker. Second, we evaluate the rules one-by-one, without bothering with interactions. Third, we only consider the syntactic rules. Concretely, this means that the material to be evaluated corresponds to utterances with dialectal syntax, but Standard German words.[23]

As mentioned above, the syntactic transformation rules require morpho-syntactically annotated Standard German input data, in the form of CoNLL-X style dependency structures. We start by using the reference implementation of the CoNLL-X format for Standard German, based on the manually annotated TIGER treebank (as discussed in Section 5.4.1). This allows us to minimize the impact of annotation errors. In a second step, we use the same sentences, but automatically annotate them with the *Fips* parser (discussed in Section 5.4.4). This will allow us to assess the robustness of our rules with respect to noisy syntactic annotation.

6.6.1. Corpus frequencies

The selection process of the syntactic constructions to be implemented has been described in Section 3.4.3; the list of selected syntactic constructions figures in Table 3.2 (page 110). In order to get an idea of the frequency of the selected constructions, we started by searching the TIGER treebank for the corresponding syntactic patterns. Table 6.16 shows frequency counts of the respective phenomena. Note that these figures should be taken with a grain of salt. First, the TIGER corpus consists of newspaper text, which is hardly representative of everyday use of Swiss German dialects. Second, it is difficult to obtain reliable recall figures without manually inspecting the entire corpus.

6.6.2. Testing the syntax rules with the TIGER treebank

For each syntactic construction, a development set and a test set were extracted from the TIGER treebank, each of them comprising at most 100 sentences showing that construction. We

23 Part of the work presented here has been published as Scherrer (2011b).

CONSTRUCTION	SENTENCES
Preterit tense	13439
Genitive case	15351
Person name determiners	5410
Verb raising	3246
Verb projection raising	2597
Prepositional dative marking	2708
Article doubling	61
Complementizer in wh-phrases	478
Relative pronouns	4619
Final clauses	629
Pronoun sequences	6
Predicative adjectives	2784
Total TIGER sentences	40000

Table 6.16.: Number of sentences in the TIGER corpus that trigger the corresponding transformation rule.

did not evaluate the accusative-dative pronoun sequences because of their small number of occurrences. Predicative adjective agreement was not evaluated because the author did not have native speaker's intuitions about this phenomenon. After achieving fair performance on the development sets, the held-out test data was manually evaluated.

Table 6.17 shows the accuracy of the rules on the test data. Each line represents a transformation rule. Some syntactic phenomena are thus covered by more than one rule. For example, transforming preterit tense to present perfect tense is rather different in matrix clauses and in subordinate clauses, and it was simpler to devise two independent rules for these two transformation contexts.

Recall also that some rules cover different dialectal variants, each of which may show different types of errors. The performance of every variant is thus indicated separately in the second column of Table 6.17. The number of variants depends on the rule; it ranges from 1 to 4. Variants that do not require any syntactic change of the Standard German source trivially achieve accuracy figures of 100%; these variants are marked with parentheses in Table 6.17. In summary, the overall performance of the transformation rules lies at 85% accuracy and above for most rules.

Construction	*Accuracy per variant*	*Sentences*
Preterit tense in matrix clauses	88%	100
Preterit tense in subordinate clauses	90%	100
Genitive in prepositional phrases	95%	100
Genitive in common noun attributes	92%	100
Genitive in proper noun attributes	68% 90% 91% 91%	100
Person name determiners	80% (100%)	100
Verb raising (VR) in matrix clauses	93% (100%)	91
VR in dependent clauses with modal verb	98% (100%)	100
VR in dependent clauses with aux. verb	97% (100%)	100
VR in dependent clauses with auxiliary and modal verb	100% 100% (100%)	8
Verb projection raising (VPR) in matrix clauses	86% (100%)	65
VPR in dependent clauses with modal verb	90% (100%)	89
VPR in dependent clauses with aux. verb	89% (100%)	83
VPR in dependent clauses with auxiliary and modal verb	50% 50% (100%)	8
Prepositional dative marking (PDM) with personal pronouns	100% 100% (100%)	3
PDM with interrogative pronouns	100% 100% (100%)	4
PDM with full NPs	93% 93% (100%)	100
Article doubling	100% 100% (100%)	31
Complementizer in wh-phrases	69% (100%)	100
Nominative/accusative relative pronoun	100%	100
Dative relative pronoun	91% 96% 96% (100%)	54
Prepositional relative pronoun	69% 94% 99% (100%)	100
Final clauses	92% 100% 100% (100%)	100

Table 6.17.: Accuracy of the syntactic transformations. Each line represents a transformation rule, and accuracy is computed separately for each variant. Figures in parentheses refer to variants that do not require any syntactic change with respect to Standard German.

For example, the rule "Final clauses" (last line in Table 6.17) contains four variants, symbolized by four accuracy values (in the present case, the four variants are *für…z*, *zum*, *zum…z*, *um…z*). The last one is syntactically identical with Standard German and trivially yields 100% of accuracy. Two other variants obtain 100% of accuracy, while the fourth variant (*zum*, in the present case) yields 92% accuracy.

The rightmost column of Table 6.17 shows the size of the test corpus: for some phenomena, the TIGER corpus only contained a small number of examples. The verb projection rising subcorpora are subsets of the verb raising subcorpora where the verb projection rule yields different results from the verb raising rule (i.e. when the dependent verb has arguments).

Note that the evaluation was performed on variants, not on inquiry points. The mapping between the variants and the inquiry points is supported by the SADS data and is not the object of the present evaluation.

6.6.3. Analysis of transformation errors

Even if the performance of the rule base is rather satisfactory, we could distinguish four major types of errors.

Annotation errors The annotation of the TIGER treebank has been done semi-automatically and is not exempt of errors, especially in the case of out-of-vocabulary words. These problems degrade the performance of rules dealing with proper nouns. In (7a), the first name *Traute* is wrongly analyzed as a preterit verb form *traute* 'trusted, wedded', leading to an erroneous placement of the determiner (7b). The correct transformation would have yielded (7c):

(7) a. Traute Müller
b. *traute die Müller
c. die Traute Müller

1	gegen	gegen	APPR	PREP	–	17	PP
2	Bousquet	Bousquet	NE	N	Acc.Sg.Masc	1	PN
3	und	und	KON	KON	–	2	KON
4	Papon	Papon	NE	N	Acc.Sg.Masc	3	CJ

Figure 6.7.: Excerpt of a noun phrase conjunction. The head of the first conjunct (*Bousquet*) is the preposition (no. 1). However, the head of the second conjunct (*Papon*) is the conjunction *und* (no. 3), although it ultimately also depends on the preposition. If a transformation rule looks for nouns headed by a preposition, it will typically find the first conjunct but not the second.

Imperfect heuristics Some rules rely on a syntactic distinction that is not explicitly encoded in the TIGER annotation. Therefore, we had to resort to heuristics, which do not work well in all cases. For instance, the genitive replacement rule needs to distinguish human from non-human NPs. We posited a heuristic that classifies masculine and feminine named entities as human, as well as two-word named entities (supposed first name / last name pairs). For example, the first heuristic fails for *Iran*, while the second heuristic fails for *Tel Aviv*.

Likewise, adding a complementizer to wh-phrases overgenerates because the TIGER annotation does not reliably distinguish between clause-adjoined relative clauses and interrogative clauses introduced as complement of the main verb.

Conjunctions Many rules look for a specific combination of part-of-speech tag and dependency relation type. For a given word node, these two features are conveniently expressed in the POSTAG and DEPREL fields of the node. However, things get complicated when the dependent is a conjunction of several elements. In this case, the CoNLL-X guidelines specify that the second conjunct is headed by the conjunction, which is itself headed by the first conjunct (see Figure 6.7). As a result, the "real" syntactic head of the second conjunct often could not be retrieved. However, extending the search patterns to the most frequent types of conjunctions should not be too difficult.

Word order errors Appositions and quotation marks sometimes interfere with transformation rules and lead to typographically or syntactically unfortunate sentences. Example (8a) shows an original sentence, and (8b) its transformation; the word order was successfully changed, but the quotes did not follow the words they surround.

(8) a. Das habe ihr einen Startvorsprung verschafft, denn sie habe "auf bestimmte Milieus zurückgreifen" können.[24]

b. ? Das habe ihr einen Startvorsprung verschafft, denn sie habe " " können auf bestimmte Milieus zurückgreifen.

In other cases, the linguistic description is not very explicit. For example, in the verb projection raising rule, we found it difficult to decide which constituents are moved and which are not. Again, scope effects may prevent polarity items from moving; see example (9a) for an original sentence, (9b) for wrongly generated verb projection raising, and (9c) for the correct dialectal word order. Different types of adverbs also tend to behave differently.

(9) a. S. soll dem Chef der Bauernpartei P. als Ministerpräsident nachfolgen, der seit seiner Berufung vor einem Monat keine tragfähige Koalition hat bilden können.[25]

b. *S. soll dem Chef der Bauernpartei P. als Ministerpräsident nachfolgen, der seit seiner Berufung vor einem Monat hat können *keine tragfähige Koalition* bilden.

c. S. soll dem Chef der Bauernpartei P. als Ministerpräsident nachfolgen, der seit seiner Berufung vor einem Monat *keine tragfähige Koalition* hat können bilden.

6.6.4. Testing the syntax rules with *Fips* annotations

In the preceding sections, we have relied on the manual annotation of the TIGER treebank. This guaranteed that the errors we encountered were always due to our implementation of the transformation rules, not to the initial syntactic analysis. In this section, we take the same sentences, but we syntactically analyze them with *Fips* and evaluate the performance of the

24 'That's what gave her a head start because she was able to "rely on certain milieus"'.

25 'S. is to follow the leader of the Peasant Party P. as prime minister, who has not been able to form a viable coalition since his appointment a month ago.'

Construction	*Accuracy per variant*	*Matched*
Preterit tense in matrix clauses	43%	99%
Preterit tense in subordinate clauses	16%	16%
Genitive in prepositional phrases	50%	63%
Genitive in common noun attributes	62%	85%
Genitive in proper noun attributes	33% 37% 37% 37%	38%
Person name determiners	23% (100%)	33%
Verb raising (VR) in matrix clauses	27% (100%)	27%
VR in dependent clauses with modal verb	5% (100%)	5%
VR in dependent clauses with aux. verb	13% (100%)	13%
VR in dependent clauses with auxiliary and modal verb	0% 0% (100%)	0%
Verb projection raising (VPR) in matrix clauses	0% (100%)	0%
VPR in dependent clauses with modal verb	0% (100%)	0%
VPR in dependent clauses with aux. verb	0% (100%)	0%
VPR in dependent clauses with auxiliary and modal verb	0% 0% (100%)	0%
Prepositional dative marking (PDM) with personal pronouns	100% 100% (100%)	100%
PDM with interrogative pronouns	0% 0% (100%)	0%
PDM with full NPs	48% 48% (100%)	61%
Article doubling	83% 83% (100%)	83%
Complementizer in wh-phrases	1% (100%)	3%
Nominative/accusative relative pronoun	2%	2%
Dative relative pronoun	15% 17% 17% (100%)	17%
Prepositional relative pronoun	0% 0% 0% (100%)	1%
Final clauses	29% 31% 31% (100%)	31%

Table 6.18.: Accuracy of the syntactic transformations when analyzed by *Fips*. Each line represents a transformation rule, and accuracy is computed separately for each variant. The rightmost column shows in how many cases the syntactic analysis matched the conditions of the rule.

transformation rules on this automatically annotated data. A rather large performance drop is expected from this change: the accuracy figures of the German *Fips* parser – although never empirically verified (but see for example Scherrer (2008a)) – are assumed to lie considerably below state-of-the-art parsers, and *a fortiori* below the manually verified annotations of the TIGER treebank.

Table 6.18 summarizes the result of this experiment. Its layout resembles the one of Table 6.17. The last column shows in how many cases the syntactic analysis of *Fips* matched the conditions of the respective rule (these values would always be 100% in the experiment of Table 6.17). The second-last column shows in how many of these cases the performed transformation was correct. The subcorpus sizes in absolute numbers are identical to the ones reported in Table 6.17.

For example, from the 100 sentences containing a genitive prepositional phrase (as indicated by the original TIGER annotation), *Fips* only analyzes 64 sentences in such a way that the genitive PP rule is triggered. Among these 64 sentences, 52 are transformed correctly, while the remaining 12 contain some error. Note that the variants that are identical to Standard German (percentages displayed in parentheses) always perform correctly in 100% of cases, even if the transformation rule is not triggered at all due to an incorrect *Fips* analysis.

While the results are satisfying for local phenomena (genitive noun phrases, dative noun phrases, article doubling contexts), they are close to 0% for most other phenomena. We discuss the main reasons for this poor performance in the following section.

6.6.5. Analysis of *Fips* parsing errors

In the following paragraphs, the main types of errors are reported for each rule.

Preterit tense in matrix clauses In incompletely analyzed sentences, *Fips* fails to correctly identify the right verb bracket and misplaces the past participle. Moreover, verb particles, which provide an easy localization of the right verb bracket, are not always identified.

Preterit tense in subordinate clauses *Fips* often fails to attach subordinate clauses to the main

clause; in consequence, the transformation rule does not recognize them as such.

Genitive in prepositional phrases *Fips* suffers from an unification problem, in the sense that many dependents of the preposition are not unambiguously tagged with genitive case.

Genitive in common noun attributes On the one hand, *Fips* overgenerates genitive attributes in contexts where there should not be any. On the other hand, *Fips* misanalyzes some genitive attributes as ethical datives (there is no morphological distinction for feminine noun phrases).

Genitive in proper noun attributes *Fips* is unable to predict the case of out-of-vocabulary words (which most proper nouns are). As a result, a lot of genitive attributes are missed.

Person name determiners Like above, *Fips* is unable to predict the gender of out-of-vocabulary words, which is used by our heuristic to identify person names. Additionally, appositions like job titles are often incorrectly attached.

Verb raising and verb projection raising As explained above, subordinate clauses are not always attached to a main clause and are therefore not detected. This systematically holds for complex subordinate clauses which would trigger verb projection raising.

Prepositional dative marking Many dative noun phrases are misanalyzed as accusative or genitive noun phrases, which prevents the triggering of the prepositional dative marking rule. Moreover, the CoNLL conversion algorithm does not cover interrogative pronouns.

Complementizer in wh-phrases The CoNLL format conversion is incomplete since some *Fips* attachment rules that are relevant for wh-phrases have not been documented.

Relative pronouns As mentioned above, subordinate clauses are not reliable attached by *Fips*; relative clauses also fall under this restriction. Furthermore, *Fips* prefers to analyze some relative pronouns as demonstrative pronouns or articles.

Final clauses This phenomenon is again hampered by incomplete analyses as well as incompletely implemented CoNLL conversion rules.

The poor performance of *Fips* is the sum of several problems. First, the coverage and precision of the German grammar is clearly insufficient. Lacking coverage is witnessed by the high number of incompletly analyzed sentences. Lacking precision mainly results from the fact that *Fips* does not use any type of frequency information: this means that rare, but less constrained readings are preferred to frequent, but more constrained readings. Second, there are some evident bugs which are not necessarily related to particular linguistic phenomena, but which may have a big impact on performance. Third, the CoNLL conversion rules are not quite complete and further degrade the effective performance.

One might argue that our syntactic transformation rules are stricter than required. For example, it shouldn't matter if one deals with a nominative or an accusative relative pronoun since they are both rendered the same in Swiss German. However, in our view, a useful German parser should be able to assign grammatical roles and thereby disambiguate case, which is why we strive to keep the rules as precise as possible. A slightly different case arises with proper names, where our rules rely on gender and case information. A parser will always encounter problems with out-of-vocabulary words, and one should probably turn to an external Named Entity Recognition system for this task.

In the light of these results, alternatives to *Fips* should be explored. But with other parsers, other difficulties are likely to arise:

- To keep the CoNLL-X format, one would ideally use a dependency parser. But most German parsers are constituent parsers, and a conversion tool would again be required, with the same potential pitfalls than the ones encountered with *Fips*.

- A popular parser that outputs dependency structures in the CoNLL-X format is *MaltParser*. However, we have not been able to train it on the TIGER treebank after several attempts.[26]

- Our different transformation rules require lemmatization, compound splitting, part-of-speech tagging, morphological disambiguation and parsing of the Standard German input. *Fips* provides all these tasks in a single tool. With alternative parsers, one would likely have to use additional tools (like the *Morphisto-TreeTagger* combination), and unifying

26 On this occasion, we would like to thank Johan Hall for his assistance.

potentially contradictory analyses would be an additional challenge.

Despite these potential difficulties, further work should be directed in improving parsing performance in order to take full advantage of the syntactic transformation rules.

6.6.6. Summary

We have shown that a small number of manually written transformation rules can model the most important syntactic differences between Standard German and Swiss German dialects with high levels of accuracy, provided that the syntactic annotation of the Standard German source text is sound. The major limitation we found with our approach is the insufficient precision of the *Fips* parser, which prevents most transformation rules from applying correctly.

For this experiment, we have utilized sentences from the TIGER treebank, which contains newspaper texts. This is hardly a genre frequently used in Swiss German. Spoken language texts would be more realistic to translate. The *TüBa-D/S* treebank (Hinrichs, Bartels, et al. 2000) provides syntactically annotated speech data, but its lack of morphological annotation and its diverging annotation standard have prevented its use in our research for the time being.

6.7. Putting the pieces together

So far, we have evaluated the word-level transfer rules (Sections 6.3, 6.4 and 6.5) and the syntactic transfer rules (Section 6.6) separately. However, the ultimate goal of Swiss German sentence generation is to combine the two sets of rules. Here, we briefly report on the benefits of this combination.

The word-level rules required a morpho-syntactic analysis of the Standard German source words consisting of part-of-speech tagging, lemmatization and compound splitting. In our experiments, this analysis was either given (e.g. in the TIGER treebank) or carried out by the *TreeTagger-Morphisto* pipeline. On the other hand, the syntactic rules require full syntactic

GENERATED	ORIGINAL				
	BA	*BE*	*OS*	*WS*	*ZH*
STANDARD	80.60%	**82.77%**	80.68%	**77.61%**	82.53%
BA	**80.94%**	81.29%	79.81%	75.88%	81.33%
BE	76.82%	79.91%	75.81%	72.50%	77.97%
OS	77.40%	79.57%	78.13%	72.78%	78.97%
WS	76.57%	79.17%	75.04%	75.82%	78.16%
ZH	80.71%	**82.51%**	**81.26%**	**76.54%**	**83.30%**

Table 6.19.: LCSR values obtained with the *TreeTagger*, and word-level rules only (same as Table 6.5).

GENERATED	ORIGINAL				
	BA	*BE*	*OS*	*WS*	*ZH*
STANDARD	**80.60%**	**82.77%**	**80.68%**	77.61%	**82.53%**
BA	80.18%	81.05%	77.89%	75.55%	80.20%
BE	79.16%	81.80%	77.02%	75.03%	80.51%
OS	79.96%	81.94%	79.33%	75.55%	81.76%
WS	79.95%	81.77%	77.67%	**78.53%**	81.34%
ZH	**80.25%**	**82.30%**	**79.70%**	76.27%	**82.26%**

Table 6.20.: LCSR values obtained with *Fips*, and word-level rules only.

GENERATED	ORIGINAL				
	BA	*BE*	*OS*	*WS*	*ZH*
STANDARD	80.60%	82.77%	80.68%	77.61%	82.53%
BA	83.26%	83.04%	81.67%	78.07%	82.37%
BE	81.80%	83.70%	80.73%	77.32%	82.27%
OS	83.05%	84.06%	83.36%	77.96%	84.37%
WS	82.98%	83.59%	81.50%	**81.57%**	83.34%
ZH	**83.32%**	**84.22%**	**83.73%**	78.70%	**84.76%**

Table 6.21.: LCSR values obtained with *Fips*, including word-level and syntactic transfer rules.

parsing; in the preceding experiments, we either relied on the gold standard annotation of TIGER, or on the *Fips* parser. For the combined system, the latter will be required as well. In this section, we use the experimental setup of Section 6.4, in which translation quality is measured in terms of Longest Common Subsequence Ratio with respect to a dialectal reference text. We again use the *Wikipedia* corpus presented in Section 6.2.

As a baseline, we refer to the system proposed in Section 6.4.3, consisting of the *TreeTagger-Morphisto* preprocessing pipeline and the finite-state implementation of the word-level transformation rules. The results of this system, first reported in Table 6.5 on page 212, are reproduced for convenience in Table 6.19.

In a first step, the preprocessing pipeline is replaced by *Fips*, keeping the transformation rules unchanged (i.e., the syntactic transfer rules are not added yet). This change allows us to measure the impact of the preprocessor. Indeed, the experiments reported in the preceding section suggests that *Fips* performs less well on this task than the *TreeTagger-Morphisto* combination, for the main reason that *Fips* does not use frequency information to disambiguate homographic forms. The results of this system are reported in Table 6.20. As expected, it performs worse than the baseline: for four (instead of two) dialects, the best solution is the one without any transformation (i.e., the Standard German original), and the only improvement concerns the Wallis dialect.

Furthermore, the results for the different dialects lie closer together, which means that this system is less able to discriminate different dialects. This finding may sound surprising since the only change lies in the preprocessing of the Standard German source text, not in the transformation rules properly speaking. The reason of this effect probably lies in verb morphology, which is subject to high ambiguity in Standard German as well as high dialectal variation in Swiss German. If the preprocessor wrongly disambiguates the Standard German forms, the wrong dialect forms will be generated in turn.

In a second step, the syntactic transformation rules are added to the word-level rules, keeping the *Fips* parser for preprocessing. By comparing the results of this second step, shown in Table 6.21, to the ones of the first step, the impact of the syntactic transformations can be measured. It turns out that the integration of the syntactic rules improves the accuracy of the generated dialect texts by 2-3% for all dialects. Thereby, the gap between the generated dialect and the

Standard German original increases. This means that the syntactic rules do help in creating more adequate dialect text, but they do not help in discriminating one dialect from another. This is not surprising since the most important syntactic transformations (genitive replacement and past tense replacement, according to Table 6.16 on page 233) do not constitute clear isoglosses between the five dialects of our test corpus. Nonetheless, we are pleased to see that the syntactic transformations are efficient despite the parsing inaccuracies encountered with *Fips*.

6.8. Conclusion

In this chapter, we have reported some experiments intended to evaluate different aspects of the dialect generation system. The results obtained in the various experiments are encouraging: it is indeed possible to translate Standard German words into Swiss German, so that the latter can be correctly analyzed or identified. Since there is no previous work on translating into multiple Swiss German dialects, there is no benchmark with which we could compare our findings. Still, we believe that better performance can and should be obtained.

First, the impact of each rule is insufficiently known. We intend to provide an automated procedure that is able to quantify the error reduction (or error increase) obtained for every single rule. This will allow us to identify problematic rules and re-engineer them so that they become truly useful.

Second, the *Fips* parser, required for the correct application of the syntactic rules, does not quite reach state-of-the-art performance, and alternatives should be investigated. Using a parser that can directly be trained on the TIGER corpus would obviate the need for error-prone and tedious to write format conversion scripts.

7. Dialectometry

7.1. Introduction

In the last few decades, **dialectometry** has emerged as a field of linguistics that investigates the application of statistical and mathematical methods in dialect research. Also called **quantitative dialectology**, one of its purposes is to discover the regional distribution of dialect similarities, to characterize it mathematically and to display it visually. Dialectometric research is generally based on linguistic atlas data. Such atlases provide information for a large number of linguistic phenomena at a large number of inquiry points. Thus, dialectometric methods are said to be based on aggregated dialect data.

Dialectometry has been first introduced by Séguy and Goebl in order to address some shortcomings of traditional dialectological research (see Section 2.6). At that time, dialectologists studied single linguistic phenomena and their geographical distribution in isolation. Even if dialect atlas projects provided them with large sets of linguistic phenomena, they did not have adequate tools to synthesize this mass of data into comprehensive pictures of entire dialect landscapes. Rather, dialect classifications were obtained on the basis of a small number of isoglosses, chosen according to the intuitions of the dialectologist. The resulting dialect landscapes therefore could hardly be considered objective. While of undeniable intellectual quality, the existing classifications of the Swiss German dialects follow this traditional approach: Bohnenberger (1953); Haas (2000); Hotzenköcherle (1984).

The availability of large data sets (atlases) and powerful computers at the same time made possible and necessary the search for new methods that allow to describe dialect landscapes in a synthetic, visual manner. The key concept of dialectometrical studies are **linguistic distances**:

the linguistic distance between two dialects is measured numerically on the basis of a large, unweighted set of linguistic phenomena. These distance values allow to detect and visualize spatial patterns in dialect data, providing dialect classifications that avoid subjective decisions on particular isoglosses.

Astonishingly, German-speaking Switzerland – one of the liveliest dialect areas of Central Europe – has been largely absent from this strand of research. A notable exception is the study by Kelle (2001), which will be discussed below. This lack is mainly due to the absence of digitally usable data sets: the SDS was completed entirely without the help of computers, and the SADS data have only become available very recently. The SDS data that we have digitized for the use in our transfer rules (Section 3.3) can thus be reused profitably for novel dialectometric research. Other data sets recently made available, such as the SADS data (Section 3.4) and the Archimob corpus (Section 7.3.5), are also studied in this chapter. The intent of this chapter is thus to provide new insights into the Swiss German dialect landscape by using data sources that have been originally intended for the dialect generation engine described in this thesis.

In Section 7.2, we explain some of the most popular techniques and methods used in recent dialectometric research. In Section 7.3, these techniques are applied to the different data sets mentioned above. Section 7.4 concludes the chapter.

7.2. Methods

The main element of a dialectometric study is a **distance matrix** that quantifies the linguistic distances between all pairs of data points. In Section 7.2.1, we explain how distance matrices are derived from dialectological raw data such as those found in dialect atlases. We then show how numerical analyses can be used to compare different data sets and different measures of linguistic distance, and how these linguistic distances correlate with the underlying geographical distances between inquiry points (Section 7.2.2). Furthermore, we present two techniques of data analysis and visualization: hierarchical clustering (Section 7.2.3) and multidimensional scaling (Section 7.2.4).

The following studies are largely inspired by the Groningen school in terms of techniques and

Location	*"hier"*	*"Leute"*	*3 Pl*
Köniz (BE)	*hie*	*-ü-*	*-e*
Wolhusen (LU)	*daa/doo*	*-ü-*	*-e*
Hergiswil (UW)	*hie*	*-i-*	*-id*
Niederwald (WS)	*hie*	*-i-*	*-end*
Horgen (ZH)	*daa/doo*	*-ü-*	*-ed*
Flawil (SG)	*daa/doo*	*-ü-*	*-ed*
Pratteln (BA)	*daa/doo*	*-ü-*	*-e*

Table 7.1.: Data matrix containing a small sample of SDS data: three linguistic variables at seven locations.

tools used (Heeringa and Nerbonne 2001; Nerbonne and Kleiweg 2005; Nerbonne and Siedle 2005). In particular, we use the freely available RuG/L04 tool[1] and its online successor Gabmap (Nerbonne, Colen, et al. 2011) for our analyses and map drawings.

Note that the density kernel estimation method used in Chapter 4 is, like other dialectometrical techniques, a method developed in spatial statistics. However, it is not a dialectometrical technique in its narrow sense, since it yields a separate density map for every linguistic phenomenon and thus does not aggregate the data of different phenomena.

7.2.1. Distance matrices

The reader may be familiar with geographic distance tables that indicate distances between the major cities of a country and that are commonly found on the back of road maps. Such tables are symmetric (i.e., the distance from A to B is equal to the distance from B to A, for any A and B) and contain a diagonal full of zeros (i.e., the distance from A to A is zero, for any A).[2] Linguistic distance matrices are based on the same principle; but while geographic distance tables measure the distances in terms of kilometers or miles, linguistic distance tables use other, more abstract units of distance.

1 http://www.let.rug.nl/kleiweg/L04/

2 For formal definitions of *distance* and *metric*, see for example J. Kruskal (1999).

	Köniz	Wolhusen	Hergiswil	Niederwald	Horgen	Flawil	Pratteln
Köniz	**0.00**	0.33	0.67	0.67	0.67	0.67	0.33
Wolhusen	0.33	**0.00**	1.00	1.00	0.33	0.33	0.00
Hergiswil	0.67	1.00	**0.00**	0.33	1.00	1.00	1.00
Niederwald	0.67	1.00	0.33	**0.00**	1.00	1.00	1.00
Horgen	0.67	0.33	1.00	1.00	**0.00**	0.00	0.33
Flawil	0.67	0.33	1.00	1.00	0.00	**0.00**	0.33
Pratteln	0.33	0.00	1.00	1.00	0.33	0.33	**0.00**

Table 7.2.: Distance matrix obtained from the dialectological data presented in Table 7.1. The zero values on the symmetry axis are highlighted.

In this section, we show how dialectological raw data can be transformed into a distance matrix. Table 7.1 shows a simple example containing the realizations of three linguistic variables at seven locations. The first variable is lexical and shows two possible realizations of Standard German *hier* 'here' (SDS VI/098).[3] The second variable is phonetic and concerns the roundedness of the monophthong in words like Standard German *Leute* 'people' (SDS I/156, see also Appendix A.2.1, p. 292). The third variable is a morphological one and shows different realizations of the third person plural ending of regular verbs (SDS III/033, see also Appendix A.4.6, p. 357). Such a table, which represents locations on the rows and linguistic phenomena on the columns, is called **data matrix**.

The distance between two locations can now be defined as the proportion of variables whose realizations are different. For example, the locations Köniz and Wolhusen only differ with respect to the first variable. Hence, their distance value is $\frac{1}{3} = 0.33$. Likewise, the distance between Wolhusen and Niederwald is $\frac{3}{3} = 1$, since all three phenomena are realized differently in these two locations. Horgen and Flawil share all variants and thus have a distance of 0. The distance between a location and itself is always defined as 0. The complete distance matrix of this example is reproduced in Table 7.2.

Aside from the small number of locations and linguistic phenomena, this example relies on some simplifying assumptions that do not necessarily hold for the larger data sets used for our

3 The distinction between the variants *daa* and *doo* is purely phonetic and does not concern us here. See below for a discussion.

experiments.

First, we assume that one unambiguous answer has been reported at every location. In other words, we suppose that a single informant has been asked at every location, and that all informants committed themselves to a single answer. In the SDS data set, this assumption mostly holds. However, there are a few cases where two or more variants are reported at the same inquiry point. We simply consider the one variant at each point that yields the minimal distance value. The SADS data set is more complex since the number of informants per inquiry point is variable. Each inquiry point is represented by a frequency distribution vector over all possible variants. Distances between two inquiry point vectors are defined as Euclidean distances.

Second, we assume that the observations are categorical: two variants are either identical or different. For instance, we do not admit intermediate values between *ü* and *i*, and we consider that the phonetic distinction between *daa* and *doo* is not relevant for the lexical variable. In the SDS and SADS data sets, the raw data has already been categorized by the atlas editors (if the linguistic material was not of categorical nature in the first place).

Third, we assume that all linguistic phenomena are equally important. One might argue that the *ü–i* alternation appears in many lemmas and should therefore weighted more heavily than the *hie–daa/doo* alternation, which only appears in a single lemma. However, it is difficult to find an objective measure of the frequency of a phenomenon. Therefore, we treat all phenomena with equal importance.

Nerbonne and Heeringa (2001) discuss such issues in more detail. In particular, they show several methods of assessing the phonetic distance of word pairs by using metrics based on Levenshtein distance. They also present several frequency weighting schemes.

7.2.2. Spatial autocorrelation

A general postulate of spatial analysis is that "on average, values at points close together in space are more likely to be similar than points further apart" (Burrough and McDonnell 1998, p. 100). This idea that the distance of attribute values correlates with their geographical distance is known

as **spatial autocorrelation**. The same idea has been coined the **fundamental dialectological postulate** by Nerbonne and Kleiweg (2005, p. 10): "Geographically proximate varieties tend to be more similar than distant ones."[4]

This postulate can be used to compare different measures of linguistic distance D_A and D_B: if D_A correlates better with geographic distance than D_B does (applied to the same data set), then D_A can be considered better than D_B. Likewise, if one subset of data correlates better with geographic distance than another subset (keeping the distance measure constant), the former subset can be considered more reliable than the latter.

We use two tests to compute spatial autocorrelation values between two distance matrices: local incoherence, and the Mantel test.

Local incoherence has been proposed by Nerbonne and Kleiweg (2005). The idea of this measure is that the correlation between linguistic and geographic distances is local and does not need to hold over larger geographical distances. In practice, for every data point, the eight linguistically most similar points[5] are inspected according to their linguistic distance value. Then, the geographic distance of these pairs of points is measured and summed up. This means that high incoherence values represent poor measurements, while lower values stand for better results.

The **Mantel test** (Sokal and Rohlf 1995, pp. 813-819; Heeringa 2004, ch. 3) is a general statistical test which applies to data expressed as dissimilarities. It is often used in evolutionary biology and ecology, for example, to correlate genetic distances of animal populations with the geographic distances of their range. The Mantel coefficient Z is computed by computing the Hadamard product (see Section 5.2.12) of the two matrices. The statistical significance of this coefficient is obtained by a randomization test. A sample of permutations is created, whereby the elements of one matrix are randomly rearranged. The correlation level depends on the proportion of samples whose Z-value is higher than the Z-value of the reference matrix.

4 Of course, this assumption is simplified, as factors like topography and socio-cultural factors often participate in explaining dialect change. In the studies reported here, we do not take into account these additional factors.

5 The restriction to eight points is the key of the local component of this measure. The exact value of this parameter has been determined empirically by the authors of the measure.

7.2.3. Clustering

Cluster analysis is an important technique of data analysis and is used in a variety of fields such as biology, psychology and marketing (Jain and Dubes 1988). A variant of cluster analysis, hierarchical clustering, has become very popular in dialectometry:

> The research community is convinced that the linguistic varieties are hierarchically organized; thus, e.g., the urban dialect of Freiburg is a sort of Low Alemannic, which is in turn Alemannic, which is in turn Southern German, etc. This means that the techniques of choice have been different varieties of hierarchical clustering […] (Nerbonne, Kleiweg, et al. 2008)

Hierarchical clustering works as follows:

> A hierarchical cluster analysis identifies clusters of similar objects in a distance matrix by initially assigning each observation to its own cluster and by then repeatedly combining the two most similar clusters to form larger and larger clusters until all of the objects have been combined to form one large cluster. Various methods exist for measuring the similarity between clusters consisting of multiple observations […] (Grieve to appear)

While the similarity between clusters consisting of single observations can be read off directly from the distance matrix, the similarity between complex clusters (consisting of several observations) may be defined in one of the following ways:

Single linkage: Minimum value of all pairwise distances.

Complete linkage: Maximum value of all pairwise distances.

Average linkage: Arithmetic mean of all pairwise distances.

Centroid: Distance between the centroid of cluster *A* and the centroid of cluster *B*. The centroid of a cluster is obtained by averaging the elements it consists of.

Ward's method: Distance between the centroid of cluster *A* and the centroid of cluster *B*, weighted in order to minimize the increase of variance as the number of clusters drops.

The clustering algorithms can be weighted by the size of the clusters, so that patterns in small clusters are weighted more heavily than patterns in large clusters. The most popular weighted algorithms are **weighted average linkage** and **weighted centroid**. More detailed descriptions of these clustering algorithms can be found in Jain and Dubes (1988).

The result of hierarchical clustering is a **dendrogram**: a tree that shows the successive steps of the cluster analysis. Each node represents a cluster, consisting of the two sub-clusters on its left. The position of a cluster on the *x*-axis represents the distance between its two subclusters; this value is called **cophenetic distance**. The dendrograms in Figures 7.1, 7.2 and 7.3 illustrate the outcomes of different clustering algorithms. They are all based on the distance matrix of the toy example presented above. The differences between the algorithms show up most clearly in the cophenetic distance value between the Flawil/Horgen cluster and the Köniz/Pratteln/Wolhusen cluster: this value is 0.33 with Single Linkage, 0.66 with Complete Linkage, and 0.44 with Average Linkage.

The groups obtained by clustering can be projected onto a geographical map. The dendrogram is cut vertically so that a fixed number of clusters is determined. For example, cutting the dendrogram of Figure 7.2 at 0.5 yields 3 clusters;[6] cutting it at 0.8 yields 2 clusters;[7] cutting it at 0.2 yields 5 clusters.[8] Then, a different color is chosen for each cluster. All data points (or rather, their corresponding Thiessen polygons) are drawn in the color of the cluster they pertain to.

There are two major drawbacks with hierarchical clustering. First, there are no guidelines as to which clustering algorithm performs best in what circumstances, nor to how many clusters should be displayed on the map. Results obtained with a single algorithm should therefore not be overinterpreted. Second, clustering is known to be unreliable, in the sense that small changes in the distance matrix may lead to drastic changes in the dendrogram. Third, the color maps

6 Cluster 1: Flawil, Horgen; cluster 2: Köniz, Pratteln, Wolhusen; cluster 3: Hergiswil, Niederwald.

7 Cluster 1: Flawil, Horgen, Köniz, Pratteln, Wolhusen; cluster 2: Hergiswil, Niederwald.

8 Cluster 1: Flawil, Horgen; cluster 2: Köniz; cluster 3: Pratteln, Wolhusen; cluster 4: Hergiswil; cluster 5: Niederwald.

resulting from hierarchical clustering suggest clear-cut boundaries at the cluster borders. This does not necessarily reflect the dialectological reality, where dialect change is often viewed as gradual.

These issues can been alleviated by bootstrap clustering or noisy clustering. **Bootstrap clustering** creates different samples of the original data set and uses each of them for a clustering run. The reliability of a cluster is then defined as the number of runs in which it has been obtained. As an alternative approach, Nerbonne, Kleiweg, et al. (2008) propose **noisy clustering**. The idea is that the stability of the clusters can be tested by adding small amounts of random noise. The clustering process is repeated n times, and at each run, different amounts of noise are added to the different cells of the distance matrix. Again, this allows to obtain reliability figures for each cluster. This yields a **composite dendrogram** which reports at the same time the mean cophenetic distances between clusters and their reliability (see Figure 7.16 for an example). These results can be visualized on color maps with a combination of noisy clustering and multidimensional scaling. We explain this technique below.

7.2.4. Multidimensional scaling

On world maps, the areas around the Poles are often distorted. This is a consequence of projecting a three-dimensional object (the Earth) onto a two-dimensional object (a sheet of paper). While this reduction of dimensionality leads to distortion, it still depicts the surfaces of Greenland and the Antarctica in a reasonably faithful way. This is the idea of multidimensional scaling: reduce the dimensionality of the data by minimally distorting the distance values.

Linguistic distance matrices generally define high-dimensional spaces: in order to display n locations proportionally to their distances, one needs at most $n - 1$ dimensions. For instance, the SDS data set contains over 600 locations, resulting in a high-dimensional space that is rather impractical to visualize. Multidimensional scaling (MDS) allows high-dimensional spaces to be reduced onto two or three dimensions that can then be displayed easily. The following example explains the idea of MDS in more detail.

According to the linguistic distances of Table 7.2, the three locations Hergiswil, Köniz and

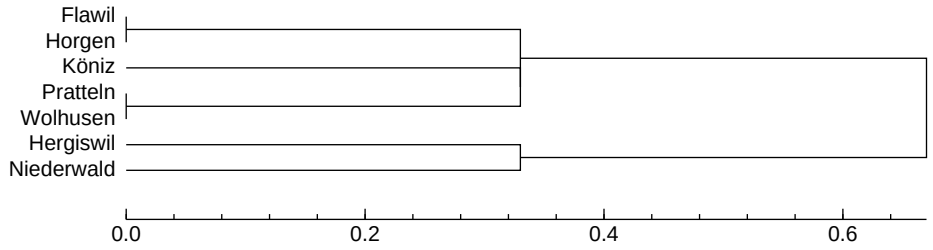


Figure 7.1.: Dendrogram obtained with the Single Linkage algorithm.

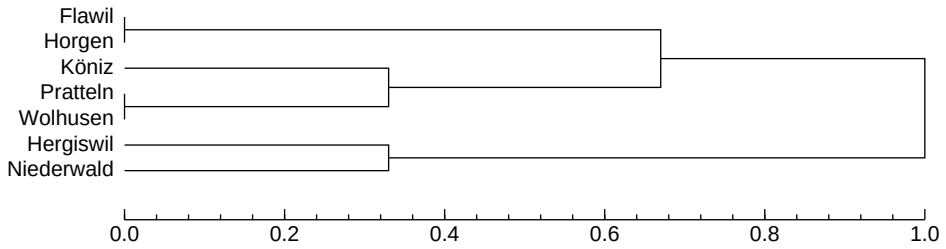


Figure 7.2.: Dendrogram obtained with the Complete Linkage algorithm.

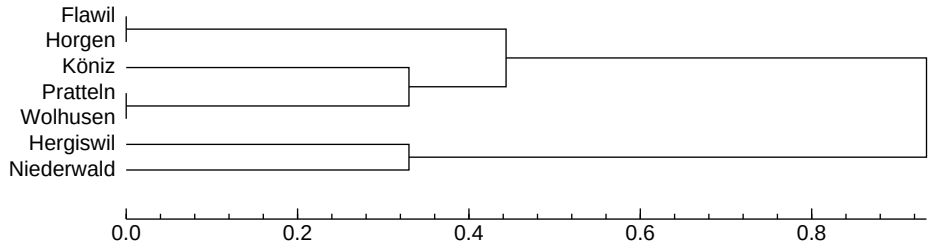


Figure 7.3.: Dendrogram obtained with the Average Linkage algorithm.

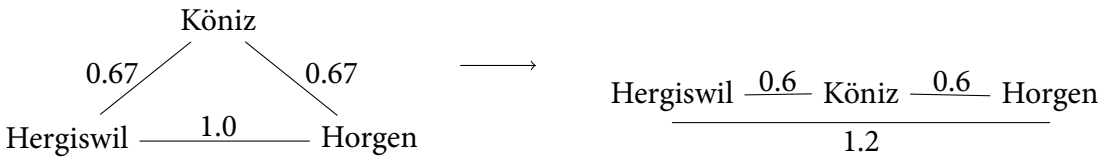


Figure 7.4.: Example of multidimensional scaling: the distance values between vertices in the (two-dimensional) triangle are modified so that they can be placed on a (one-dimensional) line.

Horgen define a triangle (see Figure 7.4, left). By reducing the distance values between Hergiswil and Köniz and between Köniz and Horgen (from 0.67 to 0.6), and by increasing the distance value between Hergiswil and Horgen (from 1.0 to 1.2), the three locations can be placed on a straight line (see Figure 7.4, right). In consequence, the dimensionality of this graph is reduced from 2 dimensions (triangle) to 1 dimension (line). The exact definition of which distance values are adapted in which way depends on the MDS algorithm. The degree of change required to accommodate the locations in a lower-dimensional space is known as **stress**.

The general algorithm for multidimensional scaling is presented in T. F. Cox and M. A. Cox (1994). In dialectometry, MDS has been first applied by Embleton (1993). The values obtained by MDS are continuous. In contrast to cluster analysis – where each cluster is a discrete entity –, MDS yields more gradual representations of dialect landscapes.

Each dimension extracted by multidimensional scaling represents a specific pattern of regional variation and can thus be interpreted in isolation. However, it is more common to display two or three dimensions simultaneously. The most straightforward visualization technique reduces the original data to two dimensions, plots the resulting values on a two-dimensional grid and compares the latter to the equally two-dimensional geographical distribution of the data points. An example of this technique is given in Figure 7.18 below. Another representation type has been introduced by Heeringa (2004, ch. 6). He reduces the data to three dimensions and associates each location on the map with a color whose RGB (red, green, blue) parameters are defined by the three dimension values obtained for that location by MDS. This type of map has been coined **rainbow map** ("Regenbogenkarte", Nerbonne and Siedle 2005). Figure 7.8 shows an example.

A different visualization technique combines noisy clustering and multidimensional scaling. First, noisy clustering is applied to create a set of dendrograms. Each dendrogram can be viewed, in fact, as another distance matrix where each pair of locations is associated with its cophenetic distance. All these distance matrices are averaged and submitted to multidimensional scaling. The resulting three dimensions are interpreted as RGB color values and drawn on a geographic map (see Figure 7.9 for an example).[9]

7.3. Experiments

7.3.1. Inferring dialect landscapes with SDS data

To our knowledge, Kelle (2001) is the only dialectometric study using data from German-speaking Switzerland. Its author digitized a subset of the SDS data in order to perform a cluster analysis and confront the findings with traditional dialect classifications. Kelle selected about every second map of the SDS volumes I to III (170 maps in total) and reduced the number of inquiry points to 101 (of roughly 600 in the original SDS data). He draws maps with up to six clusters, obtained by means of the Complete Linkage clustering algorithm. Figure 7.5 reproduces one of his images.

The resulting dialect landscape is characterized by four major regions: North West, South West, North East, and South East. Two minor regions are added: Wallis on the one hand and the Alemannic-speaking locations in Northern Italy on the other hand (displayed on the inlay square). This classification roughly corresponds to traditional dialectological knowledge (see also Section 1.2.5):

> Es ergibt sich mit den Grenzen der anderen Cluster zusammen ein starker, nur im Bereich Luzern etwas getrübter Anklang an die vielzitierte "Brünig-Napf-Reuß-Linie", die an der Aaremündung ihren Ausgang nimmt und am Furka-Paß endet.[10]

9 The algorithm of this technique is described in more detail in http://www.let.rug.nl/kleiweg/L04/Tutorial/t07.html.en.

10 "The result, together with the borders of the other clusters, is a strong convergence – blurred only in the region

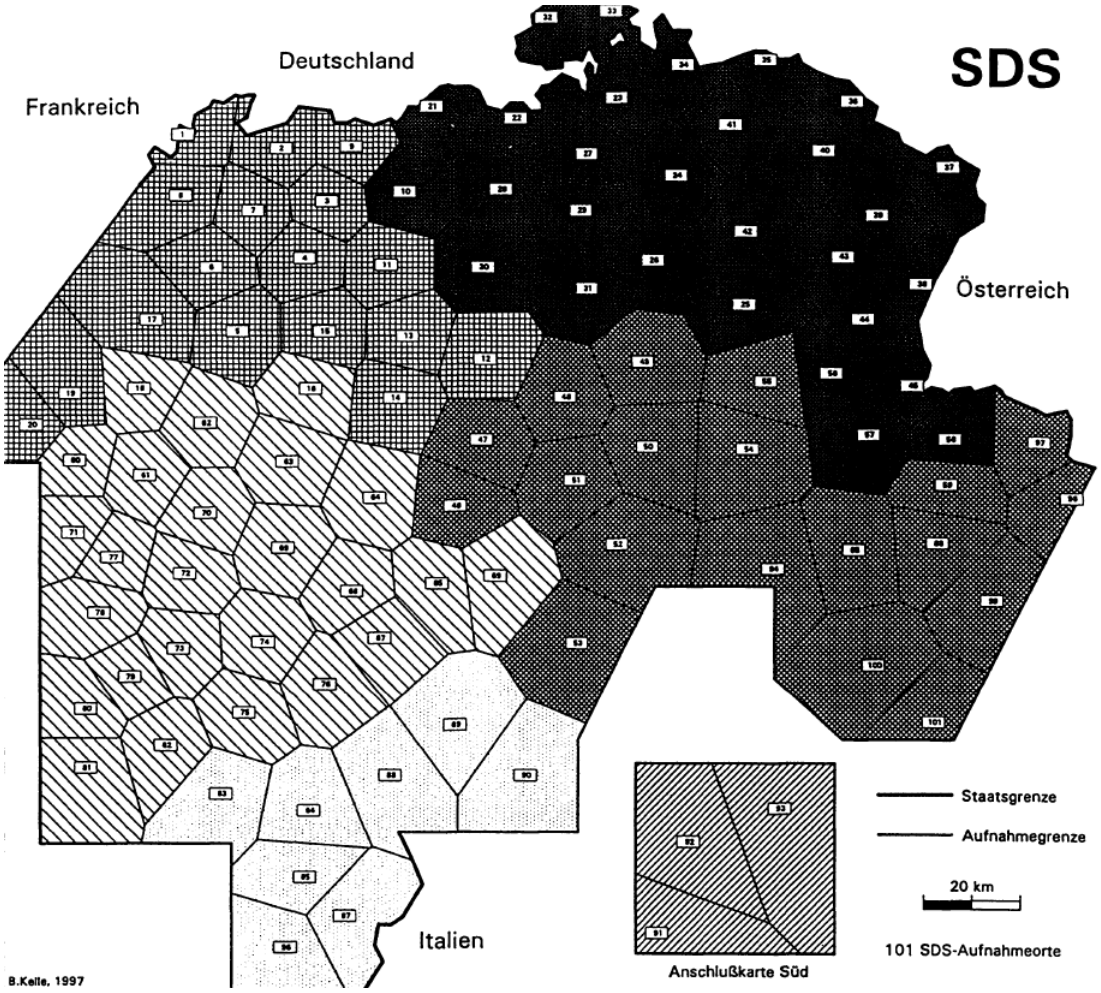


Figure 7.5.: Cluster map by Kelle (2001), representing six major dialect regions obtained with the Complete Linkage algorithm.

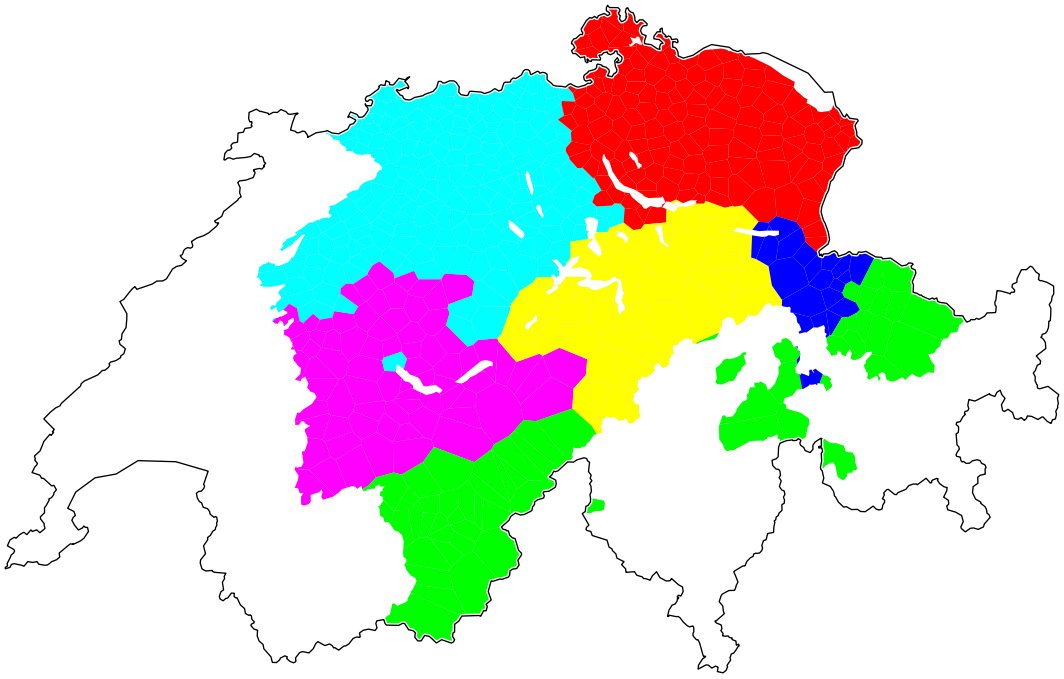


Figure 7.6.: Six clusters obtained with the Complete Linkage algorithm on the SDS data set.

(Kelle 2001, p. 29)

Our SDS data sample differs from Kelle's in several regards. First, we use the entire set of inquiry points except the 8 Italian locations. Second, the selection of the maps is more biased since maps were chosen according to their relevance for the text generation system. Our sample consists of 59 phonetic maps (Volumes I and II), 110 morphological maps (Volume III) and 27 lexical maps (Volumes IV to VIII). Globally, the number of maps is similar (196 in our sample, 170 in Kelle's).

In order to compare our data set with the one used by Kelle, we start with an identically created cluster map, based on the same algorithm and the same number of clusters (Figure 7.6). The global disposition of the different dialect area coincides quite neatly between the two maps. However, there are numerous differences on a smaller scale: the Walser dialects are attached to

of Lucerne – towards the much-cited 'Brünig-Napf-Reuss line', which originates at the mouth of the Aare and ends on the Furka pass."

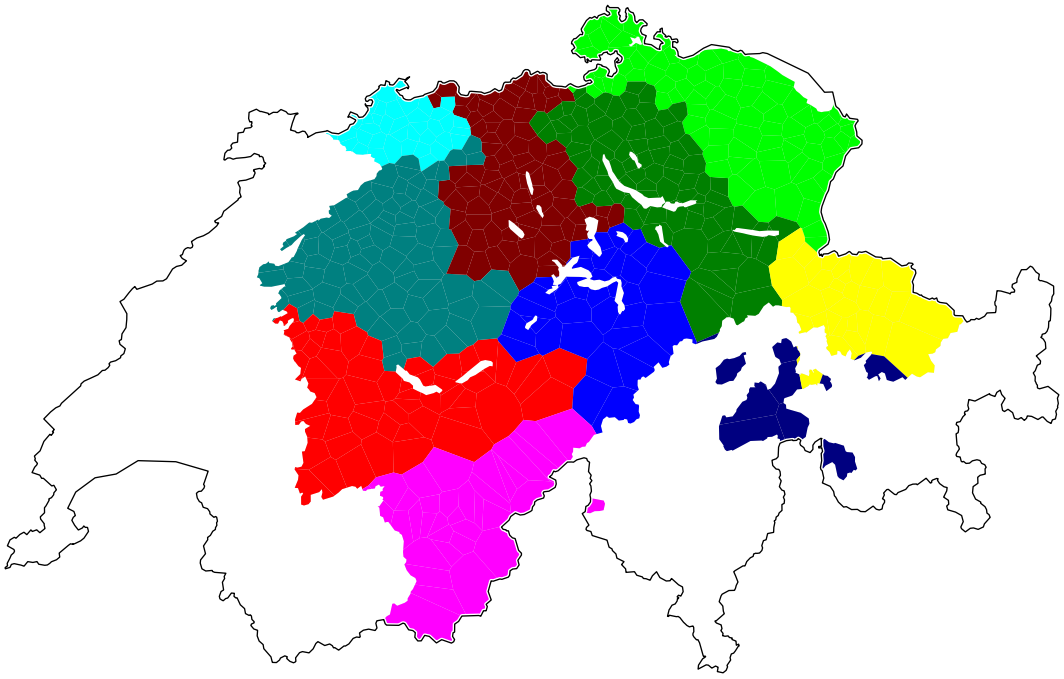


Figure 7.7.: Ten clusters obtained with the Weighted Average algorithm on the SDS data set.

the Wallis dialects; a small cluster contains the dialects in the upper Rhine Valley; the four main areas on the North of the Alps are arranged differently than on Kelle's map. A slightly different data sample thus results in a rather different cluster structure, even when the algorithm and the number of clusters are kept constant.

The Complete Linkage algorithm is not a common choice for dialectometric studies. Other methods, like Weighted Average Linkage, have been found to yield more satisfying results when applied to dialect data (Nerbonne and Siedle 2005). Moreover, even with the relatively small sample size, we believe that it can be instructive to increase the number of displayed clusters. Figure 7.7 shows a map with ten clusters obtained with Weighted Average.

In this map, the following borders concur with traditional dialectological knowledge:

- the border between Wallis and Bern (see Hotzenköcherle 1984, ch. 1, 9),
- the Northern half of the border between Bern and Luzern (see Hotzenköcherle 1984, ch. 2),

- the border separating Berner Oberland and Fribourg from the Bern lowlands regions (see Hotzenköcherle 1984, ch. 1, 10),
- the small dialect region around Basel (see Hotzenköcherle 1984, ch. 4),
- the border between Zürich and Northeastern Switzerland including Schaffhausen (see Hotzenköcherle 1984, ch. 6),
- the border between Aargau and Zürich (see Hotzenköcherle 1984, ch. 1, 5),
- a compact Central Swiss area (see Hotzenköcherle 1984, ch. 12).

Other borders are more surprising:

- the Southern part of Luzern is attached to Bern;
- the grouping of Zürich and Glarus dialects may be contested;
- the bipartition of Grisons does not strictly follow the bipartition in Rhine Valley dialects and Walser dialects.

As explained above, cluster maps may create the illusion of stable, clear-cut dialect borders. We therefore created a rainbow map with multidimensional scaling (Kruskal's method, $r = 0.969$). It is shown in Figure 7.8.

The MDS map contains gradual and smooth transitions between dialects. One of the few clear dialect boundaries is the one between Bern and Wallis. However, a small number of rather compact dialect areas can easily be identified, such as Basel (light blue), Glarus (purple), and the Upper Rhine Valley (red). Again, they reflect traditional dialect regions.

Finally, we experimented with noisy clustering to obtain a more reliable picture of the Swiss German dialect landscape than in the conventional cluster map. The resulting map is reproduced in Figure 7.9.[11] It shows a very clear distinction between alpine and lowland dialects. Among the lowland dialect groups, some borders are rather fuzzy (e.g. between Bern and Berner Oberland, between Zürich and Aargau/Luzern) while others are very sharp (e.g. between Bern and Luzern, between Zürich and St. Gallen). Intuitively, this mapping technique seems to offer the greatest

11 Weighted Average Linkage clustering was repeated 50 times and (Unweighted) Average Linkage clustering again 50 times, all with a noise level of 0.5 times the standard deviation of the data. The color values are determined by multidimensional scaling on the average of the 100 cophenetic distance matrices.

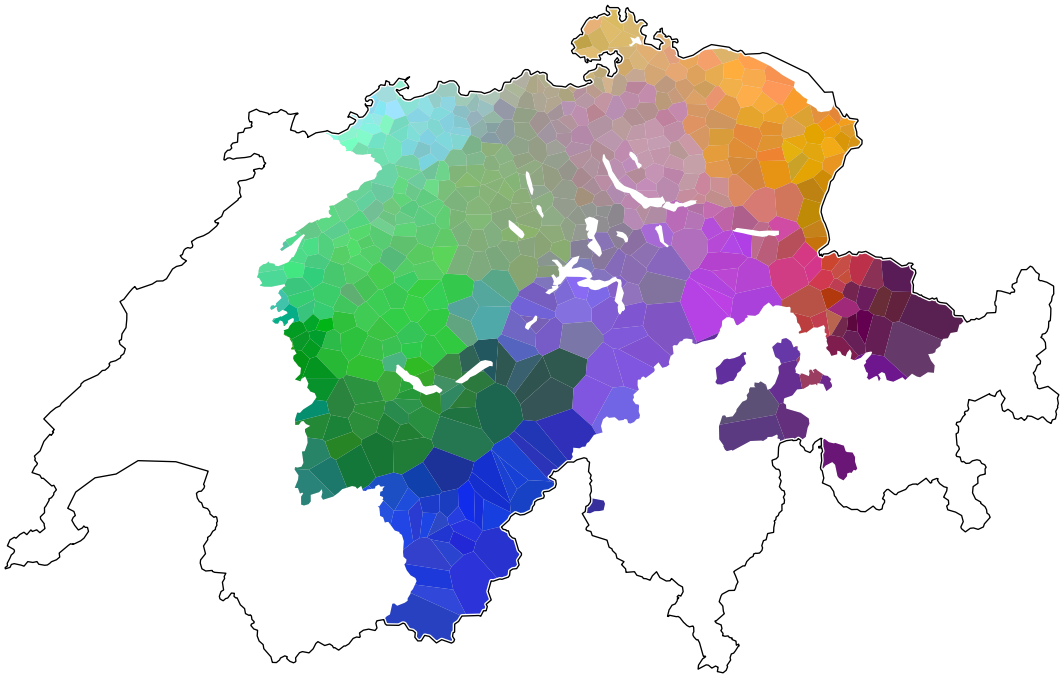


Figure 7.8.: Map resulting from multidimensional scaling applied to the SDS data set.

potential for interpretation.

7.3.2. The influence of linguistic levels on the inferred dialect landscapes

Our SDS sample contains maps for different linguistic levels: phonetics and phonology, morphology, and the lexicon. We argue that each of these levels has specific variation patterns that define different dialect landscapes.

In a first experiment, we define five map subsets: consonants, vowels, verb morphology, other morphology (noun and adjective inflection, pronouns, derivation), lexicon. For all subsets, an average distance matrix is computed and compared with geographical distances by calculating spatial autocorrelation measures. Results are reported in Table 7.3. Local incoherence turns out to be very sensible to the sample size: small samples (consonants, lexicon, vowels) yield high incoherence values. With both correlation measures, the two phonetic subsets obtain lower

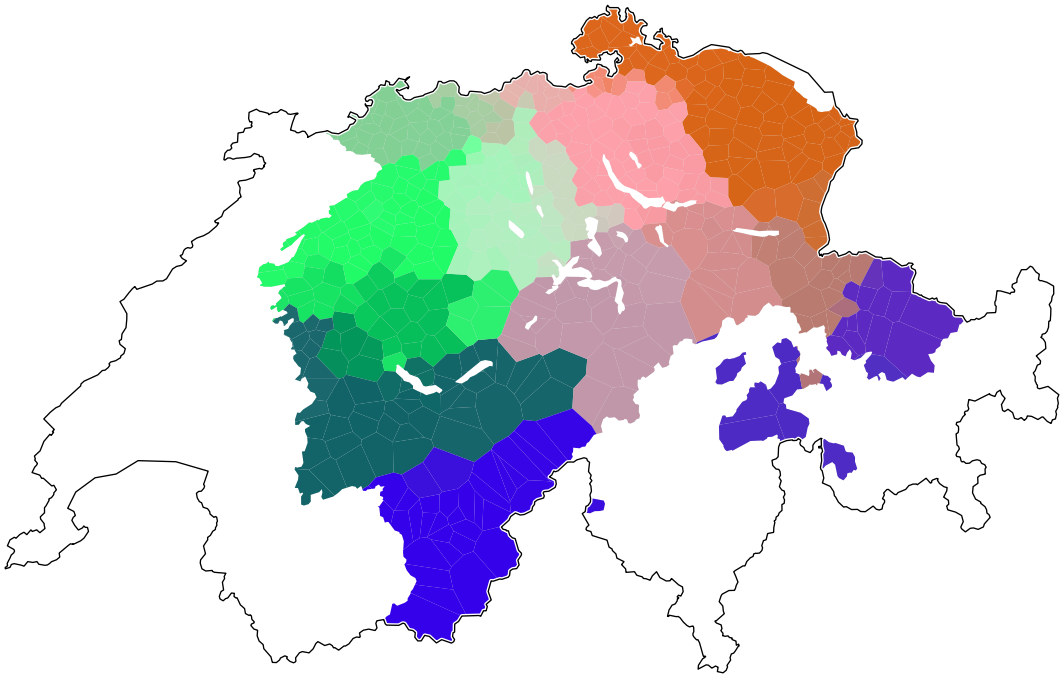


Figure 7.9.: Map created by noisy clustering and multidimensional scaling.

figures than morphological subsets. The lexical subset lies in between them, but its small size only permits tentative conclusions.

In order to homogenize the sizes of the level-specific data sets, we keep all 59 phonological variables as one subset, and build a morphological subset by randomly selecting 59 variables from both verbal and other morphology. Spatial autocorrelation measures for these homogeneous subsets are reported in Table 7.4. They confirm that morphological variables correlate better with geographic distance than the phonological variables.

An explanation for this result may be the different average number of variants per variable (see last column of Table 7.4). As a consequence of the relatively coarse-grained transcription system used in the digitization process, many phonological variables only contain binary oppositions, such as closed vs. open vowel. In contrast, most morphological variables have more than two variants. Keeping in mind that every variant corresponds to an isogloss, one may argue that variables with more variants discriminate better than variables with few variants.

	Maps	Local incoherence	Mantel test r
Consonants	22	3.02	0.45
Vowel	37	1.99	0.47
Verb morphology	71	0.99	0.68
Other morphology	39	1.52	0.62
Lexicon	27	1.65	0.63
All maps	196	0.71	0.70

Table 7.3.: Correlation measures for five subsets of SDS data. All Mantel test values are significant at $p = 0.01$ (99 simulations).

	Maps	Local incoherence	Mantel test r	Variants
Phonology	59	1.27	0.55	2.78
Morphology	59	1.02	0.65	3.18

Table 7.4.: Correlation measures for two balanced subsets of SDS data (correlations significant at $p = 0.01$). The last column reports the average number of variants per variable.

7.3.3. Inferring dialect landscapes with SADS data

The raw data of the SADS survey may also be used for dialectometric experiments. We used all 118 questionnaire items, independently of the inquiry technique used and of the linguistic phenomenon.[12]

First, we created a rainbow map by multidimensional scaling (Figure 7.10). According to this map, the syntactic landscape is fairly uniform in the lowlands of Switzerland; the most important distinctions are located in the Highest Alemannic dialect regions. As with the SDS material, the Wallis region is clearly distinguished. Furthermore, the Fribourg dialect and the dialects of the Eastern Berner Oberland are clearly separated from the rest. Graubünden dialects and Appenzell dialects stand out as well.

12 We did not take into account the "unexpected answers" that were noted manually in the database. We neither used the answers about the most natural variants here (but see Section 7.3.4).

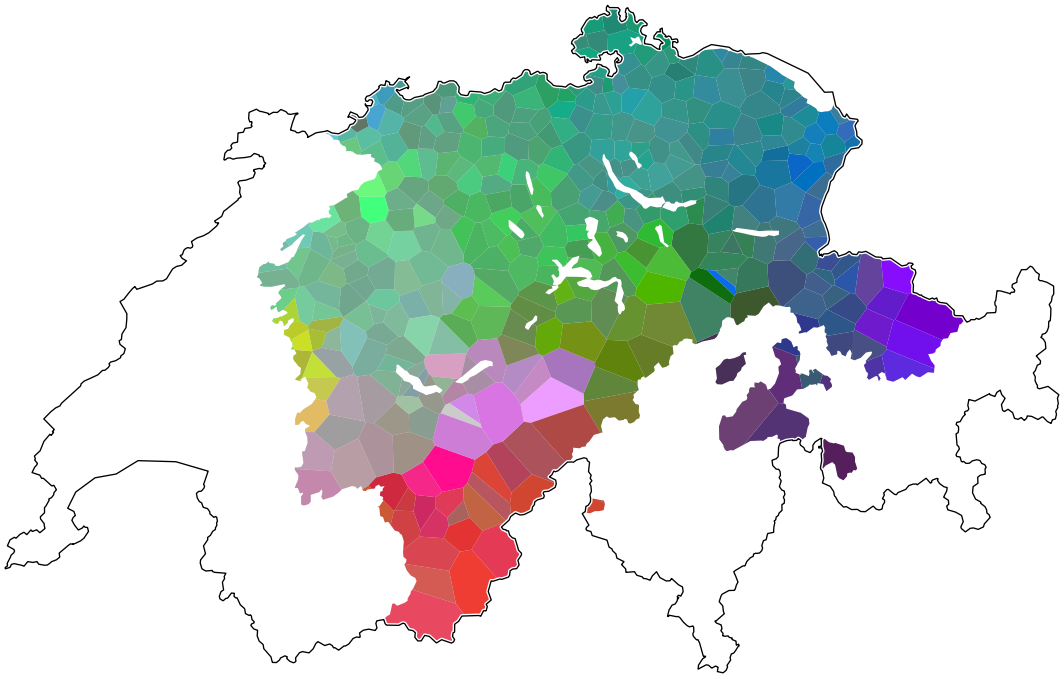


Figure 7.10.: Map resulting from multidimensional scaling applied to the SADS data set.

In contrast, the map obtained by noisy clustering (Figure 7.11) shows a different picture: East-West borders (one between Bern and Luzern, and one between Aargau and Zürich) clearly stand out, while the Southern dialect areas are less clearly distinguished. For instance, the Wallis, Fribourg and Berner Oberland regions are all drawn in very similar color hues. Further research will be required to explain these differences.

7.3.4. The influence of the inquiry technique on the inferred dialect landscapes

In the SADS dialect syntax survey, many questions consisted of choosing dialectal variants from a given list.[13] First, informants had to mark all variants that they deemed acceptable (multiple choice). Then, they had to select the single variant that they considered the most natural one (single choice).

13 Other questions involved free translation of Standard German sentences as well as completion of dialectal sentences. We do not use this material here.

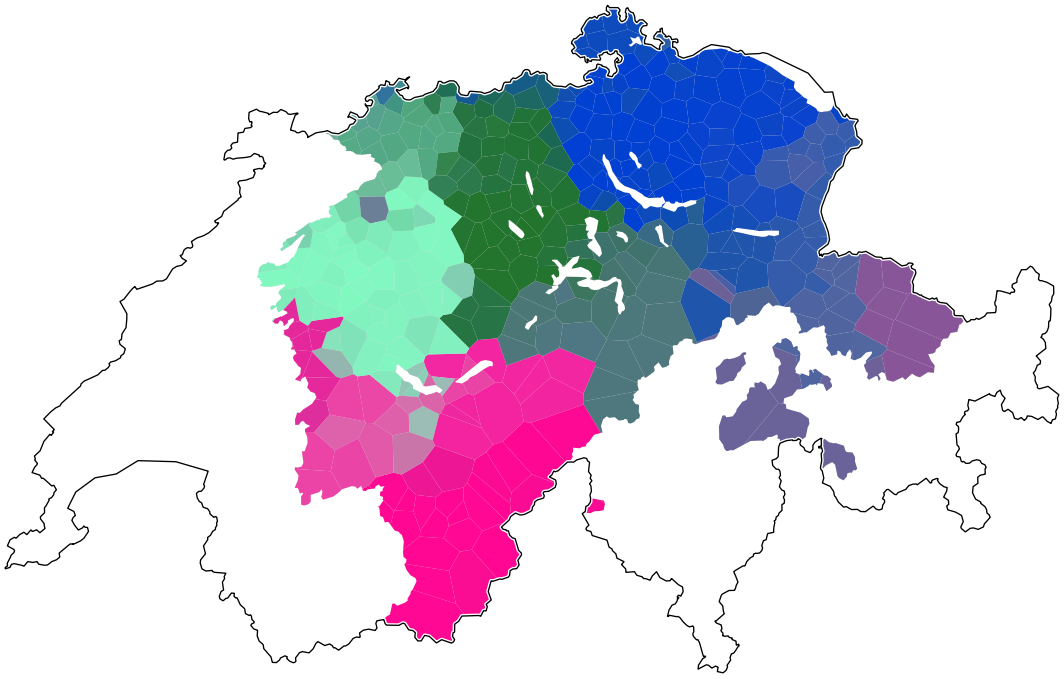


Figure 7.11.: Map created by noisy clustering and multidimensional scaling of the SADS data.

Our hypothesis is that the multiple choice material leads to "smoother" dialect landscapes: syntactic change usually is gradual, and the allowance for multiple acceptable answers better reflects this fact. On the other hand, we assume that single choice answers are more prone to outliers, as it might be difficult for informants to select a single "most natural" dialectal variant without being unduly influenced by Standard German grammar taught at school.

76 items of the SADS questionnaire contained data for both question types. We used these items

	Local incoherence	*Mantel test r*
Multiple choice	1.77	0.48
Single choice	2.37	0.45

Table 7.5.: Correlation figures for the two inquiry techniques. Lower local incoherence values are better, while higher Mantel test values are better. All Mantel test correlations are significant at $p = 0.01$.

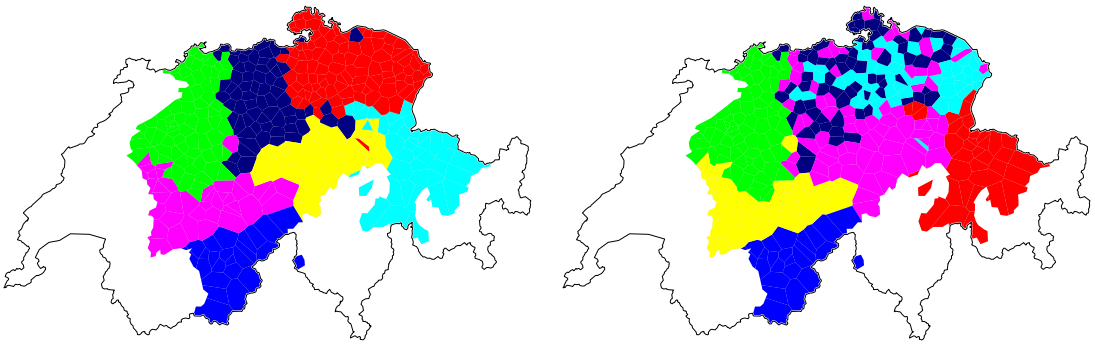


Figure 7.12.: Cluster maps inferred from multiple choice data (left) and single choice data (right).

to create two average distance matrices, one for each question type. Spatial autocorrelation figures are reported in Table 7.5. We find that the multiple choice inquiry technique correlates better with geographical distance than single choice, under both reported measures.

This observation is confirmed by the cluster maps shown in Figure 7.12 (7 clusters, Ward's Method). While the main dialectological borders are successfully induced on both maps, they show a crucial difference: the multiple choice map contains homogeneous dialect clusters, while the corresponding single choice map is much more scattered. It might be argued that the scattering is a result from choosing too high a number of clusters, or from choosing a particular clustering algorithm. This is not the case. Scattering is already visible from four clusters onwards with Ward's method. When using Weighted Average, scattering occurs in the Western regions from five clusters onwards. In summary, these results support the hypothesis formulated at the beginning of the section.

7.3.5. Localizing dialect texts

In the previous experiments, we used aggregated data collected in dialectological surveys like SDS or SADS. The work presented in this section aims to apply dialectometric analysis and visualization techniques to a different type of raw data, the comparable multidialectal Archimob

corpus.[14]

We derive a distance matrix based on the ratio between identical word pairs and cognate word pairs occurring in two texts. This ratio is shown to correlate well with geographic distance. The visualization of the resulting data allows us to recover certain characteristics of the Swiss German dialect landscape.

The Archimob corpus

The Archimob corpus used in our experiments is a corpus of transcribed speech, containing texts from multiple Swiss German dialects. The Archimob project was started in 1998 as an oral history project with the aim of gathering and archiving the people's memory of the Second World War period in Switzerland.[15] 555 surviving witnesses were interviewed in all Swiss language regions. The interviews of the German-speaking witnesses were conducted in their local dialect.

With the goal of obtaining spontaneous dialect data to complement ongoing work on dialect syntax (Bucheli and Glaser 2002; Friedli 2006; Steiner 2006), researchers at the University of Zurich selected a subset of the Swiss German Archimob interviews and transcribed them.[16] The selection process ensured that only interviews from non-mobile speakers (speakers that have not spent long periods of their life outside of their native town) were retained, and that the most important dialect areas of German-speaking Switzerland were represented. As a result, 16 interviews were selected for transcription, amounting to 26 hours of speech. All texts were anonymized. In order to ensure consistency, all texts were transcribed by the same person. The interviews were transcribed using the Dieth spelling system (Section 3.2). In our experiments, we discarded the interviewer's questions and only used the witnesses' turns. The whole corpus contains 183 000 words, with individual interviews ranging from 6 500 to 16 700 words. The

14 This section draws on Scherrer (2012), whose anonymous reviewers are thanked for many helpful remarks. We would also like to thank Prof. Elvira Glaser, Alexandra Bünzli, Anne Göhring and Agnes Kolmer (University of Zurich) for granting access to the Archimob corpus and giving detailed information about its constitution.

15 Archimob stands for "Archives de la mobilisation"; see www.archimob.ch.

16 The corpus is not yet publicly available, awaiting the completion of further annotation layers.

BE1142:	*de vatter ìsch lokomitiiffüerer gsìì / de ìsch dispensiert gsìì vom dienscht nattürlech / und / zwo schwöschtere / hani ghaa / wobii ei gsch / eini gschtoorben ìsch u di ander ìsch ìsch ime autersheim / u soo bini ufgwachse ir lenggass / mit em / pruefsleer / mit wiiterbiudig nächheer / (?)*
Translation:	the father has been a train driver / he has been dispensed from military service of course / and / two sisters / I have had / where one / one has died and the other is is in a home for the elderly / this is how I have grown up in the Lenggass / with a / apprenticeship / with further education afterwards / (?)

ZH1270:	*min vatter isch / eh eeh / schlosser hät er gleert / und und isch aber dän schoffööor woorde dur en verwante wo bim S. z züri / gschafft hät und dè hät gsait / chum tue doch umsattle bim S. vediensch mee / und dän hät dèè schoffööor gleert und das isch doozmaal ja na eener en sältene pruef gsii / dän hät dè das gleert und ich bin schtolz gsii das min / vatter en / pruef ghaa hät wo französischsch töönt hät oder schoffööor / ich han gfunde das seig en waansinige pruef*
Translation:	my father has / eh eeh / been a locksmith apprentice / and and has then become a driver through a relative who has worked at S. in Zurich and he said / come and switch jobs, at S. you earn more / and then he was a driver apprentice and this was rather a rare job at that time / so he learned that and I was proud that my / father / had a job which sounded French, you know, *chauffeur* / I found that this was an extraordinary job

Figure 7.13.: Excerpts of two informants' turns in the Archimob corpus.

place of residence of the witness was given in the corpus metadata. Excerpts of two interviews are shown in Figure 7.13.

The Archimob corpus can be qualified as comparable (McEnery and Xiao 2008): all texts deal with the same subject and the same time period (life in Switzerland at the outbreak of the Second World War), and they are collected in the same way, in the form of oral interviews guided by an interviewer.

It should be stressed that our data set is very small in comparison with other studies in the field: it contains 16 data points (texts) from 15 different locations. Moreover, some dialect areas are not represented in the sample (e.g., Graubünden in the South-East and Fribourg in the West).[17] Therefore, the goal of the present study cannot be to induce a precise dialect landscape

17 For an overview of the geographic distribution of the texts, see Figure 7.17.

of German-speaking Switzerland. Rather, we aim to find out if geographically close texts can be shown to be linguistically close, and if the classification of the texts reflects the major dialectal divisions of German-speaking Switzerland.

Computing the linguistic similarity of two comparable texts

The hypothesis put forward in this paper is that the linguistic similarity of two comparable texts can be approximated by the degree of similarity of the cognate word pairs occurring in the texts. Thus, computing the similarity of two texts amounts to the following two tasks:

1. Given two texts, extract the set of word pairs that are considered cognates. This corresponds to a task called **cognate identification** (Simard, Foster, and Isabelle 1992; Kondrak and Sherif 2006).

2. Within the set of cognate word pairs, determine the proportion of word pairs that are considered identical.

The underlying intuition is that identically pronounced cognate words account for evidence that the two dialects are closely related, whereas differently pronounced cognate words are evidence that the two dialects are distant. Word pairs that are not cognates are not relevant for our similarity measure.

Let us illustrate the idea with an example:

(1) a. *es schtòòt nìd*
b. *wil si nìd schtoot*

Intuitively, two cognate word pairs can be found in the texts (1a) and (1b): ⟨*schtòòt*, *schtoot*⟩ and ⟨*nìd*, *nìd*⟩.[18] The words *es, wil, si* do not have cognate equivalents in the other text. The two texts have a similarity of $\frac{1}{2}$ since one of the two cognate pairs consists of identical words.

18 The Archimob corpus uses narrow Dieth spelling with diacritics (see Section 3.2). We keep these conventions for our experiments and consider accented and unaccented characters as different. See also footnote 19.

In the example above, we have assumed informal meanings of **cognate word pair** and **identical word pair**. In the following section, we define these concepts more precisely in terms of Levenshtein distance thresholds.

Cognate word pairs and identical word pairs

Most recently proposed cognate identification algorithms are based on variants of Levenshtein distance (Heeringa, Kleiweg, et al. 2006; Kondrak and Sherif 2006). We have defined Levenshtein distance, together with a variant that is normalized by the alignment length, in Section 6.4.1. We use the normalized variant in our experiments here.[19]

A cognate identification algorithm based on normalized Levenshtein distance requires a threshold such that only those word pairs whose distance is below the threshold are considered **cognates**. We propose several thresholds between 0.20 and 0.40. In Scherrer (2012), we evaluate the quality of the induced cognate pairs according to these thresholds; Figure 7.14 sums up these findings.

In common understanding, an **identical word pair** is a pair of words whose Levenshtein distance is 0. In some of the following experiments, we adopt this assumption. However, we found it useful to relax this definition in order to avoid minor inconsistencies in the transcription and to neglect the smallest dialect differences. Therefore, we also carried out experiments where identical word pairs were defined as having a normalized Levenshtein distance of 0.10 or lower.

Normalization by text length

A major issue of using comparable corpora is the large variation in text length and vocabulary use. This has to be accounted for in our experiments. First, all counts refer to types of word pairs, not tokens: we argue that the frequency of a word in a given text depends too much on

19 Note that we treat all characters in the same way: replacing *o* by *k* yields the same cost as replacing it by *u* or by *ȯ*. This simple approach may not be the optimal solution when dealing with similar dialects. This issue will have to be addressed in future work.

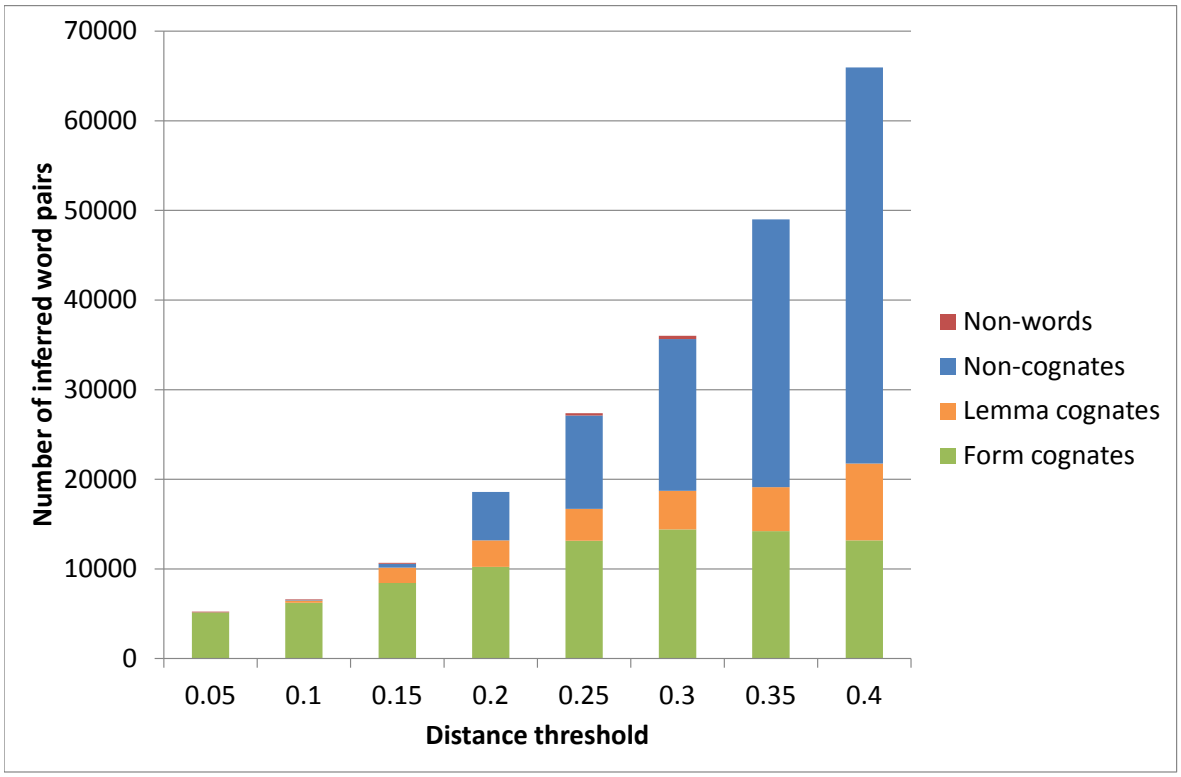


Figure 7.14.: Distribution of cognate pairs according to the cognate threshold. Form cognates are word pairs that represent the same inflected forms of the same lemma; lemma cognates are word pairs that represent different inflected forms of the same lemma.

the content of the text and is not truly representative of its dialect. Second, finding only few identical words does not necessarily mean that the two texts are dialectally distant, but may just be the consequence of one text being much shorter than the other. Hence, the proportion of identical words is normalized by the number of cognate words contained in the shorter of the two texts.

Evaluation and visualisation

Recall that we propose to measure the linguistic distance of two texts by the ratio of identical word pairs among the cognate word pairs. By computing the linguistic distance for all pairs of texts in our corpus, we obtain a two-dimensional distance matrix. Three dialectometrical analysis techniques are applied to this data. First, we measure how well the linguistic distances correlate with geographic distances. Second, we group the texts into clusters. Third, we use multidimensional scaling to plot the texts on a two-dimensional graph and visually compare it with the geographical locations of the texts.

We follow the spatial autocorrelation hypothesis (Section 7.2.2) to evaluate the different threshold combinations of our dialect similarity measure: the better a threshold combination correlates with geographic distance (i.e., places of residence of the interviewees), the better it is able to discriminate the dialects. As above, we compute local incoherence and Mantel test correlation coefficients.

Table 7.6 shows the results of both correlation measures for all experiments. Recall that for local incoherence, lower values are better. All Mantel test correlations are significant at $p = 0.001$ (999 simulations). At first, increasing the cognate pair threshold leads to more data, and in consequence, to better results. Above 0.35 however, the added data is essentially noise (i.e., non-cognate pairs), and the results drop again. According to local incoherence, the best threshold combination is $\langle 0.10, 0.35 \rangle$. In terms of Mantel test correlation, the $\langle 0.10, 0.25 \rangle$ threshold performs slightly better. Adopting an identical pair threshold of 0.00 results in slightly inferior correlations. Figure 7.15 displays the same results graphically.

The distance matrix can also be used as input to a clustering algorithm. Figure 7.16 shows

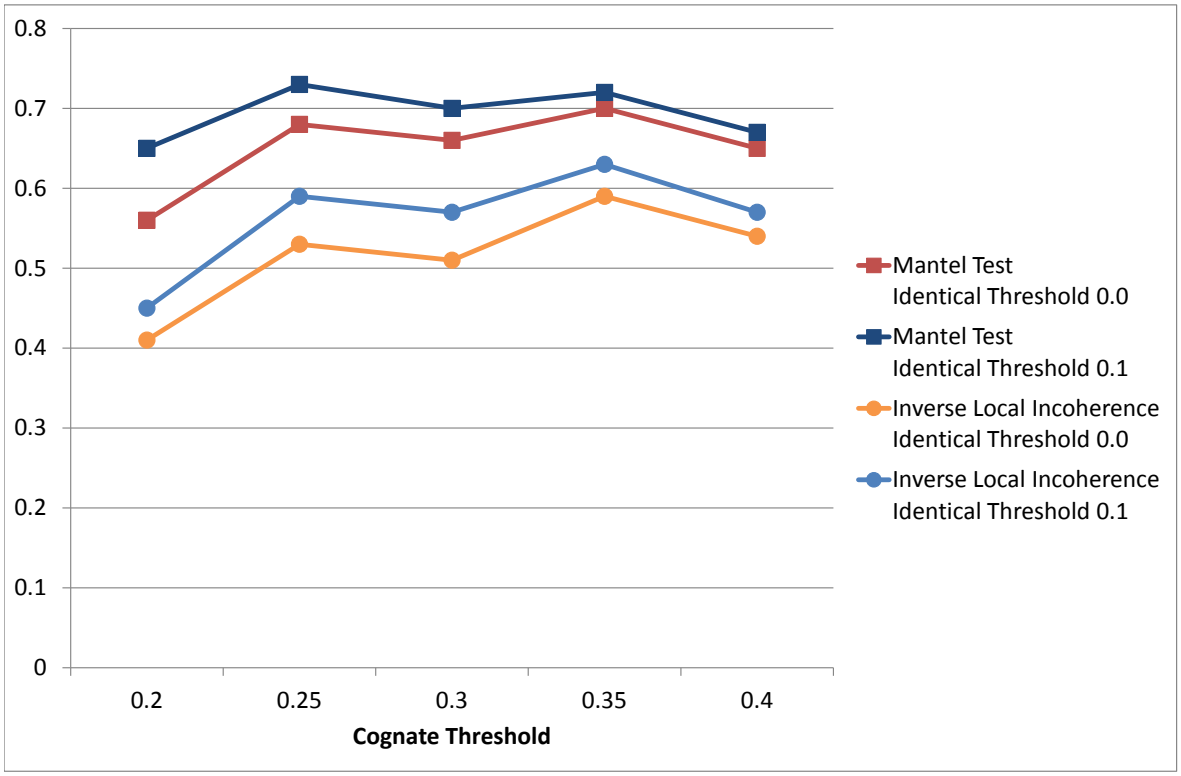


Figure 7.15.: Graphical representation of the correlation figures. Local incoherence values are inverted.

Distance thresholds		Local	Mantel
Identical	Cognate	incoherence	test r
0.00	0.20	0.59	0.56
	0.25	0.47	0.68
	0.30	0.49	0.66
	0.35	**0.41**	**0.70**
	0.40	0.46	0.65
0.10	0.20	0.55	0.65
	0.25	0.41	**0.73**
	0.30	0.43	0.70
	0.35	**0.37**	0.72
	0.40	0.43	0.67

Table 7.6.: Correlation values for the different experiments. The first and second columns define each experiment in terms of two Levenshtein distance thresholds.

a dendrogram obtained with noisy clustering.[20] Figure 7.17 localizes the data points on a geographical map. All clusters show a reliability score of 92% or above.

Clustering allows us to recover certain characteristics of the Swiss German dialect landscape. First, texts from the same canton (whose IDs contain the same two-letter abreviation) are grouped together with high reliability. Second, the dendrogram shows – albeit with lower reliability scores – a three-fold East-West stratification with blue regions in the West (BE), green regions in Central Switzerland (AG, LU) and yellow areas in the East (ZH, SZ, GL). Third, three dialects are clearly considered as outliers: the Northwestern dialect of Basel (BS1057), the Northeastern dialect of St. Gallen (SG1198), and most clearly, the Southwestern Wallis dialect (VS1212). Again, these observations are in line with common dialectological knowledge: BS1057 is the only Low Alemannic dialect in the corpus, and VS1212 the only Highest Alemannic dialect.

The close relation obtained among texts from the same canton may suggest that the distance measure is biased towards proper nouns. For example, two Zurich German texts are more likely

20 We used both group average and weighted average clustering algorithms with a total of 100 repetitions. The noise level was set at 0.25 times the standard deviation. These are the default settings of the Gabmap program Nerbonne, Colen, et al. 2011.

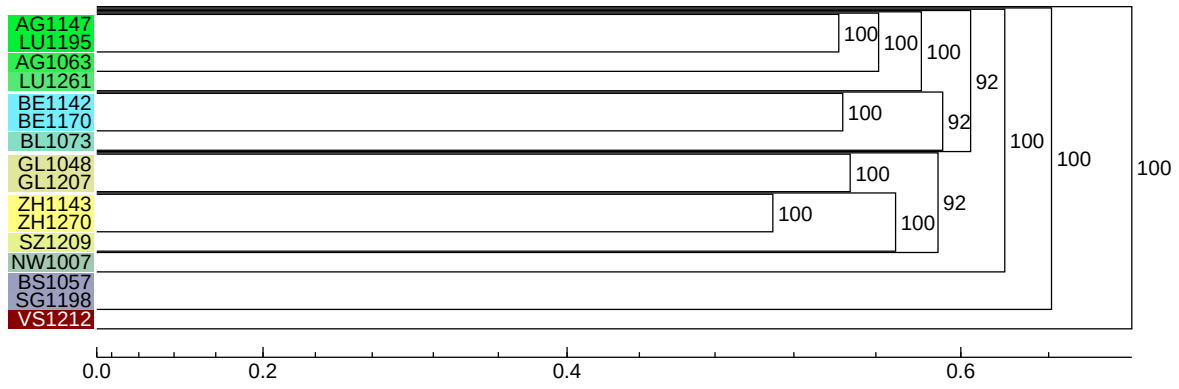


Figure 7.16.: Dendrogram obtained with a threshold setting of $\langle 0.10, 0.35 \rangle$. The scale at the bottom of the graphics represents the distance of the clusters, while the numbers on the vertical lines represent the reliability of the clusters (i.e. in how many of the 100 runs a cluster has been found).

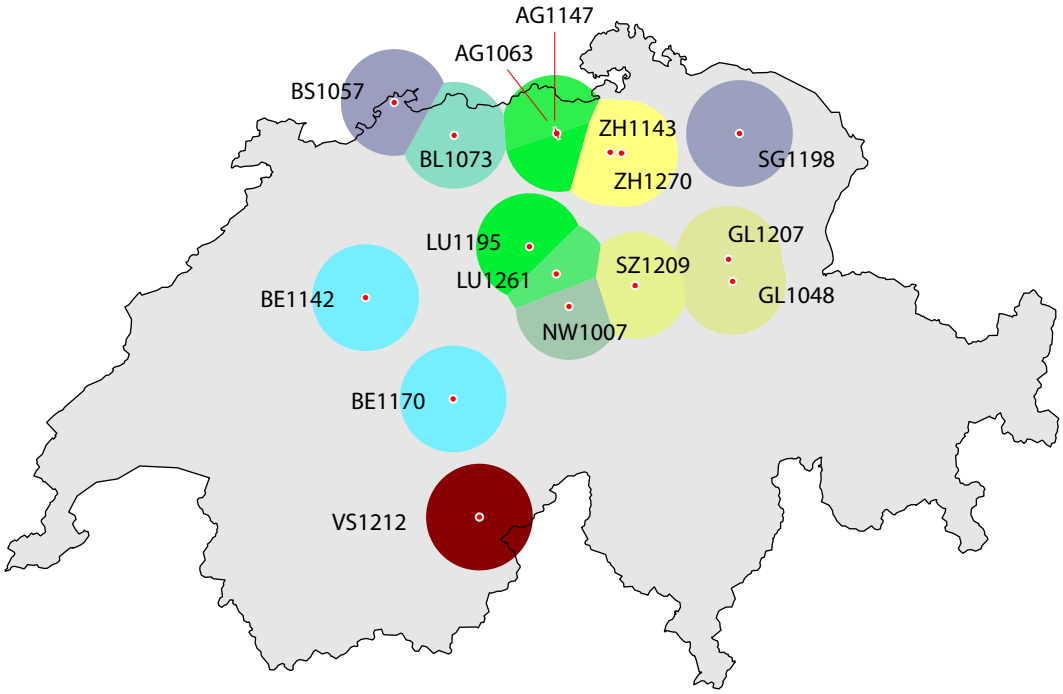


Figure 7.17.: Geographic localization of the Archimob texts, according to the place of residence of the interviewed persons. The colors represent the linguistic distance between texts; they correspond to the colors used in the dendrogram of Figure 7.16.

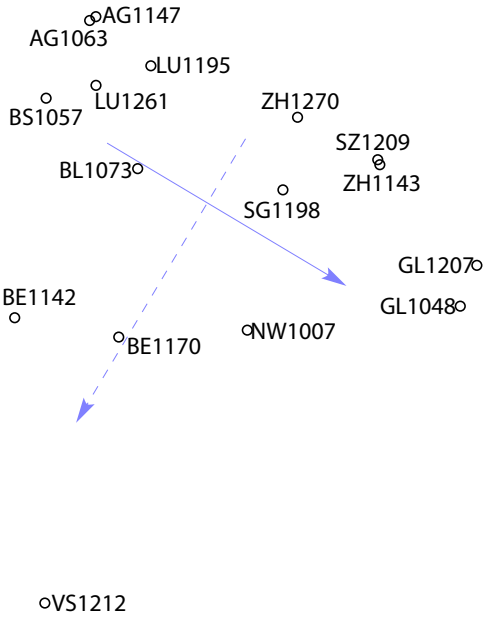


Figure 7.18.: Plot representing the first two dimensions of multi-dimensional scaling applied to the experiment with $\langle 0.10, 0.35 \rangle$ thresholds.

to use toponyms from the Zurich region than a Bernese German text. If there are many of these (likely identically pronounced) toponyms, the similarity value will increase. However, manual inspection of the relevant texts did not show such an effect. Region-specific toponyms are rare.

Multidimensional scaling allows to represent the Archimob texts in a two-dimensional space according to the mutual distance values. Our hypothesis is that the resulting space resembles the geographic localization of the texts. Figure 7.18 shows the resulting graph for one experiment. One indeed observes that the localization of data points according to linguistic distances closely corresponds to their geographic location illustrated in Figure 7.17: the major North-South divisions as well as some East-West divisions are recovered.

In conclusion, given the small amount of data and the noisy cognate identification method used in this experiment, the main characteristics of the Swiss German dialect landscape are represented in a surprisingly faithful way. The methods proposed here are rather simple and leave a lot of room for further improvements.

7.4. Conclusion

In this chapter, we have presented several dialectometrical analyses of Swiss German data. By using different data sets and different methods of analysis, we have given an overview of the state of the art in dialectometry, and of the Swiss German dialect landscape. The results presented here only mark the beginning of further dialectometrical work on Swiss German. Indeed, several issues should be addressed in future research.

First, the map selection criteria should be rendered more objective and the resulting data sets should be more balanced. In the case of the SDS data set, this means that maps that are not relevant for the text generation system should be included, and that the predominance of morphological data should be mitigated by digitizing additional lexical maps. In the SADS data set, some linguistic phenomena are weighted more than others, simply because they are represented by more questionnaire items. The data should thus be selected on the basis of linguistic phenomena rather than questionnaire items to avoid such an undesirable bias.

Second, the interpretation of the resulting maps is difficult. On the one hand, different methods or different data yielded results that were difficult to compare. On the other hand, the aggregated dialect data, at the heart of the dialectometric approach, creates a sort of "black box" whose inner workings are difficult to analyze.

Third, other methods and techniques remain to be explored. For instance, beam maps (Goebl 1982) and difference maps[21] have been widely used. More recently, factor analysis and principal component analysis have been proposed (e.g., Clopper and Paolillo 2006; Shackleton Jr. 2007; Grieve to appear) to address the "black box" problem and to gain more detailed insights into the internal structures of the resulting dialect landscapes.

Nevertheless, our experiments have already shown promising results and given interesting insights into the Swiss German dialect landscape. In part, they have confirmed traditional accounts, and in part they have suggested new dialect classifications. We therefore strongly believe that this line of research is worthwhile pursuing.

21 http://www.let.rug.nl/kleiweg/L04/Tutorial/t07.html.en

8. Conclusion

8.1. What we have achieved

In this thesis, we have studied the relation between Standard German and the various Swiss German dialects from a computational point of view and proposed a system that transforms Standard German sentences into dialectal ones. In contrast to other pieces of research, this thesis does not provide a new solution to an existing problem, but rather combines existing solutions to treat a new problem.

On the one hand, our thesis tackles a new problem in its field. Indeed, we have found few pieces of computational research about multi-dialectal systems, and even less work has dealt with the specificities of the Swiss German dialect landscape (see Chapter 2).

On the other hand, the methodological innovation chiefly lies in the original combination of a multitude of well-known techniques. For instance, we rely on key findings of previous dialectological work about Swiss German (Chapter 3, Appendix A); we have used geographical information systems to process the maps (Chapter 4); we have used finite-state toolkits to formalize and implement the transfer rules (Chapter 5); and we have drawn on existing resources for Standard German such as the TIGER corpus, the morphological analyzer *Morphisto*, or the *Derewo* word list (Chapter 5). Finally, we have applied geostatistical analysis and visualization techniques to interpolate the maps (Chapter 4) and to carry out dialectometrical investigations (Chapter 7).

In the introduction, we have distinguished two basic approaches to human language technology: the Natural Language Processing paradigm, and the Computational Linguistics paradigm. Ret-

rospectively, we consider most of our work as an instance of the former. The main goal of our research was to determine the most adequate computational tools and knowledge resources to accomplish a specific task, namely the generation of Swiss German text on the basis of Standard German text. To a large extent, the linguistic facts about Swiss German were taken for granted – after all, we were fortunate to draw on a lot of existing dialectological research. The exception to this general picture is Chapter 7: the dialectometric studies presented there are – we believe – of direct interest to dialectologists. They yield intuitive, yet theoretically founded visualizations of different aspects of dialect variation, taking advantage of the digitized SDS data. In this sense, dialectometry may be viewed as a subdomain of Computational Linguistics.

The work presented in this thesis has been time-consuming and challenging, yet always fascinating. At times, it resembled a tightrope walk: the rules had to be formulated very precisely in order to capture the slightest of dialectal nuances, but at the same time, they had to be tolerant enough towards the inevitable variation in spelling and diachronic evolution. While the results reported in Chapter 6 show that the optimal balance between these two tendencies is not quite reached, we believe that the general conception of our approach is sound and that minor improvements can be made within this framework.

Throughout this thesis, we took the stance that precisely annotated Standard German input is the key for high translation quality. However, obtaining such precise annotations turned out to be harder than imagined. All systems and resources used are error-prone to some extent: the compound word decomposition algorithm based on *Morphisto* may produce unwanted results; the part-of-speech tagger may show insufficient performance; the named entities in the TIGER corpus may not be correctly annotated; the *Fips* parser may not be able to analyze some sentences, etc. All these errors may have a negative impact on the dialect generation process. Paying less importance to the well-formedness of the Standard German input might alleviate this issue, but likely introduce more ambiguity elsewhere.

Another issue with this hypothesis is that morpho-syntactically annotated Standard German words are underspecified: some dialectal features cannot be inferred on the basis of the Standard German annotation alone. We did not sufficiently anticipate this fact, which in the end required us to integrate additional knowledge sources: lexical rules to override default phonetic transformations, word lists to distinguish phonemes that have fallen together in Standard German, and

word lists to determine the Swiss German noun inflection classes.

Despite these challenges, we have achieved some significant advances in the field of multi-dialectal processing of Swiss German. First, we have made available a significant subset of the SDS data in electronic format. The dialectometric experiments carried out with these data give a hint of the potential of such data sources. Second, we have shown how rule-based approaches to morphology can be extended in a cross-lingual and multi-dialectal way, by introducing the concept of georeferenced transfer rules. Again, we propose various interactive interfaces to this system on the web. Third, we have presented several experiments that show how such an original system may be evaluated thoroughly. We maintain that the evaluation of a multi-dialectal system must operate on two levels, a linguistic one and a geographical one. It must not only be ensured that the produced output forms are linguistically correct, but also that they are predicted to occur in the adequate geographic area. Evidently, such a two-fold evaluation is more demanding than the standard type of evaluation.

8.2. What we could have done but haven't

A doctoral dissertation consists of a set of methods and results, but it usually also features its share of missed opportunities and renouncements. This thesis being largely a solo effort, some interesting ideas could not be pursued because of work force and time constraints. In this section, we would like to review some of them.

Create a parallel corpus In order to evaluate the transfer from Standard German to Swiss German, a parallel corpus containing Standard German text and literally translated Swiss German text in multiple dialects is required. We have fulfilled this requirement rather minimally: the *Wikipedia* corpus presented in Section 6.2 contains 100 sentences of five dialects, together with their Standard German translation. A larger, cleaner corpus would have allowed us to obtain more reliable results, but also to pursue data-driven approaches.

Over the last few years, quite a lot of monolingual Swiss German text resources have been created and made available. Some of these resources are transcripts of recorded speech (the

Archimob corpus presented in Section 7.3.5, the corpus assembled by Christen (1998b)), while other resources are characteristic of a particular genre of electronic communication (the *chat* corpus of Siebenhaar (2005), the *sms4science* corpus of Stähli, Dürscheid, and Béguelin (2011)). Currently, none of these data sets contain Standard German translations or transcriptions, although this is planned for the *Archimob* and *sms4science* corpora. A multi-dialectal parallel corpus could have drawn on one of these dialect corpora, or on a different data source altogether. In any way, this would have been too time-consuming a task for a single person.

Use machine learning methods Most current work in Natural Language Processing uses machine learning methods to infer models for a specific task on the basis of large amounts of data; these data may or may not be specifically annotated for that purpose. Our work does not currently use machine learning techniques. There are two main reasons for this methodological choice. First, the dialectological atlases used as primary resources already contain linguistically interpreted data: the legend of each map can be interpreted as a rule with its conditions of applicability in a relatively explicit way. By using hand-written rules, we can fully take advantage of these data. Second, machine learning approaches require large training data – moreover, parallel training data for the translation task –, which are notoriously hard to find for small-scale language varieties like Swiss German, as we have mentioned in the previous paragraph. Even if a such a large, multi-dialectal parallel corpus had existed, we doubt that it would have permitted to recover the same type of georeferencing precision that could be achieved by using dialectological survey maps.

Create monolingual dialect models In the absence of parallel data, the availability of multi-dialectal but monolingual text collections could have been used profitably for tasks requiring monolingual data, for instance to learn morphology,[1] to create a part-of-speech tagger or even a full parser for Swiss German. Two approaches can be imagined. In the first approach, all data is thrown in the same bucket, and a single model that is robust enough to handle all dialects is learned. Likely, such a model would not be very precise, given the large dialectal variety. In the second approach, a specific model would be created for every dialect (area). It is unclear

1 For a recent overview of unsupervised learning techniques of morphology, see Hammarström and Borin (2011).

if enough data could be provided to ensure a satisfactory learning process, and if the resulting system could compete with a system based on dialectological maps. Using training data from various sources would have one particular advantage though: robustness to spelling variation would essentially come for free if trained on texts from various authors.

Evaluate models by native speakers Another way out of the parallel data bottleneck would consist of having the results of our system evaluated by native speakers of different dialects. For example, one could let dialect speakers guess the target dialect aimed at by the system, and measure the distance between the guesses and the effective target dialect. In this way, the quality of the translations could be measured in kilometers instead of BLEU points or LCSR scores, which would completely do justice to our multi-dialectal approach with georeferenced rules.

We have renounced this type of evaluation mainly for practical reasons. Besides the organizational challenge to find speakers of multiple Swiss German dialects in a city, university and faculty with relatively few native Swiss Germans, dialectality judgements are difficult to obtain from non-linguistically formed persons: Swiss German dialect speakers do not benefit from any grammatical education about their native dialect nor about other Swiss German dialects. Thus, we reckon that if several native speakers of the same dialect had been asked to judge the quality of our system's output, inter-annotator agreement would be quite low.

Our models have been evaluated either by the author, a native speaker of one Swiss German dialect, or on the basis of original Swiss German data. We believe that this setup is sufficient to obtain reasonably reliable data.

Use speech data Since Swiss German dialects naturally occur in spoken form, it would have been logical to use speech data in our investigations. This would have constituted an entirely different strand of research, which appeared less evident to pursue given our background and competences. For pioneering work such as the one presented here, we preferred sticking to written data, which are somewhat easier to handle from an Natural Language Processing point of view. Nevertheless, we find the idea of integrating speech into our model appealing, but this integration would have to be conducted in collaboration with experts in the field.

8.3. What remains to be done

The points listed in the previous section all require relatively heavy investments; each of them might qualify for a future research project. Besides these points, some minor modifications may be sufficient to improve the results of the translation system, or to acquire a better understanding of the Swiss German dialect landscape and its dynamics. The following paragraphs present some propositions.

More reliable preprocessing The current preprocessing pipeline for the morphosyntactic analysis of the Standard German source text relies on *Fips* or on the *TreeTagger-Morphisto* combination. Both approaches are error-prone in various ways. Such analysis errors propagate to the transfer system and are partly responsible for transfer errors. Alternative systems should thus be evaluated for all steps of the pipeline.

Translate from dialect into Standard German It has been suggested on several occasions that the translation from the dialects into Standard German could be more useful and more practical. We do not deny this claim, and would indeed have liked to investigate this translation direction. However, complexity issues have prevented us from doing so up to now. Streamlining the transfer rules to reduce their ambiguity could thus be a rewarding path for future research.

Extend the evaluation to larger and more dialectally diverse texts Most of the experiments reported in this thesis are based on the small *Wikipedia* corpus. This data set may not be large enough to truly assess the translation quality because its content is written by a very small number of authors who use idiosyncratic spelling and writing styles. Moreover, it only covers a small part of the Swiss German dialect landscape and so only partially does justice to the coverage of our translation system. In other words, some of the rules are never applied in the five dialects of the *Wikipedia* corpus, and their correctness and localization potential is not measured at all in the current experiments.

Investigate the impact of individual rules Currently, we only have a limited understanding of the importance and impact of the particular transfer rules. We have mainly chosen them according to our native speaker's intuition and to the availability of the respective geographical data. By quantifying the applications of each particular rule during the translation of a gold-standard corpus, two types of information may be obtained. First, the detection of unexpected applications of particular rules would allow us to spot and correct errors and imprecisions in the formulation of the rules. Second, the importance of a particular rule for a particular dialect can be measured by counting its correct applications in the corpus. In this way, one may determine the most characteristic rules for each dialect – in other words, its *shibboleths*. Such information is interesting from a dialectological point of view, but may also help improve our dialect identification methods.

Investigate better interpolation techniques The interpolation methods used in Section 4.6.2 aimed at extending the SDS data to places that had not been covered by the survey. Incidentally, interpolation also minimized the impact of perception and transcription errors. In our view, a successful interpolation technique for dialectal data should take into account two additional properties. First, the Swiss German dialects have undergone diachronic change since the SDS inquiry. Some aspects of this change may be simulated by judicious interpolation techniques, providing a "virtual update" of the SDS. Second, topographic information should be included in the interpolation algorithm. The current algorithm assumes that the influence of adjacent inquiry points is proportional to the distance as the crow flies. While this interpretation is reasonable in areas without major topographical obstacles (such as the Netherlands, or large parts of Germany), it is not adequate for the more or less secluded Alpine valleys of Switzerland.

8.4. Epilogue

It is somewhat unusual to propose a thesis in Computational Linguistics that relies so much un cartography and spatial analysis. Throughout our work, we were fascinated by the possibilities of this approach; we hope that the reader similarly appreciates the visual appeal of the resulting illustrations.

The integration of the above-mentioned techniques in a Natural Language Processing task constitutes a piece of truly interdisciplinary research. We hope that the ideas put forward in this thesis will motivate researchers from dialectology, computational linguistics and spatial statistic to continue together on this path.

This thesis draws on previous work from an unusually broad scientific field. We hope that our ideas stimulate this path of interdisciplinary research and brings together researchers from dialectology, computational linguistics and spatial statistics.

A. A comparative multi-dialect grammar of Swiss German

A.1. Introduction

The aim of this appendix is to provide a description of the most salient features of the different Swiss German dialects in a **comparative**, **multi-dialectal** manner. This grammar is designed to serve as a reference for the implementation of a computational system that derives Swiss German dialect data from Standard German data. Accordingly, it presents the relevant linguistic features in terms of **transformations** from Standard German to Swiss German. This synthesis draws on general descriptions of Swiss German (Haas 2000; Baur 1983; Lötscher 1983; Rash 1998), specific dialect grammars (A. Weber 1948; Suter 1976; Fischer 1989; Marti 1985), case studies (Penner 1995) and the two linguistic atlases presented in Chapter 3 (Hotzenköcherle, Schläpfer, et al. 1962-1997; Bucheli and Glaser 2002). It expands the general considerations on the Swiss German dialect landscape of Section 1.2.5.

The dialectal phenomena will be treated according to different linguistic levels, starting with phonetics (Section A.2). We found it difficult to neatly separate inflectional morphology from syntax. We have therefore chosen to combine these two levels in a section called morphosyntax (A.4). In contrast, derivational morphology is independent of syntax and in our case deals mostly with phonetic issues; we have placed this section after the phonetics section (A.3). Lexical variation will be discussed briefly at the end of this chapter (A.5), but semantic and pragmatic issues will not be addressed.

SDS	*ISO*	*German Name*
AG	AG	Aargau
AP	AR	Appenzell Ausserrhoden
	AI	Appenzell Innerrhoden
BA	BL	Baselland
	BS	Basel-Stadt
BE	BE	Bern
FR	FR	Freiburg
GL	GL	Glarus
GR	GR	Graubünden
LU	LU	Luzern
UW	NW	Nidwalden
	OW	Obwalden
SH	SH	Schaffhausen
SZ	SZ	Schwyz
SO	SO	Solothurn
SG	SG	St. Gallen
TG	TG	Thurgau
UR	UR	Uri
WS	VS	Wallis
ZG	ZG	Zug
ZH	ZH	Zürich

Table A.1.: Canton abbreviations according to the SDS and ISO conventions.

A.1.1. Geographical references

In the following sections, we need a principled system to refer to the regions in which the presented dialect differences occur. The cantons, the main polities of Switzerland, have turned out to be convenient and intuitive entities for dialectological description. Nevertheless, the surfaces of the cantons vary so much that many linguistic phenomena concern either parts of large cantons, or sets of smaller cantons.

We have chosen the following guidelines:

- We always use the German designation of the cantons, even if English names exist, and even if the French name is more common: 'Zürich' instead of 'Zurich', 'Luzern' instead of 'Lucerne', 'Freiburg' instead of 'Fribourg'.

- We use the two-letter abbreviations of cantons as used in the SDS. These differ slightly from the abbreviations used on license plates and in other administrative uses (norm ISO-3166-2:CH), as shown in Table A.1. For reference, a map of the cantons is displayed in Figure A.1.

- Cardinal points (and their German abbreviations N, S, O, W) are employed to describe

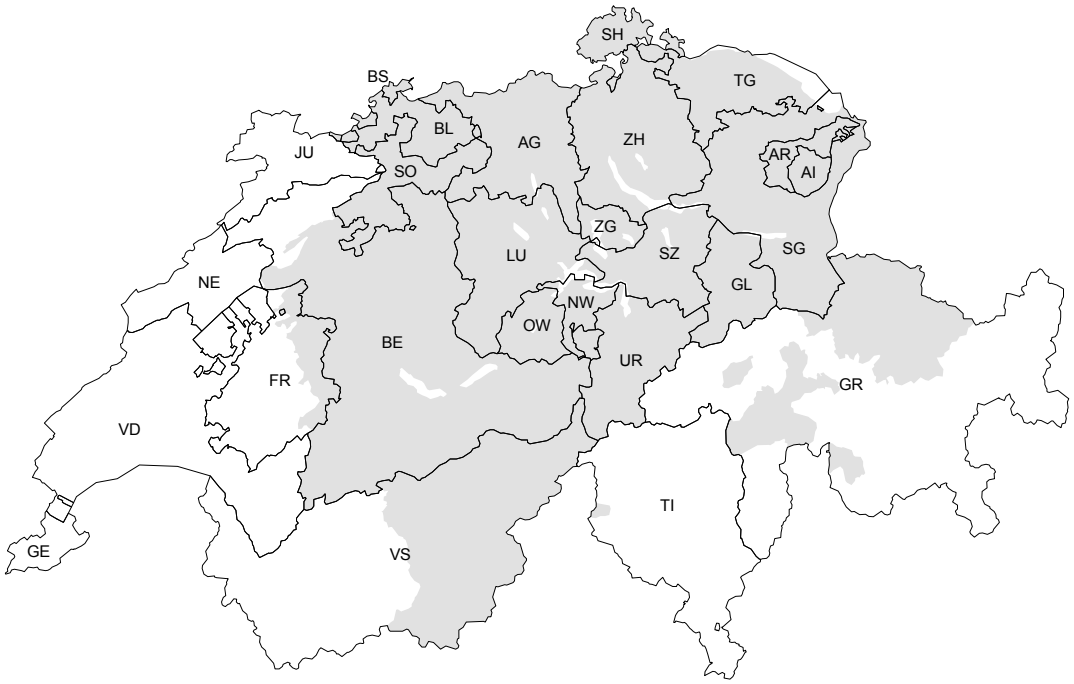


Figure A.1.: Map of the cantons of Switzerland. Labels follow the ISO convention. The grey surface represents the traditionally German-speaking area.

large-scale phenomena.

- To specify subregions inside a canton, we use the name of that region, prefixed by the canton abbreviation to help the reader unfamiliar with the geography: *UW-Engelberg, WS-Lötschental, BE-Oberland, GR-Walser*. The common term *Oberland* is abbreviated as O: BEO, ZHO, SGO.

A.2. Phonetics and phonology

This section is based on accounts in Haas (2000, pp. 87-92), Baur (1983, pp. 22-30), Lötscher (1983, pp. 80-89, 137-180), Rash (1998, pp. 138-140), and Siebenhaar and Voegeli (1997). We will also refer to the first two volumes of the SDS (Hotzenköcherle, Schläpfer, et al. 1962-1997).

Throughout our work, we will not distinguish explicitly between phonetics and phonology. This distinction is difficult to maintain in the context of dialect geography. A specific sound difference may have phonemic value in one dialect, but not in another. Hence, the same sound law would be classified under phonetics in one dialect, and under phonology in another. In a dialect continuum, such a separation would complicate the presentation without obvious benefit to the linguistic description.

We first present the vowel system and then turn to the consonants. Vowel correspondences show rather irregular variation patterns, while consonant correspondences are more uniform.

A.2.1. Vowels

The most important vocalic variation between Standard German and Swiss German is between diphthongs and monophthongs. We will present these variation patterns first, although they are difficult to describe because of massive lexical and interdialectal variation. Roughly speaking, Swiss German dialects preserve the vocalic patterns of Middle High German (Paul 2007, pp. 74-86; Weddige 2007, pp. 31-41), while Standard German has undergone some innovations.

Standard German diphthongs

Standard German has three types of diphthongs: *ei* [aɪ], *au* [aʊ], and *eu/äu* [ɔʏ]. The outcome of these diphthongs in Swiss German is not uniform and depends on the historic origin of the sounds. There are two major classes of diphthong transformations.

Class 1 The Standard German diphthongs of this class are an innovation of the 13th century. They reflect the Middle High German monophthongs *î*, *û* and *iu* (pronounced [y]). The outcome of these sounds in Swiss German depends largely on the phonetic context. Two major contexts are worth noting: (a) the Standard German diphthong is followed by a consonant or by the

suffix *-er*; (b) the Standard German diphthong is followed by another consonant *(hiatus)* or occurs at the end of the word or morpheme.

Class 1a In the first context, most Swiss German dialects have conserved the older long monophthongs. Only in some alpine dialects, these words show diphthongs, albeit with different qualities than in Standard German (SDS I/104-107):[1]

(1) a. Zeit [aɪ] – Ziit, Zeit [eɪ] (NW, GR-Schanfigg) 'time'
b. Haus – Huus, Huis (OW, NW, WS-Lötschental), Hous (GR-Schanfigg) 'house'
c. Häuser – Hüüser, Heiser (NW), Höüser (GR-Schanfigg) 'houses'

Interestingly, the GR-Schanfigg dialect developed diphthongs as a result of substratum influence from the Romansh dialect spoken in that region.

In the context of a following *er*-suffix, the *schwa* of the suffix is dropped. Otherwise, the same transformations apply:

(2) a. Feier – Fiir, Feir 'celebration'
b. Bauer – Puur, Puir, Pour 'farmer'
c. Feuer – Füür, Feir, Föür 'fire'

In Bern dialect, the monophthongs are short if they occur in open syllables (SDS II/71-72):

(3) a. schneiden – schniide, schnide (BE) 'to cut'
b. Taube – Tuube, Tube (BE) 'dove'

Class 1b If the Standard German diphthong occurs at the end of the stem of a word, the plateau dialects use some sort of diphthong, while alpine (i.e. Southern) dialects keep the older monophthongs (SDS I/148-157):

1 In the following examples, the word to the left of the dash will always refer to its Standard German spelling. The word(s) to the right of the dash exemplify different Swiss German dialectal variants. In the absence of geographic information in parentheses, the given variants are those that are valid in the majority of plateau dialects. In order to distinguish small-area variants from default plateau variants, we present the default variant first, followed by the small-area variants with their respective geographical extent specified in parentheses.

(4) a. frei – frei (W), fräi (Z), frai (O), frii (S) 'free'
b. Sau – Sou, Sau (O), Suu (S) 'sow'
c. neu – nöi, noi (O), nüü (S) 'new'

When the diphthong is followed by another vowel, a subset of these alpine dialects introduce a *w* to bridge that hiatus:

(5) a. bauen – buue (NW, SZ, GL, GR), buuwe (BEO, WS, OW, UR) 'to build'
b. neuer – nüüer (NW, SZ, GL, GR), nüüwer (BEO, WS, OW, UR) 'newer'

Class 2 These Standard German diphthongs reflect the Middle High German diphthongs *ei, ou,* and *öu*. Again, the resulting Swiss German forms depend on the phonetic context (SDS I/104-132).

If followed by another vowel *(hiatus)*, all Swiss German dialects keep the diphthong. Its quality may differ though:

(6) a. Eier – Eier, Äier, Aier 'eggs'
b. Frauen – Froue, Fraue 'women'
c. heuen – höie, hoie, heie, haie 'to hay'

As in Class 1, semi-vowels can be introduced in alpine dialects:

(7) Frauen – Frouwe, Frauwe 'women'

In all other contexts, the diphthong is reduced to a long monophthong in BE-Oberland and FR dialects, while other dialects keep the diphthong:[2]

(8) a. Geiss – Geiss, Gäiss, Gaiss, Geess (BEO, FR) 'goat'
b. Auge – Oug, Aug, Oog (BEO, FR) 'eye'
c. Baum – Boum, Baum, Boom (BEO, FR) 'tree'
d. räuchern – röichere, roichere, rööchere (BEO, FR) 'to cure, to fumigate'

2 However, the spatially limited monophthongization is receding in favor of the large-scale diphthong variant (Christen 1997, p. 355).

e. Bäume – Böim, Boim, Bööm (BEO, FR) 'trees'

Monophthongization tendencies are also found in Northeastern dialects, but they are usually restricted to contexts of a following nasal consonant. In these dialects, the resulting monophthong is short:

(9) a. heim – hamm (TG, SG), hää (AP) 'home' (Adv)
b. Baum – Bòmm (SH, TG, SG, AP) 'tree'
c. Bäume – Bömm (SH, TG, SG, AP) 'trees'

In the same Northeastern dialects, the monophthong variants of *ei* extend to non-nasal contexts (albeit with a lot of lexical variation). In these cases, the monophthongs are long, and they show characteristic nasalisation in AP-Innerrhoden dialect:

(10) a. Geiss – Gaass (TG, SG), Gääss (AR), Gẽẽss (AI) 'goat'
b. Stein – Staa (TG, SG), Schtää (AR), Schtẽẽ (AI) 'stone'

Conclusion The phonetic configuration of a Standard German word form only partially determines its class affiliation. As a consequence, phonetically similar Standard German words can have very different outcomes in Swiss German:

(11) a. weiss – wiiss, weiss (NW) 'white'
b. weiss – weiss, waiss, waass (SG, TG), wääss (AP), weess (BEO) 'knows'

(12) a. Zeit – Ziit, Zeit (NW) 'time'
b. breit – breit, brait, braat (SG, TG), bräät (AP) 'broad'

Even when looking at a single dialect, a given Standard German diphthong may lead to differing pronunciations according to its class:

(13) a. frei (Class 1) – frei (ZH)
b. Stein (Class 2) – Schtäi (ZH)

(14) a. Sau (Class 1) – Sou (ZH)
b. Baum (Class 2) – Baum (ZH)

Class 1 will be used as a default class for all Standard German diphthongs. Words belonging to class 2 will have to be marked specifically. The different subclasses can be determined by the phonetic context.

Standard German long monophthongs

In the previous section, we have presented cases of Standard German diphthongs becoming monophthongs in Swiss German. In other contexts, the evolution has gone in the opposite direction: Middle High German diphthongs have evolved into long monophthongs in Standard German, but have been conserved in Swiss German:

(15) a. gut – guet 'good'
b. müde – müed 'tired'

The qualities of the first element of the diphthong and the quality of the *schwa* may change in different dialects. These changes will be discussed later. The diphthongization tendency is rather universal across Swiss German dialects (SDS I/140-I/146). We neglect the (obsolete) exceptions in front of nasal consonants in SH and parts of TG.

The problem is more complex in the case of the Standard German long monophthong *ie*.[3] Some occurrences of *ie* reflect Middle High German *io/ia*; in parallel to the *u* and *ü* cases of the preceding paragraph, they show a diphthong in all Swiss German dialects:

(16) a. lieb [iː] – lieb [iə] 'dear'
b. nie [iː] – nie [iə] 'never'

Other occurrences of Standard German *ie* result from Middle High German *iu*. They show various types of diphthongs and monophthongs in Swiss German dialects (SDS I/134-138):

(17) a. tief – tief (W), töif (ZH), tüüf (O) 'deep'
b. Fliege – Fliege (W), Flöige (ZH), Flüüge (O) 'fly'

3 One factor of complexity is due to spelling. The Standard German grapheme *ie* is used for the monophthong pronounced [iː]. In Swiss German spelling, it is customary to use the grapheme *ie* for the diphthong [iə].

The above correspondences are valid for the Standard German long vowels *i, u, ü*. The long vowels *e, o, ö* are traditionally diphthongized only in the cantons of Schwyz and Zug (SDS I/95-103):

(18) a. Schnee – Schnei (SZ, ZG) 'snow'
b. Lehrer – Leirer (SZ, ZG) 'teacher'
c. Brot – Brout (SZ, ZG) 'bread'
d. Löhne – Löun (SZ, ZG) 'wages'

Closing

The Middle High German *u* in front of nasal consonants has evolved into *o* in Standard German, but has remained *u* in some Swiss German dialects (SDS I/45-46):

(19) a. Sommer – Sommer, Summer (BA, ZH, SH) 'summer'
b. geschwommen – gschwomme, gschwumme (BA, ZH, SH) 'swum'
c. Wolle – Wole, Wule (BA, ZH, SH) 'wool'

In word-final contexts (mostly before a dropped final *n*), Glarus dialect has *u* instead of *o* (SDS I/68-72):

(20) a. Lohn – Loo, Luu (GL) 'salary'
b. gehen – goo, guu (GL) 'to go'

Opening

The opposite evolution can also be seen in some Swiss German dialects. *u* is transformed to *o*, *ü* to *ö*, and *i* to *e* in the Western half of the Plateau, as well as in the Appenzell region (SDS I/48-52):

(21) a. Hund – Hund, Hond (BE, SO, AG, LU, AP) 'dog'
b. Hunde – Hünd, Hönd (AP) 'dogs'
c. Schlitten – Schlitte, Schlette (BE, SO, AG, LU, AP) 'sled'

In Bern dialect tradition, it is customary to spell the words with the closed vowel: *Hund, Hünd, Schlitte.*

Unrounding

In the alpine dialects of Wallis, Uri, Unterwalden, parts of the Berner Oberland, and the GR-Walser dialects, the following shifts have taken place: *u* → *ü*, *ü* → *i*, and *ö* → *e*. The second shift is also characteristic of Basel dialect:

(22)	a.	Fuchs – Fuchs, Függs (WS, UW, UR, GR)	'fox'
	b.	Füchse – Füchs, Figgs (WS, UW, UR, GR, BA)	'foxes'
	c.	Löcher – Löcher, Lecher (WS, UW, UR, GR)	'holes'

This unrounding is applied after the abovementioned monophthongization and diphthongization changes, as shown in the following examples:

(23)	a.	Haus – Huus > Hüüs	'house'
	b.	Häuser – Hüüser > Hiiser	'houses'
	c.	Blume – Blueme > Blüeme	'flower'
	d.	grün – grüe(n) > grie(n)	'green'

These changes are described in maps I/47-60, I/102-107, I/120-132, I/142-146 of the SDS.

Shortening

Vowels in open syllables were short in Middle High German, but have been lengthened in Modern Standard German. Northwestern dialects have long vowels as in Standard German, but the other dialects have kept the original short vowels (SDS II/1-44).

As we focus on the transformation from present-day Standard German to present-day Swiss German, we refer to this process as shortening, even if historically, there has never been a

transition from long to short vowels. Note also that the Standard German spelling does not always explicitly indicate vowel length:

(24)
a. zählen [ɛː] – zele 'to count'
b. Schnabel [aː] – Schnabel, Schnaabel (BA, SO) 'beak'
c. übel [yː] – übel, üübel (BA, SO) 'evil'
d. Hose [oː] – Hose, Hoose (BA, SO) 'trousers'
e. Regen [eː] – Räge, Rääge (BA, SO, AG) 'rain'

As we focus on written dialect representations, the transformation to be applied looks backwards: words whose pronunciation is the same have to be written differently (*Hose – Hoose* with a long vowel in both cases), but words with different pronunciations have to be spelled the same (*Hose – Hose* with long vowel in Standard German, but short vowel in Swiss German).

Furthermore, words with long vowels that enter a compound noun as the first element tend to have shortened vowels (SDS II/79):

(25)
a. grooss + Vatter > Grossvatter 'grandfather'
b. Fiir + Oobig > Firoobig 'closing time'

Lengthening

In front of consonant clusters, Standard German vowels are short, but are lengthened in most Swiss German dialects (depending on the context, these vowels had been short or long in Middle High German). The geographic distribution of long and short vowels is less clear-cut than in the previous case. It seems restricted to alpine areas like Uri, Unterwalden, St. Galler Oberland, and others (SDS II/57-66):

(26)
a. gebracht – praacht, pracht (UR, UW, SGO) 'brought'
b. Dorf – Doorf, Dorf (UR, UW, SGO) 'village'
c. Berg – Bäärg, Bärg (UR, UW, SGO) 'mountain'
d. Salz – Saalz, Salz (UR, UW, SGO) 'salt'

In this case again, spelling issues have to be discussed: while the Dieth spelling requires long vowels to be spelled as double vowels (as we have done in the examples above), most dialect writers would rather use simple vowels (*Dorf, Bärg, Salz*) as in Standard German spelling. It seems thus that this phenomenon is less important when it comes to generating readable Swiss German: any variant, with simple or double vowels, can be considered acceptable.

Schwa syncope and apocope

Two issues have to be discussed in relation with the reduced vowel *schwa* in Swiss German. This section deals with the presence or absence of the *schwa*. The following section presents regional variation of the quality of the *schwa*.

All Swiss German dialects drop word-final *schwa* (apocope):

(27) a. Rede – Reed 'speech'
b. Straße – Straass 'road'

However, *schwa* is maintained in final syllables when followed by a consonant (independently of the fact that the final consonant may have dropped in Swiss German):

(28) a. sitzen – sitze 'to sit'
b. Wasser – Wasser 'water'

Syncope is also frequent, especially for the unstressed prefixes *ge-* (for the use of *ge-* in past participles, see page 357) and *be-* (see also page 318):

(29) a. beschlossen – bschlosse 'decided'
b. gesungen – gsunge 'sung'
c. Gesellschaft – Gsellschaft 'society'

Schwa quality

In the plateau Swiss German dialects, the *schwa* is slightly more open than the Standard German *schwa*. To mark this difference in spelling, many dialect writers use *ä* instead of *e*:

(30) sitzen – sitzä

We believe that this slight difference of sound quality should not be reflected in spelling, as plateau Swiss German *schwa* remains a reduced vowel. Following Dieth (1986, p. 33), we continue to write *e* in these cases:

> In unbetonter Silbe sollst du den Vokal nur dann anders als *e* (in dieser Stellung ein farbloser Laut) schreiben, wenn er ausgesprochene Farbe (*i a ä o u*) zeigt.

These "colored" variants appear in alpine dialects, with the quality [æ] from Wallis to Central Switzerland, and with the quality [ɐ] typical for Graubünden dialect (SDS III/1):

(31) sitzen – sitzä (UR, UW, WS), sitza (GR)

These considerations are also valid for the second element of diphthongs (SDS I/140-141):

(32) a. lieb – lieb, liäb (UR, UW, WS), liab (GR)
b. gut – guet, guät, guat
c. müde – müed, müäd, müad

Semi-vowel insertion in hiatus

Some Standard German verb stems end in a vowel, which causes a hiatus with the verb ending *-en*. This hiatus is sometimes bridged by the insertion of *h* (which is seldom pronounced however). Some Swiss German dialects insert a *j* instead, while others keep the two subsequent vowels (SDS II/161-162):

(33) a. mähen – määje, mäe 'to mow'

b. säen – sääje, säe ‘to sow’
c. brühen – brüeje, brüe ‘to scald’

Note that the long *ü* is diphthongized if followed by *j*. The insertion of *w* or *b* to bridge the gap between *u* and *e* (SDS II/157-158) is marginal and will not be discussed here.

General vowel quality

All vowels tend to be more open in Western dialects than in Eastern dialects (SDS I/15-42):

(34) a. Bett – Bètt (W), Bett (O) ‘bed’
b. sterben – schtäärbe (W), schtèèrbe (O) ‘to die’
c. Brot – Bròòt (W), Broot (O) ‘bread’
d. hübsch – hübsch (W), hübsch (O) ‘beautiful’
e. richtig – rìchtig (W), richtig (O) ‘right’

As explained above, we have chosen not to use diacritics to mark vowel quality. Therefore, this important interdialectal difference will not be expressed by our system.

The system of *e*-like vowels is particularly complicated. Recall that the Standard German system has (besides *schwa*, which we do not discuss here) two degrees of openness, of which only one accepts short vowels. Swiss German dialects have two or three degrees of openness, and all of them exist in short and long variants. We compare the vowel systems of three Swiss German dialects with Standard German in the table below (SDS I/15-39, I/73-81, I/95-98, II/25-31):

(35)

	Closed		*Mid-Open*		*Open*	
	Long	*Short*	*Long*	*Short*	*Long*	*Short*
	[eː]	[e]	[ɛː]	[ɛ]	[æː]	[æ]
Standard	✓		✓	✓		
SG	✓	✓	✓	✓		
ZH	✓	✓	✓	✓	✓	✓
BE			✓	✓	✓	✓

The spelling of these different variants depends on the dialect. Dieth proposes to use *e* for the most closed vowel of the dialect, *ä* for the most open vowel, and *è* for intermediate vowels if they exist in that dialect. This collapses some interdialectal differences. In broad spelling variants like the one that we use, the *è*-forms are either spelled *e* or *ä*. As a result, there is no need for transformation rules in most cases.

Umlaut as a morpho-phonological process

The term *umlaut* commonly refers to the German graphemes containing a diaeresis: *ä, ö, ü*. But *umlaut* also refers to a morpho-phonological process that yields the phonemes associated with these graphemes. Historically, the presence of a *i* in a word caused the preceding vowel to be fronted: [a] yielded [ɛ], [o] yielded [ø], and [u] yielded [y]. These correspondences were later systematized in spelling with the grapheme pairs *a – ä, o – ö, u – ü*. The *i* appeared in diminutive endings, noun plural endings, and other inflection markers. While the *i* is still visible in some suffixes today (e.g. the dialectal diminutive ending *-li*), it has been reduced to *schwa* in other contexts. However, its action on the stem vowel has persisted.

In present-day Standard German, *umlaut* consists of the transformations *a* → *ä, o* → *ö, u* → *ü, au* → *äu*. It is used in diminutive derivation (36a), for noun plural forms (36b), adjective gradation (36c), and irregular verb inflection (36d). *Umlaut* also appears lexemically, without any obvious trigger of the transformation (36e):

(36)	a.	Hund / Hündchen	‘dog / little dog’
	b.	Haus / Häuser	‘house / houses’
	c.	gross / grösser	‘big / bigger’
	d.	ich falle / du fällst	‘I fall / you fall’
	e.	grün	‘green’

This process and its outcomes are relatively similar in Swiss German dialects. Use of *umlaut* to mark noun plural and adjective gradation is more frequent in Swiss German than in Standard German, as will be discussed in the morphology section. From a phonetic point of view, some dialectal particularities may be pointed out.

- Due to monophthongization, the vowel pair *au – äu* often corresponds to *uu – üü* in Swiss German.

- Standard German *a* yields Swiss German *o* in some dialects (SDS I/61-72). When such words undergo *umlaut*, the resulting forms logically contain *ö* instead of *ä*. However, Zürich dialect uses non-umlauted form *a* together with the umlauted form *ö*:

  (37) Rat / Räte – Root / Rööt (NW, NO), Raat / Räät (SW), Raat / Rööt (ZH)
  'counsellor / counsellors'

- Many dialects distinguish the mid-open *è* [ɛ] from the open *ä* [æ]. Both sounds can figure as umlauted versions of *a*, depending on the dialect and the word:

  (38) a. Hand / Hände – Hand / Händ 'hand / hands'
  b. zart / zarter – zaart / zèèrter 'delicate / more delicate'

  Due to the intermediate nature of the sound *è*, its non-diacriticized spelling varies between *e* and *ä*. By collapsing both pronunciations into the grapheme *ä*, we can avoid the need for word lists to generate correct umlauted forms.

A.2.2. Consonants

The pronunciation of *ch* and *s*

In Standard German, the grapheme *ch* is pronounced [ç] after the vowels *e, i, ä, ö, ü*, but [x] after the vowels *a, o, u*. It is well known that this distinction is not made in Swiss German. Rather, there is a one-to-one correspondence between the grapheme *ch* and the pronunciation [x].

The situation for the grapheme *s* is similar: it is voiced [z] in some Standard German contexts and voiceless [s] in others. Swiss German speakers always pronounce it voiceless.

As these facts does not influence the spelling, no transformation rules are required.

The outcomes of Standard German *k*

The Standard German graphemes *k* and *ck*[4] are pronounced as an aspirated plosive [kʰ]. The most similar sound in the majority of Swiss German dialects is [kx], which the Dieth spelling associates with the graphemes *k* and *ck*. However, this is not to say that all Standard German occurrences of *k, ck* [kʰ] correspond to *k, ck* [kx] in all Swiss dialects and in all contexts; the situation is more complicated.

Let us first examine the word-medial and word-final contexts. Many words containing *ck* [kʰ] in Standard German have *ck* [kx] in most Swiss German dialects, but *gg* [k] in Northeastern Switzerland. This law also holds for most loanwords containing Standard German *k* (SDS II/95):

(39) a. wecken – wecke, wegge (TG, SG) 'to awake'
b. Sack – Sack, Sagg (TG, SG) 'bag'
c. Doktor – Tokter, Toggter (TG, SG) 'doctor'

In words that already had *gg* in Middle High German, *gg* remains in all Swiss dialects while Standard German has innovated to *ck* (SDS IV/41):

(40) Rücken – Rugge (O), Rügge (W) 'back'

Most words containing *k* in word-initial position have *ch* [x] in most dialects. Exceptions occur in BS and some GR, SG and AP dialects with *kh* [kʰ] (SDS II/94). Moreover, if initial *k* is followed by *l, n, r*, it becomes *g* in Basel dialect (SDS II/165):

(41) a. Kalb – Chalb, Khalb (BS, GR, SG, AR) 'calf'
b. klein – chlii, glai (BS), khlii (GR, SG, AR) 'small'
c. Krieg – Chrieg, Grieg (BS), Khriag (GR, SG, AR) 'war'

Another class of words has *k* [kx] in all dialects except in Basel and Graubünden. Among these words are also recent loanwords and re-borrowings from Standard German. The *k* → *ch* law is

4 Standard German spelling uses *ck* instead of *kk*.

thus no longer productive.[5]

(42) Kamel – Kameel, Khameel (BS, GR) 'camel'

These examples show that the Standard German phoneme-grapheme correspondences differ from those defined in the Dieth spelling. The phoneme [kʰ] is spelled *k* in Standard German, but *kh* in the rare Swiss German dialects that use it. The phoneme [kx] is not used in Standard German, but corresponds to *k* in Swiss German.

In some words borrowed from a third language, Standard German *k* corresponds to Swiss German *g* or *gg* in all dialects:

(43) a. Kutsche – Gutsche 'coach'
b. Apotheke – Apoteegg 'pharmacy'

These rules require thus that loanwords can be distinguished from inherited words. It remains to be seen if this distinction can be made effectively.

The grapheme *qu* is pronounced [kxw] in most Swiss German dialects, but its spelling usually remains *qu*. In Basel dialect, it is pronounced [gw] and written *gw*:

(44) Quelle – Quäll(e), Gwälle (BA) 'source'

Aspiration of initial *p* and *t*

Initial *p* and *t* are usually aspirated in Standard German, but not in Swiss German. This phonetic difference is not reflected in spelling. However, there are some loanwords which retain the aspirated pronunciation in Swiss German and which are consequentially spelled with *ph* or *th*. In the case of *t*, the aspiration concerns mainly Greek loanwords which are already spelled with *th* in Standard German. In the case of *p*, the distribution is less clear and does not seem to follow

5 Contrary to the expectations, the word *kein* 'no' is also pronounced with initial [kx]. This is due to assimilation with a preceding *d*: mhd. *dehein* > *d chein* > *kein* (Staub et al. 1881-, Vol. 3, col. 318-319).

precise rules.[6] This phenomenon applies in all Swiss German dialects without geographical differences.

(45) a. Polizei [pʰ] – Polizei [p] 'police'
b. Punkt [pʰ] – Phunkt [pʰ] 'point'

(46) a. Tiger [tʰ] – Tiger [t] 'tiger'
b. Tee [tʰ] – Thee [tʰ] 'tea'
c. Theater [tʰ] – Theater [tʰ] 'theater'

Lenition and Fortification

There is a lot of fluctuation between the lenis consonants *d* and *b* and the fortis consonants *t* and *p*, depending on the phonetic context.

In most parts of Switzerland, word-initial *d* and *t* are collapsed into *t*, except in Northwestern Switzerland, where both consonants are collapsed into *d* (SDS II/164-165; Suter 1976, p. 59).

(47) a. Tag – Taag, Daag (BA, SO) 'day'
b. tragen – traage, trääge, draage (BA, SO) 'to carry'
c. Deckel – Teckel, Deckel (BA, SO, WS, GR) 'lid'
d. drücken – trucke, drugge (BA, SO) 'to press'
e. donnern – tonnere, donnere (BA, SO) 'to thunder'

Some words starting with *d* keep *d* in alpine regions (in addition to the Northwest) (47c), but others keep *d* everywhere (48).

(48) a. Donnerstag – Donnschtig 'Thursday'
b. dreissig – driissg 'thirty'

6 Still, it seems that aspiration is a distinctive feature of Swiss German. Consider the following minimal pair:

(i) teere (AP) [t] / theere [tʰ] 'to torrefy / to tarmac'

There is no general collapse of initial *b* and *p*. If they precede a vowel, the outcome depends on the lexeme (49a)-(49c). If a *p* precedes a consonant (*l, n, r, w*), it is lenited to *b* in Northwestern Switzerland (49d)-(49e). If a *b* precedes a consonant, it is fortified in a discontiguous area consisting of Zürich, Uri and Graubünden dialects (49f) (SDS II/164-169):

(49) a. Pass – Pass 'pass, passport'
b. Bart – Baart 'beard'
c. Bauer – Puur 'farmer'
d. Platz – Platz, Blatz (BA, SO) 'place'
e. Prämie – Präämie, Bräämie (BA, SO) 'prime'
f. Blatt – Blatt, Platt (UR, ZH, GR) 'sheet, leaf'

Let us now turn to word-final contexts. The lenis consonants *b, d, g* are fortified in alpine regions (II/172-175, II/117-122). This same phenomenon occurs in Standard German under the name *Auslautverhärtung* (with the exception of the *ng*-context), but is not reflected in spelling.

(50) a. Rad [t] – Rad, Rat (FR, WS, UR) 'wheel'
b. Auge > Aug – Aug, Augg (FR, WS, UR, GR) 'eye'
c. Kalb [p] – Kalb, Kalp (BEO, WS, UR, GR) 'calf'
d. lang – lang, langg (WS, GR) 'long'
e. Hund [t] – Hund, Hunt (WS, UR) 'dog'

Finally, special transformations apply when plosives occur as part of a consonant cluster in word-medial contexts. The clusters *bl* and *ng* are fortified in alpine dialects (SDS II/171, SDS II/117). In contrast, *rb* is fortified in Northwestern Switzerland (SDS II/170). The cluster *rg* is fortified both in the Northwest and in the Alps, but not in the lowland regions in between (SDS II/202). The last two phenomena contrast thus sharply with the lenition tendencies observed in Northwestern Switzerland.

(51) a. hobeln – hoble, hople (WS, UR, GR) 'to plane'
b. Hunger – Hunger, Hungger (WS, GR) 'hunger'
c. Garbe – Garbe, Garpe (BA, SO) 'sheaf'
d. würgen – würge, würgge (BA, SO, FR, BEO, WS, GR) 'to choke'

Intervocalic consonant gemination

In intervocalic and word-final contexts, fricatives are geminated in Standard German like in Swiss German. The difference lies in spelling: in Standard German, gemination is only explicitated with *s*, where it is also for *f* in Swiss German. However, we avoid the awkward spellings *schsch* and *chch* (or *sch*, *ch*).

(52) a. Straße, Strasse [sː] – Schtraass, Schtròòss 'street'
b. Haufen [fː] – Huffe 'heap'
c. waschen [ʃː] – wäsche [ʃː] 'to wash'
d. machen [xː] – mache [xː] 'to make'

l-vocalization

Vocalization of *l* is perceived as a typical phenomenon of Bern dialect, although it spreads to neighboring dialects of Solothurn and Luzern, and is not traditionally used by the upper class of the city of Berne. *l* becomes an *u*-like vowel in front of a consonant and in word final contexts. In geminated contexts, it becomes a *w*-like semi-vowel (SDS II/147-150, II/197-199):

(53) a. alt – alt, aut (BE, SO, LU) 'old'
b. viel – vil, viu (BE), vöu (SO, LU) 'much'
c. stellen – schtelle, schteue (BE), schtöue (SO, LU) 'to put'

Several spelling conventions have been proposed for this phenomenon: *l, u, w, ɬ, u̯,* (Marti 1985, p. 55). We will use *u* as it is most widely used today and looks the most natural to us.

Note that *l*-vocalization may have an impact on the quality of the preceding vowel (SDS I/112, I/164-165; Marti 1985, p. 57):

(54) Milch – Milch, Miuch, Müuch, Möuch 'milk'

Nasal and glide gemination

Intervocalic geminated *nn, ll* are sometimes shortened. The shortening is characteristic for Zürich German, but extends eastwards to Thurgau and St. Gallen, and is also used in some Graubünden, Freiburg and Berner Oberland dialects (SDS II/186-200). Obviously, *ll* can only be shortened where it is not vocalized.

(55) a. Tanne – Tanne, Tane (FR, BEO, ZH, TG, SG, GR) 'pine'
b. stellen – schtelle, schtele (FR, BEO, ZH, TG, SG, GR) 'to put'
c. Keller – Cheller, Cheler (ZH, GR) 'cellar'

In contrast, intervocalic simple *n, m* is sometimes geminated in Berner Oberland and in the Northeast (SDS II/83-86, II/192-194):

(56) a. kleine – chliini, chlinni (AP, SG, BEO) 'small'
b. Blume – Blueme, Bluemme (AP, SG, BEO) 'flower'

s and *sch*

In alpine dialects (Wallis and Walser dialects of Graubünden), *s* becomes *sch* in some contexts (SDS II/144-146):

(57) a. sie – si, schi (WS) 'she'
b. Häuser – Hüüser, Hiischer (WS) 'houses'

According to Bohnenberger (1913, §99), this transformation occurs as a result of contextual *i* in the Old High German pronunciation of the word: OHG *mûsi* > *Miisch* 'mice'. As the Old High German variants are relatively difficult to reconstruct, we rely on word lists to trigger the *sch*-transformation.

st and *sp*

In Standard German, *st* and *sp* are pronounced [ʃt] and [ʃp] at the beginning of a word, but [st] and [sp] in the middle or at the end of words. In Swiss German (as in some Southern German varieties), these two graphemes are pronounced [ʃt] and [ʃp] in all phonetic contexts (except at *s-t* and *s-p* morpheme boundaries).

The Dieth spelling proposes *st* and *sp* in word-initial contexts, since readers would already be used to pronouncing them [ʃt] and [ʃp] from Standard German. In other contexts however, *scht* and *schp* should be written to emphasize the divergence from Standard German pronunciation.

We find this rule confusing and write *scht* and *schp* in all contexts. This is probably the most explicit solution, although it may be rather cumbersome:

(58)	a.	Strand [ʃt] – Schtrand	'beach'
	b.	Spinne [ʃp] – Schpinne	'spider'
	c.	Mast [st] – Mascht	'mast'
	d.	Wespe [sp] – Wäschpi	'wasp'

n-drop and *n*-insertion

A very common phenomenon in Swiss German is the drop of *n* at word ends:

(59)	a.	Ferien – Ferie	'vacation'
	b.	holen – hole	'to fetch'
	c.	von – vo	'from'

As a consequence of this drop, words ending with a vowel will often be followed by words beginning with a vowel. In order to bridge this hiatus, the *n* will be reintroduced in this context:

(60)	guten Abend – guete + Aabig > gueten Aabig	'good evening'

The additional *n* may even be introduced in cases where there had never been one etymologically:

(61) a. wie er will – wie*n* er wett 'as he wants'
b. ich denke an dich – i denke*n* a di 'I'm thinking of you'

Different spelling variants have been proposed: attaching the *n* to the first word (*wien er wett*), attaching everything together (*wiener wett*), or using hyphens (*wie-n-er wett*). Along with Dieth (1986), we choose the first convention, as exemplified in (61).

The reinsertion of *n* to bridge word boundaries is an instance of *sandhi*. *Sandhi* is defined in WordNet as "the articulatory process whereby the pronunciation of a word or morpheme changes when it is followed immediately by another (especially in fluent speech)".[7] The alternance between the article *a* and *an* is often cited as an example of *sandhi* in English.

Note however that in the area of BE-Haslital, final *n* is never deleted (SDS III/1, III/40). In this region, the corresponding rule only has to deal with the introduction of non-etymological *n*.

This is one of the rare rules that operate across word boundaries. Technically, it is implemented in two steps. In a first step, the special symbol ~ wherever the *n* may appear, without looking at the following word. At the very end of the transformation process, the tilde is replaced according to the context: by *n* if the following word starts with a vowel, or by the empty string if the following word starts with a consonant.

Staub's law

Staub (1877) first described a sound law that is since associated with his name: when a *n* is preceded by a vowel and followed by a fricative consonant, then the *n* disappears and the preceding vowel undergoes a compensatory change, i.e. lengthening or diphthongization.

(62) a. uns – üüs 'us'
b. fünf – füüf, föif 'five'

7 http://wordnetweb.princeton.edu/perl/webwn?s=sandhi, accessed 5.12.2009.

c. Fenster – Fäischter 'window'

Staub's law applies to the first two examples in the whole Swiss German area except Basel (where the words are pronounced *uns* and *fümf*). While *Fäischter* is used in a fairly large area of German-speaking Switzerland, it is in competition with the non-altered variant *Fenschter* (SDS II/128; Wolfensberger 1967, pp. 78-80).

A particular change happened in the Berner Oberland and Wallis regions, where [ŋkx]-clusters developed into [nx], which then triggered Staub's law:

(63) a. trinken – trinche > triiche 'to drink'
b. Bank – Banch > Baich 'bench'

The SDS deals with the general cases in maps II/124-II/136, and with the *nk*-cases in maps II/97-II/107. As the case of *Fäischter* and the example of *Kunscht* cited in the introduction (page 48) suggest, this law is no longer productive and is not applied to modern loanwords. Therefore, we will only use these rules on a word-by-word basis.

nd

The sequence *nd* (Standard German [nt]) is maintained as *nd* [nd] in most Swiss German dialects. Its transformation to *ng* [ŋ] is a characteristic feature of Bern dialect, and the transformation to *nn* a characteristic feature of Freiburg dialect. If this sequence occurs in word-final positions, the Uri and Wallis dialects show *nt* [nt], as a consequence of the general *d* > *t* fortification discussed earlier. In intervocalic position, these dialects keep *nd* or *nn* (SDS II/119-123):

(64) a. Hund – Hund, Hung (BE, SO), Hunn (FR), Hunt (UR, WS) 'dog'
b. Kind – Chind, Ching (BE, SO), Chinn (FR), Chint (UR, WS) 'child'
c. gefunden – gfunde, gfunge (BE, SO), gfunne (FR, WS) 'found'

Note that the *ng*-variant may lead to additional ambiguity:

(65) a. Hang – Hang 'slope'
b. Hand – Hang (BE, SO) 'hand'

As a result, the *ng*-transformation is sometimes blocked in ambiguous cases:

(66) a. Ring – Ring 'ring'
b. Rind – Rind, *Ring (BE, SO) 'cattle'

chs

In Standard German, the sequence *chs* is pronounced [ks], and not, as the spelling would suggest, [xs] or [çs]. Swiss German is more consistent in this respect. In the majority of dialects, the spelling *chs* is maintained and consistently pronounced [xs]. In alpine regions (WS, BEO, UR, GR), in Eastern regions (SG-Rheintal, TG), and in BA, the Where the Standard German pronunciation [ks] is conserved (in the alpine regions of WS, BEO, UR, GR, in the Eastern regions of SG-Rheintal, TG, and in BA), it is written as *ggs* (SDS II/113-116):

(67) Fuchs [ks] – Fuchs [xs], Fuggs [ks] 'fox'

rn, rm and *ln*

The word-final sequence *rn* shows a particular evolution in the alpine dialects. Historically, the consonant cluster *rn* was broken up by inserting a *schwa* (yielding *ren*), and then dropped the final *n*. In the other dialects, the sequence *rn* was maintained. The sequence *rm* underwent an analogous change, albeit restricted to Wallis (SDS II/137-143):

(68) a. gern – gäärn, gääre (GR, GL, UR, WS) 'gladly'
b. Arm – Aarm, Aare (WS) 'arm'

In the abovementioned cases, a full (stressed) vowel precedes the *r*. This contrasts with the Standard German verb classes ending in *-ern* and *-eln* containing a *schwa*. In Swiss German, these verbs have *-ere* and *-le* in the infinitive, but no other irregularities in the finite forms:

(69) a. wandern – wandere 'to hike'
b. wechseln – wechsle, weggsle 'to change'

A.3. Derivational morphology

Most differences in derivational morphology fall into one of two categories. On the one hand, an existing Standard German derivational marker may have a different form, which is in most cases due to the phonetic reduction of unstressed syllables. On the other hand, there are Swiss German derivational markers for which there is no precise equivalent in Standard German. In this section, we will describe differences of both types. However, the second category is less manageable in a machine translation setting, as derivational morphology is usually not very consistent. We will therefore only implement the phenomena that fall under the first category.

There is relatively little interdialectal variation in derivation. However, some suffixes are dialect-specific and are therefore mentioned in the SDS. Other data sources for this compilation (from which we have borrowed many examples) include Haas (2000, p. 96), Lötscher (1983, pp. 86, 100-104) and Rash (1998, p. 141).

Noun diminutives

One of the best-known characteristics of Swiss German dialects is the use of the diminutive suffix *-li* instead of the Standard German forms *-chen* and *-lein*:

(70) a. Hund > Hündchen – Hündli ‘dog’ (dim.)
b. Auge > Äuglein – Öugli ‘eye’ (dim.)

While the Standard German diminutive suffixes always require an *umlaut* in the stem (where applicable), the *-li* suffix can also be used without, especially in Western regions. Moreover, the alternate suffix *-eli* can be used more or less interchangeably, again with or without *umlaut*:

(71) a. Schätzchen – Schätzli, Schätzeli, Schatzli, Schatzeli ‘darling’, litt. ‘treasure’
b. Händchen – Händli, Händeli, Handli, Handeli ‘hand’

The different forms may yield slight semantic nuances; as a general rule, the *-li* suffix with *umlaut* is the most neutral and the most frequent one. In Berner Oberland, Wallis and Graubünden

Walser dialects, the *-li* suffix is replaced by *-i, -ji, -schi, -ti* (SDS III/149-154). As in Standard German, all diminutives are neuter.

(72) a. Hündchen – Hündi (BEO), Hündschi (BEO, WS), Hundschi (WS), Hundji (WS, GR), Hundi (GR) 'dog'
b. Vögelchen – Vögi (BEO), Vogelti (WS) 'bird'

We will allow some simplifications when translating diminutives. All Standard German diminutives (*-chen* and *-lein*) will be replaced by the *umlaut* version of *-li* for plateau dialects, or by the non-*umlaut* version of *-ji* for alpine dialects (SDS III/152). For words ending in *-el*, we will use the plateau dialect diminutive *-eli* (with *umlaut*) and the alpine diminutives *-i* (with *umlaut*) and *-elti* (without *umlaut*) (SDS III/154).

Furthermore, the *-i* suffix is used for diminutive names in a large area. Masculine diminutive names keep the masculine gender, while feminine diminutive names become neuter in Western dialects (see page 342):

(73) a. de Robert > de Röbi
b. d Anna > s Anni, d Anni

Verb diminutives

Standard German only provides morphological means to form diminutive nouns. Swiss German dialects instead have a distinct suffix to create diminutive verbs. This suffix, in the forms *-le* or *-ele*, is attached to the verb root and requires the root to take an *umlaut* if phonotactically possible. The meaning of diminutive verbs can be pejorative:

(74) a. schneie 'to snow' – schneiele 'to snow slightly'
b. schaffe 'to work' – schäffele 'to work a bit'
c. theääterle 'to play theater amateurishly'

This suffix can also be used to express similarity of smell or taste. In this case, the base word can be a noun or an adjective:

(75) a. Fisch 'fish' – fischele 'to smell like fish'
b. grau 'grey, moldy' – gräuele 'to molder'

As Standard German does not possess diminutive verbs, such constructions would have to be generated from scratch, according to some semantic heuristics. We do not consider this possibility for the time being.

Derivation of diminutive adjectives

Lötscher (1983, p. 103) mentions the following "word derivation algorithm":

1. A noun is converted into its diminutive form and shifts its meaning.
2. The diminutive noun is converted into a verb.
3. The past participle of this verb is used as an adjective of quality.

Let us illustrate this phenomenon with two examples:

(76) 1. Blueme 'flower' > Blüemli 'flower on a pattern'
2. Blüemli > blüemle 'to paint flowers repetitively'
3. blüemle > plüemlet 'containing a flower pattern'

(77) 1. Huus 'house' > Hüüsli 'check on paper'
2. Hüüsli > hüüsle 'to checker paper'
3. hüüsle > ghüüslet 'checkered'

Adjectives ending in *-end, -ern, -en*

Adjectives with the suffix *-end* are often (but not always) derived from present participles. They may keep this suffix in Swiss German or replace it with *-ig*:

(78) a. glänzend – glänzig, glänzend 'gleaming'

b. tausend – tuusig 'thousand'

The suffix *-ig* can also be used to generate other adjectives:

(79) a. gfüürchig 'terrifying'
b. gäbig 'convenient'

Adjectives denoting material quality end in *-ern* or *-en* in Standard German, but in *-ig* in Swiss German:[8]

(80) a. hölzern – hölzig 'wooden'
b. golden – goldig 'golden'

Verbal prefixes

The verbal prefixes *zer-* and *er-* are not in use in Swiss German. They are replaced by *ver-* instead:[9]

(81) a. zerbrechen – verbräche 'to break'
b. erzählen – verzelle 'to tell'

The verbal prefix *be-* is not native in Swiss German, but has become increasingly frequent through borrowing, either with (82a) or without (82b) apocope:

(82) a. behalten – bhalte [ph] 'to keep'
b. behandeln – behandle 'to treat'

8 As a result, the words *eisig 'icy'* and *eisern 'iron'* would fall together in Swiss German: *iisig*. The distinction is maintained by differentiating the vowel length: *iisig 'icy'* vs. *isig 'iron'*.
Moreover, the semantic differentiation between Standard German *golden* 'golden' and *goldig* 'like gold, cute' is lost in Swiss German.

9 Lötscher (1983, pp. 103-104) notes that *er-* is sometimes used in Western dialects. In the absence of precise geographical data, we do not consider this variant.

Nouns ending in *-ung*

The Standard German nominal suffix *-ung* is usually pronounced *-ig* in Swiss German. Exceptions occur in the Schaffhausen dialect, in the traditional upper-class dialect of the city of Bern, and in the BE-Oberland and Wallis dialects (SDS III/163):[10]

(83) Zeitung – Ziitig, Ziiting (SH), Ziitung (WS, FR, BEO) 'newspaper'

In the context *-igung*, the Standard German pronunciation is maintained instead of the impractical pronunciation **-igig*:

(84) Ermäßigung – Ermäässigung 'reduction'

Nouns ending in *-ling*

Some nouns with the suffix *-ling* change to *-lig* in Swiss German, while others keep the Standard German suffix:

(85) a. Frühling – Früelig, Früeling (SH) 'spring'
b. Liebling – Liebling 'darling'

Nouns ending in *-end*

Nouns ending in *-end* traditionally use the suffix *-et* in Swiss German. However, the Standard German pronunciations are becoming increasingly popular:[11]

(86) a. Tugend – Tuget, Tugend 'virtue'

10 A special case is the word *Achtung*. Its original abstract meaning 'regard' is reflected in the regular Swiss German derivation *Achtig*, while the interjective meaning 'Caution!' has been imported into Swiss German as a secondary loanword with unchanged pronunciation: *Achtung!*

11 The word *Abend* is particular in the sense that it follows the above rule only in the East, yielding *Oobet*. In the other regions, it follows the rule for *end*-adjectives (discussed above) and yields *Aabig* (SDS IV/13).

b. Jugend – Juget, Jugend 'youth'

Phonetically reduced derivational suffixes

The unstressed suffixes *-at, -eit* are often phonetically reduced to *-et*. The Standard German suffixes *-heit, -keit* are usually maintained (depending on the dialect, with the diphthong *ei, äi*, or *ai*, see page 293), but can also be reduced in some cases (Lötscher 1983, p. 86):

(87) a. Heimat – Häimet 'homeland'
b. Arbeit – Arbet 'work'
c. Gesundheit – Gsundhäit 'health'
d. Arbeitslosigkeit – Arbetslosigkäit 'unemployment'
e. Wahrheit – Woorhäit, Woret 'truth'

Abstract nouns

In Standard German, there are some feminine abstract nouns ending in *-e*. Instead of dropping the *e*, these nouns take a final *i* in Swiss German:

(88) a. Suche – Suechi 'search'
b. Größe – Gröössi 'size'

Map SDS III/187 deals with the plural form of *i*-nouns (see also page 347):

(89) a. Größen – Gröössine (W), Grössene (O) 'sizes'

However, the *i*-suffix is more productive in Swiss German and is also used in words that are constructed differently in Standard German:

(90) a. Trockenheit – Tröchni 'dryness'
b. Kleider – Aaleggi (< aalegge 'to dress') 'clothes'
c. Wahl – Weli (BE) 'choice'

Deverbal intensifying affixes

Lötscher (1983, p. 102) mentions three characteristic dialect markers (*-erei, -ete, G-* to form action nouns out of verbs; all markers may have a pejorative meaning:

(91) a. brüele 'to yell' > Brüelerei, Brüelete, Prüel (< Gbrüel) 'yelling'
b. choche 'to cook' > Chocherei, Chochete, Gchöch 'cooking'

Note that the *-erei* and *Ge-* affixes are also used in Standard German, albeit less productively.

Masculine agent nouns

The Standard German agent suffix *-er* is also used productively in Swiss German. A variant *-ler* is used to denote professional or habitual activity:

(92) a. Poscht 'postal service' > Pöschtler 'postman'
b. Hund 'dog' > Hündeler 'dog-fancier'

Another dialect suffix *-i* is mainly used in colloquial speech with pejorative connotation:

(93) bhoupte 'to allege' > Bhoupti 'dogmatist'

Feminine derivation of agent nouns

In Standard German, the suffix *-in* (plural *-innen*) is used to transform a masculine agent noun into a feminine agent noun:

(94) a. Koch – Köchin 'cook' (masc. – fem.)
b. Lehrer – Lehrerin 'teacher' (masc. – fem.)

If the word is a *-er/-ler* derivative, then the feminine suffix is *-e* in Western Swiss German, and *-i* in traditional Eastern Swiss German. However, the full Standard German suffix *-in* is also in common use (SDS III/159-161):

(95) Lehrer – Lehrere (W), Lehreri (O), Lehrerin

If the word does not end in *-er*, the suffix is more variable (SDS III/162):

(96) Wirt – Wirtin, Wirtene, Wirteni (LU, SZ)

The plural form of *Wirtene* and *Wirteni* is *Wirtene*, while the plural of the more standard-like *Wirtin* is also more standard-like *Wirtinne*.

Compound nouns and *Kompositionsfuge*

Standard German allows nearly unlimited noun composition. Starting with a head noun, adjunct nouns, adjectives, and verb stems can be added to its left. The resulting compound noun takes the gender of its head, and morpheme boundaries are usually not marked (only in complicated cases, hyphens are inserted to help the reader). Compound adjectives and compound verbs can also be constructed, but these cases are less frequent and will not be treated here.

Several problems arise for compound noun translation (see for example Popovic, D. Stein, and Ney 2006). If the entire compound noun is not available in the lexicon of the translation system, it is disassembled and reassembled. However, this is not trivial because special material may be added between the adjunct and the head. This material is called *Kompositionsfuge* and consists of either one of the infixes *-e-, -en-, -s-*. The choice of the infix is mostly arbitrary and gives rise to some interdialectal differences. The *s*-infix is more frequent in Southern varieties of Standard German:[12]

(97) a. Schwein*e*braten – Schwein*s*braten ‘pork roast’
b. Zugmitte – Zug*s*mitte ‘center of the train’

Swiss German dialects follow the Southern variants:

(98) a. Schwein*e*braten – Schwin*s*braate
b. Zugmitte – Zug*s*mitti

12 http://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/ada/runde_3/f01a-d, accessed 14.3.2012.

A.4. Morphosyntax

Many morphological characteristics of Swiss German are interdependent with problems of syntax. For example, the lack of verbal preterit forms – a problem of inflectional morphology – calls for a periphrastic expression of past tense, raising some word order issues – a purely syntactic problem. Therefore, we have chosen to discuss inflectional morphology and syntax in the same chapter. For each part of speech, we will begin by presenting its inflectional paradigm, then move on to some syntactic issues that are related to inflection, and finally disuss some syntactic patterns that are independent of the morphological realization of the words.

General descriptions of Swiss German present the most salient features of morphology and syntax (Haas 2000, pp. 96-98; Baur 1983, pp. 30-32; Lötscher 1983, pp. 104-118; Rash 1998, p. 143), but fail to discuss phenomena with smaller geographic extensions. Grammars of specific dialects may clarify some specific morphology problems, but rarely contain much information about syntax.

For a long time, syntax has been a less popular area of study in dialectology, caused by the difficulty of obtaining reliable grammaticality judgements from untrained informants,[13] and partly also by lack of interest by dialectologists. For instance, the SDS only contains about a dozen maps related to syntactic phenomena (SDS III/126, 137-139, 141, 235, 256, 259-266). The SADS project is about to fill this gap. Bucheli and Glaser (2002), Glaser (2003), and Klausmann (2006) present preliminary findings, but we will also draw on raw results obtained from further questionnaires since then.

A.4.1. Topological fields

Before discussing the details of Swiss German morphosyntax, let us define some technical terms of descriptive German syntax.[14]

13 Because of the dominance of Standard German in Swiss school curricula, there is no metalinguistic reflection whatsoever about the dialects.

14 This account is inspired by http://www.canoo.net/services/OnlineGrammar/Satz/Wortstellung/index.html (accessed 25.1.2010), but there are similar presentations in German grammars, for instance Engel 1998, Zifonun

The word order of German sentences is defined by five so-called *topological fields*. The second and the fourth field contain verbal elements and form the so-called *sentence bracket*. The other fields are filled with non-verbal elements (see Figure A.2).

Vorfeld (VF)	Linke Satzklammer (LK) Left sentence bracket	Mittelfeld (MF)	Rechte Satzklammer (RK) Right sentence bracket	Nachfeld (NF)

Figure A.2.: Topological fields in German.

We can distinguish three types of sentences:

- In declarative main clauses (99a) and interrogative sentences (99b) with a *wh*-word, the finite verb is placed in the left sentence bracket, and the *Vorfeld* is not empty. As a result, the finite verb is in second position.

- Polar interrogative sentences and exclamative sentences are characterized by the sentence-initial position of the finite verb (99c). Hence, they require the *Vorfeld* to be empty.

- In subordinate clauses, all verbal material is in the final position (i.e. in the right sentence bracket). The *Vorfeld* is empty, and the left sentence bracket is occupied by the complementizer.

(99) a. $[\text{Sie}]_{\text{VF}}$ $[\text{kauft}]_{\text{LK}}$ $[\text{ein Auto}]_{\text{MF}}$ $[]_{\text{RK}}$ $[]_{\text{NF}}$. 'She buys a car.'

b. $[\text{Wann}]_{\text{VF}}$ $[\text{kauft}]_{\text{LK}}$ $[\text{sie} \mid \text{ein Auto}]_{\text{MF}}$ $[]_{\text{RK}}$ $[]_{\text{NF}}$. 'When does she buy a car?'

c. $[]_{\text{VF}}$ $[\text{Kauft}]_{\text{LK}}$ $[\text{sie} \mid \text{ein Auto}]_{\text{MF}}$ $[]_{\text{RK}}$ $[]_{\text{NF}}$? 'Does she buy a car?'

d. $[]_{\text{VF}}$ $[\text{dass}]_{\text{LK}}$ $[\text{sie} \mid \text{ein Auto}]_{\text{MF}}$ $[\text{kauft}]_{\text{RK}}$ $[]_{\text{NF}}$. 'that she buys a car.'

Depending on the type of clause or sentence, additional constraints apply:

- The *Vorfeld* is occupied by at most one constituent (subject, object, modifier, adverbial).

- In main clauses, the *left sentence bracket* contains the finite verb. In subordinate clauses, it contains the complementizer.

et al. 1997.

- The *Mittelfeld* can hold several constituents of different types, e.g., objects, subjects, modifiers or adverbials (100). The order of the different constituents is relatively free (scrambling).

- The *right sentence bracket* contains all verbal material that does not fit into the left bracket. In main clauses, it typically contains separable verb particles and non-finite verb forms. In subordinate clauses, all verbal material is accumulated in the right bracket (101).

- The *Nachfeld* can only take one constituent, and only from a restricted set, for instance relative clauses relating to one of the constituents in the *Mittelfeld*, prepositional phrases, heavy noun phrases, subordinate clauses, comparative expressions (102).

(100) a. $[\text{Heute}]_{\text{VF}}$ $[\text{kauft}]_{\text{LK}}$ [sie | ausnahmsweise | mit ihrem Mann | ein Auto]$_{\text{MF}}$.
b. $[\text{Heute}]_{\text{VF}}$ $[\text{kauft}]_{\text{LK}}$ [sie |mit ihrem Mann | ausnahmsweise | ein Auto]$_{\text{MF}}$.
'Today, she exceptionally buys a car with her husband.'

(101) a. $[\text{Sie}]_{\text{VF}}$ $[\text{holt}]_{\text{LK}}$ [ihr Auto]$_{\text{MF}}$ [ab]$_{\text{RK}}$. 'She picks up her car.'
b. $[\text{Sie}]_{\text{VF}}$ $[\text{hat}]_{\text{LK}}$ [ein Auto]$_{\text{MF}}$ [gekauft]$_{\text{RK}}$. 'She has bought a car.'
c. $[\text{dass}]_{\text{LK}}$ [sie | ein Auto]$_{\text{MF}}$ [kaufen | will]$_{\text{RK}}$. 'that she wants to buy a car.'

(102) a. $[\text{Man}]_{\text{VF}}$ $[\text{wird}]_{\text{LK}}$ [sich | noch]$_{\text{MF}}$ [beschäftigen]$_{\text{RK}}$ [mit diesem Problem]$_{\text{NF}}$.
'One will deal with this problem.'
b. $[\text{Sie}]_{\text{VF}}$ $[\text{hat}]_{\text{LK}}$ [das Auto]$_{\text{MF}}$ [gekauft]$_{\text{RK}}$, [von dem sie träumte]$_{\text{NF}}$.
'She has bought the car she dreamt of.'

The Swiss German sentence structure maintains most of these constraints; syntactic differences mostly occur inside phrases and do not touch the general sentence structure. Major reorderings only occur with verb clusters (see page 371).

A.4.2. Determiners and pronouns

The Swiss German stock of determiners and pronouns is roughly the same as in Standard German. However, these mostly unaccented words have been phonetically reduced, leading

in some cases to cliticization. While traces of cliticization can be found in spoken Standard German, they are not reflected in spelling. Therefore, some complications may be expected when translating Standard German to Swiss German.

As in Standard German, the Swiss German determiners and pronouns agree with gender, number and case. The three genders (masculine, feminine, neuter) are only overtly distinguished in the singular forms, and a unique plural form is used for all genders. Swiss German dialects mainly use three cases: nominative, accusative, and dative, where the nominative and accusative forms are identical (except in a small area, see page 327). Use of genitive is marginal.

Practically, the Standard German pronouns will be replaced by Swiss German variants on a form-by-form basis. The correspondences are not regular enough to incorporate them into generic phonetic rules.

The following presentation, including some of the tables, is inspired by Nübling (1992, Chapter II.2.1).

The definite article

The following table sums up the Swiss German forms of the definite article. Genitive articles are marginal and are not presented here (SDS III/131-136, III/140):[15]

15 The Standard German forms are as follows:

(i)

	Nominative/Accusative	*Dative*	*Genitive*
Singular Feminine	die	der	der
Singular Masculine	der / den	dem	des
Singular Neuter	das	dem	des
Plural	die	den	der

(103)

	Nominative/Accusative	*Dative*
Sing. Fem.	d [-Adj] / di [+Adj]	dr (W), de (O)
Sing. Masc.	dr (W), dr [+V] / de [-V] (O)	em, dem (BEO), dum (WS)
Sing. Neut.	ds (S), s (N)	em, dem (BEO), dum (WS)
Plural	d [-Adj] / di [+Adj]	de

The determiner *d* is an abstraction and hardly ever pronounced as such. If the following word begins with a *m* or *f*, the determiner is realized as *p* (104a). If the following word begins with a *b, d,* or *g*, that letter is merely devoiced (104b). Before voiceless consonants (*p, t, k*), the determiner disappears altogether (104c). Despite these assimilations, it is customary to write the determiner as *d* in all cases:

(104) a. d Frau [pfraʊ] 'the woman'
b. d Blueme [pluəmə] 'the flower'
c. d Tante [tantə] 'the aunt'

The plural and feminine singular forms vary according to the presence of an adjective, of a numeral, or of a substantivized adjective:

(105) a. *d* Fläsche 'the bottle'
b. *di* gross(i) Fläsche 'the big bottle'
c. *di* drüü Fläsche 'the three bottles'
d. *di* Grosse 'the big ones'

In parts of Wallis, Graubünden, and the BE-Haslital region, traditional dialects have distinct nominative and accusative forms for masculine determiners (SDS III/136). However, there is a tendency to generalize the nominative form.

(106) a. der Baum (Nom) – dr Baum 'the tree'
b. den Baum (Acc) – de Baum (WS, GR), den Baum (BE-Haslital) 'the tree'

Determiners frequently appear right after prepositions. Depending on the phonotactic shape of the preposition (ending in a consonant or in a vowel) and the case it governs (accusative or dative), some assimilations and contractions (set in italic below) occur (SDS III/137-139, SADS

I/14).[16]

(107)

	Accusative		*Dative*	
Preposition	i 'into'	für 'for'	i 'in'	mit 'with'
Sing. Fem.	i d / i di	für d / für di	*ir* (?i dr) (W) i de (O)	mit dr mit de
Sing. Masc.	*i* (*i dr), *in* (*i de), *inn* (WS), in de (GR)	*füre* (?für dr) für de	*im* (*i em) i dem	mit em mit dem
Sing. Neut.	*is* (?i ds) is	*fürs* (?für ds) fürs	*im* (*i em) i dem	mit em mit dem
Plural	i d / i di	für d / für di	i de in de (BA, GR)	mit de

The Standard German masculine/neuter contractions of the type *in dem* → *im* thus also exist in Swiss German. But while the Standard German feminine contraction *zu der* → *zur* only occurs in this token, nearly all prepositions allow this type of contraction in Western Swiss German.[17]

The indefinite article

The following table contains the dialect forms of the indefinite article (SDS III/142-145, SADS IV/35, IV/37). As in Standard German, indefinite plural is expressed by the *null*-morpheme.[18]

16 For reasons of readability, some geographical indications are left out in the following table.

17 As a result, the form *vor* is ambiguous in Western Swiss German:

(i) vor – vor 'before, in front of'

(ii) von der – vor (< vo dr) 'from the'

18 In written Standard German, the indefinite determiner has the following paradigm:

(108)

	Nominative/Accusative	*Dative*
Singular Feminine	e, en (AG, WS)	ere, enere
Singular Masculine	e (W), en (O)	eme, emene
Singular Neuter	es, e (SG, TG, AP)	eme, emene
	e [+Adj] / es [-Adj] (ZH, GL)	
Plural	—	—

Assimilations with preceding prepositions only occur if these end in a vowel. In accusative forms, a *n* is inserted to bridge the hiatus. Dative forms show additional assimilations. After dative prepositions ending in a consonant, the short article forms (*ere, eme*) are proportionally more frequent than in non-prepositional contexts (SADS IV/34, IV/36).

(109)

	Accusative	*Dative*
Preposition	i 'into'	i 'in'
Singular Feminine	i*n*e, i*n*en (AG, WS)	*inere*
Singular Masculine	i*n*e (W), i*n*en	*imene* (*i eme)
Singular Neuter	i*n*es (W), i*n*e (O)	*imene* (*i eme)
Plural	i	i

Prepositional dative marking

Central Swiss dialects present the so-called *prepositional dative marking* phenomenon (Seiler 2002). In these dialects, the formal dative case has been reanalyzed as a prepositional case which can only occur after prepositions. In other words, it is no longer possible to use raw dative case to mark indirect objects, and indirect objects are marked with prepositional phrases headed by

(i)

	Nominative/Accusative	*Dative*	*Genitive*
Singular Feminine	eine	einer	einer
Singular Masculine	ein/einen	einem	eines
Singular Neuter	ein	einem	eines
Plural	—	—	—

Note that shortened forms are current in spoken non-dialectal German.

a or *i*,[19] depending on the region (SADS I/2, I/7, I/20, II/2, II/12, II/18). In true prepositional phrases, no additional preposition is added:

(110) a. Er hilft der_{Dat} Mutter. – Er hilft *i* de_{Prep} Mueter.
b. mit der_{Dat} Mutter – mit de_{Prep} Mueter

Hence, in these regions, the raw dative forms of tables (103) are non-existent, and only the forms of table (107) exist, either with a "real" preposition or with the dummy indirect object preposition *a* or *i*. The same applies to the indefinite article.

Historically, prepositional dative marking is thought to have arisen by the phonetic enrichment of the dative forms *em* to *am* or *im*, and *emene* to *amene* or *imene*. Thanks to the existence of the prepositions *a* 'at' and *i* 'in' in Swiss German, these forms have subsequently been reanalyzed as preposition + clitic determiner (*a=m, i=m, a=mene, i=mene*), and the preposition has spread to other forms (Nübling 1992, pp. 206-210).

Demonstrative determiners

The Standard German definite article *der, die, das* has evolved into two series in Swiss German. An unaccented variant, characterized by further phonetic reduction, has given rise to the Swiss German definite determiner presented above. An accented variant has become the most common Swiss German demonstrative determiner.[20] Its paradigm is presented below (SDS III/224):

(111)

	Nominative/Accusative	*Dative*
Singular Feminine	die [diə]	däre, där
Singular Masculine	dä	däm, dem
Singular Neuter	das, da (NO)	däm, dem
Plural	die [diə]	däne, dene

19 These prepositions are the equivalents of Standard German *an* 'at' and *in* 'in', but they are entirely grammaticalized in Central Swiss German and have lost their locative sense.
Direct objects continue to be marked with raw accusative case in all dialects.

20 Standard German demonstrative lexemes like *dieser, jener* are only marginally used in Swiss German.

Besides some phonetic details, the only relevant isogloss concerns the drop of final *s* in neuter determiners in Northeastern Switzerland.

Interrogative pronouns and determiners

In most dialects, interrogative pronouns look similar to the Standard German counterparts, with the exception of Wallis dialects which use a different lexeme. In line with the demonstrative, the interrogative non-human pronoun drops the final *s* in Northeastern Switzerland (SDS III/221-223, SADS III/2-3, IV/4).

(112)

	Nominative	*Accusative*	*Dative*
Human	wer , wels (WS)	wer , wen (O), wee (O), wem (BE), wels (WS)	wem, welem (WS)
Non-Human	was, wa (NO)	was, wa (NO)	wem

The interrogative determiner *wele* 'which' is inflected as follows:

(113)

	Nominative/Accusative	*Dative*
Singular Feminine	weli	welere
Singular Masculine	wele	welem
Singular Neuter	weles	welem
Plural	weli	welne

Personal pronouns

Standard German spelling does not distinguish stressed from unstressed personal pronouns; there is thus a single series of personal pronouns, used in any syntactic and pragmatic context.

In contrast, three series of personal pronouns have been proposed for Swiss German: stressed, proclitic, and enclitic pronouns (Baur 1983, pp. 31-32; Lötscher 1983, pp. 93-94). Because of different phonetic realizations, these distinctions are usually maintained in writing. According

to German syntax, pronouns may occupy the following positions:

1. The *Vorfeld*, immediately before the inflected verb of a main clause.

2. The first position(s) in the *Mittelfeld* of a main clause, immediately following the inflected verb, or another pronoun.

3. The first position(s) in the *Mittelfeld* of a subordinate clause, immediately following the complementizer, or another pronoun.

4. If stressed, the pronoun may occur in non-initial positions of the *Mittelfeld* of main or subordinate clauses.

Stressed Swiss German pronouns can occur in all four syntactic contexts:

(114) a. *Sii* sind cho. 'They have come.'
b. Hüt sind *sii* cho. 'They have come today.'
c. wil *sii* cho sind. 'because they have come.'
d. wil immer au *sii* mitchunt. 'because she always comes along.'

Proclitic pronouns occur in context 1 (115a), while enclitic pronouns occur in contexts 2 (115b) and 3 (115c).

(115) a. *Si* sind cho.
b. Hüt sind *s* cho.
c. wil*s* cho sind.

The Swiss German dialects distinguish different forms for nominative, accusative, and dative cases. Because of syntactic constraints, the proclitic variants of accusative or dative pronouns are very rare, and are not displayed below.[21] The following table presents the dialectal forms of

21 They may occur with verbs of psychological state, as in the following Standard German example (O. Rambow, p.c.):

(i) Ich bin sehr wählerisch. *Mir* gefallen nur Silberohrringe.'I am very picky. I only like silver earrings.'

the first and second person singular pronouns (SDS III/195-197, SADS III/4, III/11).[22]

(116)

1st Singular	Stressed	Proclitic	Enclitic
Nominative	ii, iich, iig	i, ich	i, ich, Ø
Accusative	mii, miich, miir (FR)		mi, mich
Dative	miir, mier		mer

(117)

2nd Singular	Stressed	Proclitic	Enclitic
Nominative	duu	du, de, Ø	t, Ø
Accusative	dii, diich, diir (FR)		di, dich
Dative	diir, dier		der

The *null*-realizations of clitics is commonly admitted (Cooper 1995, ch. 3) to be a result of morpho-phonological assimilation rather than an instance of the (morphosyntactic) *pro-drop* phenomenon. However, the exact distribution of overt and *null* realizations is not entirely clear. Cooper (1995, ch. 3) provides an overview of different accounts with respect to Zürich German (although I should say that I disagree with an exceptional amount of her grammaticality judgements).

The first person singular nominative enclitic is left out in many dialects if it is followed by a dative pronoun; SADS I/15 shows that this does not hold in the Olten region (parts of BE, SO, AG) and in WS:

(118) Habe ich dir das schon erzählt?
- Ha der das scho verzellt?
- Han*i* der das scho verzellt? (BE, SO, AG, WS) 'Have I already told you that?'

In the second person, the *null*-enclitic is rather used after verbs, and the *t*-enclitic rather attaches to a complementizer:

(119) a. Hüt bisch-Ø cho. 'Today you have come.'

22 Proclitic and enclitic *ich* is in fact reduced to *ch*. The latter spelling is used by Cooper (1995), but is unconventional otherwise.

b. wil*t* cho bisch. 'because you have come.'

Nübling is right to claim that demonstrative pronouns are often used instead of stressed third person personal pronouns, especially when they refer to non-human antecedents. However, we do not repeat these forms here to keep the tables readable (relevant maps: SDS III/199-210).

(120)

3rd Sing. Fem.	Stressed	Proclitic	Enclitic
Nominative	sii, schii, seie	si, schi	si, se, schi
Accusative	sii, schii, seie		si, se, schi
Dative	ire, iere		ere, re

(121)

3rd Sing. Masc.	Stressed	Proclitic	Enclitic
Nominative	er, är	er	er
Accusative	in, en		en, in, ne, nu
Dative	im		em, nim, mu

(122)

3rd Sing. Neut.	Stressed	Proclitic	Enclitic
Nominative	äs, es	es	s, es
Accusative	äs, es, ins		s, es
Dative	im, em		em, nim, mu

The dative and accusative forms coincide in the first and second person plural:

(123)

1st Person Plural	Stressed	Proclitic	Enclitic
Nominative	miir, mier, wier	mer, wer	mer, wer
Accusative/Dative	üüs, iis, insch		üs, is, nis, nisch

(124)

2nd Person Plural	Stressed	Proclitic	Enclitic
Nominative	iir, ier, diir	ir, dir	er, der
Accusative/Dative	öi, öich, euw, oi		i, ni, ech, nech, oi

(125)

3rd Person Plural	Stressed	Proclitic	Enclitic
Nominative	sii, schii, seie	si, schi	s, si
Accusative	sii, schii, seie		s, si
Dative	ine, ene		ine, ene, ne

One may note that the stressed and proclitic pronouns are usually very similar, the proclitics having a shorter vowel than the full variants. In order to reduce the complexity of these paradigms, we have chosen the following implementation strategy: we will not mark length explicitly in the stressed forms. By doing this, we deviate from the Dieth rules, but follow most dialect writers. We will thus write *ich, du, si, mir, ir*. As a result of this spelling change, most proclitics will be homograph with the stressed pronouns. This allows us to dispense with implementing a separate proclitic series. When translating from Standard German, it would be difficult to tell apart stressed and unstressed pronouns anyway.

The indefinite pronoun *me/mu*, equivalent of Standard German *man* 'one', is used with a third person singular verb form. As in Standard German, its accusative and dative forms are suppletive on the basis of the indefinite pronoun *einen/einem*. There is no distinction between stressed and clitic forms (SDS III/229-230, SADS IV/2):

(126)

Nominative	me, mu
Accusative	eim, aim, ein (UR, SZ), eis (WS, GR)
Dative	eim, aim

Let us conclude this section with three notes.

- A *n* is inserted between a verb form ending in a vowel and an enclitic pronoun starting with a vowel:

  (127) a. wir geben ihm – mer gäbe*n* im (BE) 'we give him'
  b. dies suche ich – das suechen i (W), das suech i (O) 'this is what I'm searching'

  The latter example (SDS III/25) hints at an East-West division of *-e* and *-Ø* first person singular forms (see page 357). However, this division depends on the context, and *-Ø*

forms are more common if they are not followed by a clitic (SDS III/22).

- There is a crucial word-order difference in clitic clusters. If a sentence contains an accusative and a dative pronoun at the same time, Standard German will have the accusative pronoun first, while Swiss German will have the dative pronoun first:

  (128) a. Sie hat es$_{\text{Acc}}$ mir$_{\text{Dat}}$ erzählt.
  – Si hät mer$_{\text{Dat}}$s$_{\text{Acc}}$ verzellt. 'She has told me (it).'
  b. Ich muss ihn$_{\text{Acc}}$ dir$_{\text{Dat}}$ zurückgeben.
  – I mues der$_{\text{Dat}}$en$_{\text{Acc}}$ zrugggee. 'I have to give it back to you.'

  This behavior is fairly consistent across dialects, but the Standard German word order is accepted as well in many dialects (SDS III/260; SADS III/7, III/9, IV/3, IV/8). It is unclear if *es*$_{\text{Acc}}$ triggers a different word order than *es*$_{\text{Nom}}$.

- Spelling of clitics is notoriously problematic: the question whether clitics should be attached to their base or figure as separate words has been answered differently in different languages. Likewise, the Dieth rules are not consistent in this respect. He suggests to write the *-t* enclitic as a separate word *d* (suggesting its phonetic reduction from *du*), while he proposes to attach *-s* and *-mer* (Dieth 1986, pp. 44-45). We have decided to generate all clitics as separate words; as needed, they can be attached to their base with simple heuristics.

Polite form

Standard German realizes polite forms with the (capitalized) pronoun *Sie* and the third person plural. Eastern Swiss German dialects use the same pattern. In contrast, Western Swiss German dialects use the second person plural with the pronoun *Dir* (BE, SO, BL) or *Iir* (BS, FR, BEO, WS) (SDS V/117; secondary material in SADS IV/6).

Reflexive and reciprocal pronouns

The third person reflexive pronoun is *sich/sech* in all Swiss German dialects. Sometimes, the personal pronouns are substituted, but this phenomenon is marginal (SADS III/13, 20).

The reciprocal pronoun *enand/enander* corresponds to Standard German *einander*.

Relative pronouns

Most Swiss German dialects use an invariable relative particle *wo* (or a phonetic variant *wa*, unfortunately not registered in the SDS and SADS) to introduce the subject or accusative object of the relative clause. The corresponding Standard German relative pronoun is inflected.[23] This simplification is an innovation of Swiss German (Lötscher 1983, p. 111).

(129) der Mann, *der* gestern gekommen ist – de Maa, *wo* geschter choo isch
die Frau, *die* gestern gekommen ist – d Frau, *wo* geschter choo isch
das Kind, *das* gestern gekommen ist – s Chind, *wo* geschter choo isch
die Leute, *die* gestern gekommen sind – d Lüüt, *wo* geschter choo sind
'the man/woman/child/people who came yesterday'

To introduce the dative object of a relative clause, dialect grammars (for example, Lötscher 1983) maintain that a personal pronoun should be added:

(130) der Mann, *dem* die Polizei einen Brief geschrieben hat
– de Maa, *won em* d Polizei en Brief gschribe hät
'the man to whom the police wrote a letter'

In this example, *wo* (or rather, its *sandhi* variant *won*) is followed by the personal pronoun *em*. However, recent dialectological data (SADS II/2, II/18) show that this solution is preferred only in Northwestern dialects (BA, SO, BE). Other dialects use the simple particle *wo* also for dative

23 SADS II/20 lists occurrences of *der, die, das* relative pronouns. They are scattered all over Switzerland, but do not constitute majority answers in a contiguous area. We neglect these data.

resumption:[24]

(131) der Mann, *dem* die Polizei einen Brief geschrieben hat
– de Maa, *wo* d Polizei en Brief gschribe hät

Prepositional objects referring to humans are introduced with *wo* + prepositional phrase with pronoun in the plateau dialects; alpine dialects use a Standard German-like pronoun *dem*, either with or without *wo* (SADS II/28). *woni* is a contraction of the relative particle with the nominative personal clitic *i* 'I':

(132) der Professor, *von dem* ich erzählt habe
– de Profässer, *wo*ni *von em* verzellt ha
– de Profässer, *vo dem wo*ni verzellt ha (FR, WS)
– de Profässer, *vo dem* i verzellt ha (GR) 'the professor about whom I talked'

With impersonal prepositional objects, pronominal adverbs are used instead. Simple heuristics might be sufficient to distinguish pronouns referring to humans from impersonal ones.

(133) das Waschmittel, *für das* man Reklame macht
– s Wäschmittel, *wo* me Reklame *defüür* macht
'the detergent for which they show advertisement'

The prepositional phrase may be placed in the *Mittelfeld* (132)-(133), or in the *Nachfeld* (134). The data of SADS II/28 suggest that the *Mittelfeld* variant is preferred, but it is unclear whether this a general effect of prepositional phrases or only occurs in relative clauses.

(134) a. de Profässer, *wo*ni verzellt ha *von em*
b. s Wäschmittel, *wo* me Reklame macht *defüür*

24 For a recent theoretical account of this phenomenon within the framework of Stochastic Optimality Theory, see Salzmann and Seiler (2011).

Possessive pronouns

The following table presents the inflection of the first person singular possessive pronoun. For the second person singular pronoun forms, replace the first letter *m* by *d*. For the third person singular masculine and neuter forms, replace *m* by *s* (SDS III/211-214).

(135)

	Nominative/Accusative	*Dative*
Sing. Fem.	mini, mi	mir, mire, miner, minere
Sing. Masc.	min, mi, mine	mim
Sing. Neut.	mis	mim
Plural	mini, mi, miner	mine

The third person singular feminine forms and the third person plural forms are identical (SDS III/215-216):

(136)

	Nominative/Accusative	*Dative*
Sing. Fem.	iri, irni, iru	irer, irere
Sing. Masc.	ire, iren, iru	irem
Sing. Neut.	ires	irem
Plural	iri, irni, iru	ire, irne

The first and second person plural forms are given below (SDS III/217-219).

(137)

	Nominative/Accusative	*Dative*
Sing. Fem.	üsi, öisi, üseri	üser, üsere, öisere
Sing. Masc.	üse, öise, üsere	üsem, öisem, üserem
Sing. Neut.	üses, öises, üsers	üsem, öisem, üserem
Plural	üsi, öisi, üser, üseri	üse, öise, öisne, üserne

(138)

	Nominative/Accusative	*Dative*
Sing. Fem.	öii, öieri, iiweri, euwi	öier, öiere, iiwere, euwere
Sing. Masc.	öie, öiere, iiwere, euwe	öiem, öierem, iiwere, euwem
Sing. Neut.	öies, öiers, iiwers, euws	öiem, öierem, iiwere, euwem
Plural	öii, öieri, iiweri, euwi	öie, öine, öiere, iiwere, euwe

Indefinite pronouns and determiners

The pronouns and determiners in the following table have been studied in either the SDS or the SADS:

(139)
a. jemand – öpper(t), näber(t), eswer, eswels (SDS III/225) 'someone'
b. etwas – öppis, näbis, eppis, eswas, was (SDS III/226; SADS IV/6, 16) 'something'
c. kein – kei, kein (SDS III/231, 232) 'no'
d. alles – ales, aus, alls (SDS IV/149) 'everything'
e. nichts – nüt, nüüt, nünt, nix (SDS IV/171) 'nothing'
f. ein wenig – e chli, e bitzeli, en bitz, e weeni, öppis (SDS IV/155, 164; SADS II/8) 'a bit'

Note that unlike Standard German, the lexemes *öppis, nüüt* have (facultative) dative forms *öppisem, nüütem* (Lötscher 1983, p. 95).

There are a few other pronouns and determiners for which no dialectological data are available:

(140)
a. niemand – niemer, niemert 'no one'
b. jeder – jede 'every'
c. ein paar – e paar 'some'

Alpine dialects have a pronoun *sum* which corresponds etymologically and semantically to English *some*, but did not leave any traces in other varieties of German (SDS III/228).

Pronominal partitive

A geographically restricted phenomenon concerns the presence of the partitive pronoun *re*, as illustrated in the following sentence:

(141) …Milch …Soll ich welche kaufen gehen?
– Söll i *re* go chaufe? '…milk…Shall I buy some?'

The usage of this pronoun corresponds to French *en*:

(142) Veux-tu que j'aille *en* acheter? 'Do you want me to buy some?'

The *re*-pronoun is restricted to Berner Oberland, parts of Wallis (in the form *ru*) and parts of the Walser dialects in Graubünden (SDS III/235; SADS I/18, III/6). In the other regions, this pronoun is left out altogether, or is replaced by a quantifier like *echli 'a bit'* or an prepositional pronoun like *drvo 'of that'*.

Expletive *es*

In Standard German, impersonal passive sentences require an expletive pronoun *es* if the *Vorfeld* is not filled by other material:

(143) a. *Es* wird gearbeitet. 'Work is being done', litt. 'It is worked.'
b. Heute wird Ø gearbeitet. 'Today, work is being done', litt. 'Today is worked.'

In a small area in Aargau and Luzern, the pronoun is maintained in the *Mittelfeld* position (SADS I/13; Glaser 2003):

(144) Hüt wird *s* gwärchet. 'Today, work is being done', litt. 'Today is it worked.'

Other dialects go the opposite way and allow an impersonal pronoun to be dropped where Standard German requires it, for instance a "weather-it":

(145) Am Morgen kann es noch kalt sein.

– Am Morge cha *(*s)* no chalt sii. 'In the morning, it can still be cold.'

However, the available data do not show a clear geographic delimitation of this phenomenon (SADS III/15). These variables are interesting from a syntactic and dialectological point of view, but are hardly essential to make a difference in a machine translation system.

Definite article before names

Contrarily to Standard German, but similarly to other substandard varieties (e.g. of French, Spanish, or German), person names are preceded by definite articles in Swiss German (Lötscher 1983, p. 105):

(146) a. Sandra – d Sandra 'Sandra'
b. Onkel Max – de Onkel Max 'Uncle Max'
c. Herr Professor Müller – de Herr Profässer Müller 'Professor Müller'

In the Berner Oberland and parts of Graubünden, definite articles are not traditionally used (SDS III/141). However, more recent surveys show that definite articles are now also accepted there, and sometimes even preferred (SADS II/31-32). Furthermore, in BE-Emmental and parts of BE-Haslital, there is a special accusative/dative morpheme *-en* for masculine names (Bucheli-Berger 2006):

(147) a. Anna – d Anna / Anna
b. Fritz – de Fritz / Fritz / Fritzen (Acc)

The WS-Lötschental dialect uses the accusative/dative suffix *-en* for feminine names, but the determiner *in* without suffix for masculine names (SADS II/31, 32):

(148) a. d Annen
b. in Fritz

In most Swiss German dialects (except in the Northeast), neuter pronouns are used to refer to women's names. In Wallis dialects, neuter pronouns are also used for men's names (SADS IV/38):

(149) a. …Trudi …Es hät hüt Geburtstag. (W) / Si hät hüt Geburtstag. (O)
'Today is her (Trudi's) birthday.'
b. …Paul …Er hät hüt Geburtstag. / Es hät hüt Geburtstag. (WS)
'Today is his (Paul's) birthday.'

The correct implementation of the phenomena discussed here would require a Named Entity Recognition system that is able to distinguish person names from other named entities. We have not investigated this field yet.

Indefinite articles before occupational titles and names of origin

Lötscher (1983, p. 105) argues that indefinite articles are required before occupational titles and names of origin. He illustrates this claim with the following examples:

(150) a. Er ist Zürcher. – Er isch en Zürcher. 'He is from Zurich.'
b. Er ist Arzt. – Er isch en Tokter. 'He is a doctor.'

We do not support Lötscher's categorical claim; the forms with and without articles sound similarly acceptable to us. This might be due to influence from Standard German, or to inter-dialectal differences. Unfortunately, there is no SADS data dealing with this phenomenon. In any case, we do not find the evidence and the frequency of this phenomenon compelling enough to implement it.

Adverbs and articles – word order and article doubling

If an adjective is modified by an intensity adverb, the article precedes the entire adverb-adjective complex in Standard German. This variant is only marginally used in Swiss German (151a). Instead, the article is more often placed between the adverb and the adjective (151b), or the article is doubled and appears in both positions (151c):

(151) a. ein ganz liebes Kind – *es* ganz liebs Chind 'a very nice child'
b. ganz *es* liebs Chind

c. *es* ganz *es* liebs Chind

This phenomenon only occurs with intensity adverbs like *so 'such', vil 'much', ganz 'very', echli 'a bit', rächt 'quite'* (Lötscher 1983, p. 105). Other types of adverbs are located in the same position as in Standard German. Concerning the geographical distribution, variant (151b) is used in all Swiss German dialects, while variant (151c) is preferred only in BA, LU, SZ, GL, SGO (SADS I/10, IV/1). Steiner (2006) notes that article doubling is more often preferred by younger informants than by older ones, suggesting that this Swiss German innovation may be spreading further in the future.

The situation is more complex if comparative forms of adjectives are involved. In this case, Swiss German tends to use definite articles (instead of indefinites in Standard German), again with or without article movement and doubling (SADS II/10):

(152) a. ein viel besserer Koch – *de* vil besser Choch 'a much better cook'
b. vil *de* besser Choch
c. *de* vil *de* besser Choch

There is no geographical distinction between the variants: (152b) is preferred everywhere, (152a) is accepted everywhere, and (152c) is rarely accepted (Steiner 2006).

A.4.3. Nouns

Inflection

The inflectional paradigm of Swiss German nouns is reduced in comparison with Standard German. In most dialects, there is only one singular and one plural form. Case is not marked in the nouns except for some relics discussed on page 348. In the following discussion, we adopt the five inflection classes of Lötscher (1983, pp. 89-93).

Group 1: *Umlaut* This group contains masculine and feminine nouns that form the Standard German plural with *umlaut*, or *umlaut* + *e*. It is the default plural form for inanimate masculines

and notably also contains some words that would use *e*-plural (without *umlaut*) in Standard German (153f)-(153g).

(153)
a. Gast / Gäste (M) – Gascht / Gäscht 'guest'
b. Vogel / Vögel (M) – Vogel / Vögel 'bird'
c. Schaden / Schäden (M) – Schade / Schäde 'damage'
d. Wand / Wände (F) – Wand / Wänd 'wall'
e. Faust / Fäuste (F) – Fuuscht / Füüscht 'fist'
f. Hund / Hunde (M) – Hund / Hünd 'dog'
g. Apparat / Apparate (M) – Apperaat / Apperäät 'device'

Group 2: *e*-plural The Swiss German *e*-plural corresponds to the Standard German *en*-plural. This class contains mainly animate masculine nouns, as well as many feminine nouns, and the two neuter nouns *Auge* 'eye' and *Ohr* 'ear'.

(154)
a. Bär / Bären (M) – Bäär / Bääre 'bear'
b. Sache / Sachen (F) – Sach / Sache 'item'
c. Druckerei / Druckereien (F) – Truckerei / Truckereie 'printing press'

In Western dialects, this class also contains monosyllabic masculine nouns that have a *e*-plural, and some masculine nouns ending in *-el* in Standard German (all these words have a *null*-plural in other Swiss German dialects; see SDS III/164-166):

(155)
a. Fisch / Fische (M) – Fisch / Fische (BE, SO, FR, WS) 'fish'
b. Stein / Steine (M) – Schtei / Schteine (BE, SO, FR, WS) 'stone'
c. Esel / Esel (M) – Esu / Esle (BE, SO, FR, WS) 'donkey'

Note that drop of final *n* and *l*-vocalization are applied to the inflected forms.

Group 3: *Umlaut* + *er* This group contains mostly neuter nouns, but also some masculines. Its distribution is roughly the same as in Standard German, but it contains additional neuter nouns (156e). Note that not all words in this class can form an *umlaut* for phonotactic reasons.

(156) a. Buch / Bücher (N) – Buech / Büecher 'book'
b. Huhn / Hühner (N) – Huen / Hüener 'chicken'
c. Licht / Lichter (N) – Liecht / Liechter 'light'
d. Wald / Wälder (M) – Waald / Wäälder 'forest'
e. Brot / Brote (N) – Broot / Brööter 'bread'

Group 4: *null*-plural The words in this group have the same form in singular and plural. It contains several subgroups.

First, it contains masculine and neuter nouns with *e*-plural in Standard German that cannot build an *umlaut* (with the Western dialect exceptions stated above):

(157) a. Fisch / Fische (M) – Fisch / Fisch (except W) 'fish'
b. Netz / Netze (N) – Netz / Netz (except W) 'net'

Second, there are masculine and neuter nouns which already have *null*-plural in Standard German. These words end in an unstressed syllable that resembles one of the plural suffixes (*Pseudosuffix*).

(158) a. Jäger / Jäger (M) – Jäger / Jäger 'hunter'
b. Gewitter / Gewitter (N) – Gwitter / Gwitter 'thunderstorm'

Third, *null*-plural is often used in foreign nouns where Standard German would use *s*-plural. Traditionally, there is no *s*-plural in Swiss German, but it is becoming common lately (see below).

(159) a. Akku / Akkus (M) – Akku / Akku 'accumulator'
b. Büro / Büros (N) – Büro / Büro 'office'

Fourth, feminine nouns ending in *-e/-en* in Standard German have *-e/-e* in many Swiss German dialects:

(160) Nase / Nasen (F) – Nase / Nase 'nose'

These feminine nouns are the only ones that lack any plural morpheme; indeed, not even the

definite article allows to distinguish singular from plural forms: *d Nase / d Nase*. This ambiguity leads to a "morphological state of emergency" (Wolfensberger 1967, p. 108), to which different dialects have found different solutions (SDS III/186).

- Traditional Zürich German omits the final *schwa* in the singular: *Tann / Tanne* 'pine'. However, this former innovation is threatened by the pressure from surrounding dialects which have kept the homophonous forms.

- Freiburg dialect has *-a* in the singular and *-e* in the plural: *Tanna / Tanne*. This inflection pattern has some resemblance with neighboring Romance dialects.

- Berner Oberland dialects have *-e* in the singular and *-i* in the plural: *Tanne / Tanni*.

- A more recent solution (Christen 1998a), for which no geographical data is available,[25] borrows the plural ending *-ene, -ine* from the *i*-feminines (see Group 5 below): *Tanne / Tannene*.

Finally, diminutives show a *null*-plural in most dialects, but show *-eni* in some Southwestern dialects (SDS III/180-181):

(161) Äuglein – Öugli, Öugleni (BEO, FR), Öuglini (WS) 'little eye'

Group 5: Special plural marks In most dialects, kinship terms in *-er* form the plural with *umlaut* and *-e* (SDS III/170):

(162) a. Bruder / Brüder – Brüeder / Brüeder(e) 'brother'
b. Vatter / Väter – Vatter / Vätter(e) 'father'
c. Mutter / Mütter – Muetter / Müettere 'mother'
d. Schwester / Schwestern – Schwöschter / Schwöschtere 'sister'

Feminine nouns ending in *-i* form the plural with *-ene* in most dialects, but with *-ine* in Bern dialect (SDS III/187):

25 The relevant SDS maps do not show any evidence of this suffix.

(163) a. Mühle / Mühlen – Müli / Mülene, Müline (BE) 'mill'
b. Höhe / Höhen – Hööchi / Hööchene, Höchine (BE) 'height'

_s_-plural Lötscher (1983, p. 91) maintains that "im Schweizerdeutschen unbekannt ist der Plural auf *-s* wie in *Autos, Sofas*".[26] We do not agree with this categorical statement. The *s*-plural may be a relatively recent grammatical borrowing from Standard German, but it is definitively used for foreign words in present-day Swiss German:

(164) a. Cowboy / Cowboys – Koboy / Koboys 'cowboy'
b. Tipp / Tipps – Tipp / Tipps 'hint'
c. Ami / Amis – Ami / Amis 'American'

For each Standard German plural marking, a default Swiss German plural marking will be defined. Words that deviate from this default mapping will be listed individually as exceptions.

Relics of case marking

Some Swiss German dialects have maintained specific noun endings for dative plural, genitive singular and genitive plural. ***Dative plural*** forms (SDS III/172, 179, 190) traditionally end in *-e* (*-u/-n* in Wallis dialects), where this is not already part of the plural ending (Rash 1998, p. 140):

(165) a. Büchern (Dat) – Büechere 'books'
b. Kühen (Dat) – Chüene 'cows'
c. Schlüsseln (Dat) – Schlüssle 'keys'
d. Äuglein (Dat) – Öuglene < Nom. Öugleni (BEO, FR) 'little eye'

However, these forms are becoming obsolete in Zürich German (Baur 1983, p. 45; Wolfensberger 1967, p. 111) as well as in other dialects; we will neglect them.

Relics of ***genitive case*** can still be used with persons and in fixed idioms:

26 "The plural in *-s*, as in [Standard German] *Autos, Sofas*, is unknown in Swiss German."

(166) a. Rudis Haus – s Ruedis Huus 'Rudi's house'
b. ?in (des) Pfarrers Garten – is Pfarrers Gaarte 'in the reverend's house'
c. um Gottes Willen – um s Gotts Wille 'for God's sake'

However, these forms are considered obsolete in many parts of Switzerland, and are used regularly only in alpine dialects (WS, BEO, GL, GR) (SADS II/22, 23, 30).[27] Bohnenberger (1913) presents entire noun and pronoun paradigms with genitive forms for the conservative Wallis and GR-Walser dialects:

(167) a. des Tages (Gen Sg) – ds Tagsch (WS) 'of the day'
b. der Tage (Gen Pl) – dr Tago (WS) 'of the days'

Further field work would be required to determine how productive these forms are in present-day dialect. For the time being, we do not consider them in our system.

Paraphrase of Standard German genitive case

To compensate for the loss of inflectional genitive forms, Swiss German dialects use two main patterns of paraphrase. One possibility is to express possessive attributes by a prepositional phrase, analogous to Romance languages, or to English *of*. The preposition used is *vo*, which governs the dative (Lötscher 1983, pp. 92-93):

(168) a. das Dach des Hauses – s Tach vom Huus 'the roof of the house'
b. das Resultat dieser Berechnung – s Resultat vo dere Rächnig 'the result of this calculation'

A better-known alternative is the so-called **possessive dative**, although it can only be used in restricted contexts: the head noun phrase must be definite, and the possessor must be a person.

(169) a. Rudis Vater – em Ruedi sin Vatter 'Rudi's father'
b. das Haus meines Bruders – mim Brüeder sis Huus 'my brother's house'

27 See also Russ 1990, and http://www.walser-alps.eu/mundart/das-walliserdeutsche, accessed 24.1.2010.

c. das Dach des Hauses – *em Huus sis Tach

I have been unable to find a detailed syntactic description of this phenomenon. In my work, I analyze the possessive dative construction as follows:

- The entire noun phrase takes the same structure as a simple noun phrase with a possessive determiner, of the type *mis Buech* 'my book'.
- Possessive determiners of the third persons are allowed to take a noun phrase argument. This argument obligatorily takes dative case and represents the possessor.
- The same holds for possessive pronouns, which accounts for sentences of the type *Das isch mim Vater sis* 'this is my father's'.

The base form of the possessive determiner depends on the possessor: (170a) vs. (170b); it agrees with the possessee in gender and number: (170a) vs. (170c); and it marks the case of the entire noun phrase: (170a) vs. (170d).

(170) a. em Ruedi (M) *sin* Vatter (M-Nom-Acc)
b. de Sandra (F) *iren* Vatter (M-Nom-Acc)
c. em Ruedi (M) *sis* Huus (N-Nom-Acc)
d. em Ruedi (M) *sim* Vatter (M-Dat)

In the areas of prepositional dative marking (see page 329), the dummy dative prepositions are also used in these constructions:

(171) a. am / im Ruedi sin Vatter
b. a / i mim Brüeder sis Huus

Note that the possessive dative forms can also be used together with interrogative pronouns (Bucheli and Glaser 2002, p. 48):

(172) Wessen Haus ist abgebrannt?
– Wem sis Huus isch abeprännt? 'Whose house has burnt down?'

We will implement the prepositional construction as a default rule, and the possessive dative as a secondary rule that fires only if the cited syntactic and semantic constraints are fulfilled. It remains to be shown how the latter constraint can be implemented effectively.

A.4.4. Adjectives

Inflection

As in Standard German, adjectives agree in case, number and gender with the head noun. Moreover, the inflection depends on the type of the determiner: a definite determiner triggers the so-called weak inflection, an indefinite determiner triggers the strong inflection series:[28]

(173)

	Nom/Acc Weak	*Nom/Acc Strong*	*Dative*
Singular Masculine	guet (O), gueti (W)	guete	guete
Singular Feminine	guet (O), gueti (W)	gueti	guete
Singular Neuter	guet (O), gueti (W)	guet (N), guets	guete
Plural	guete	guet (S), gueti	guete

For the weak singular forms, the Eastern dialects traditionally show *null*-ending and the Western dialects *i*-ending. However, the latter is currently spreading all over Switzerland (SDS III/254; Christen 1998a, pp. 57-58; Lötscher 1983, p. 95). The strong neuter forms traditionally have a *null*-ending in Northwestern Switzerland, but the *s*-form is gaining ground (SDS III/252). The *null*-ending in strong plural forms in alpine dialects is being replaced by the *i*-form (SDS III/253). All these evolutions tend to eliminate the *null*-ending from the adjective inflection altogether, and to reserve it for the uninflected base form.

28 For comparison, the Standard German forms are below:

(i)

	Nom/Acc Weak	*Nom/Acc Strong*	*Dative*
Singular Masculine	gute (Nom), guten (Acc)	guter (Nom), guten (Acc)	guten
Singular Feminine	gute	gute	guten
Singular Neuter	gute	gutes	guten
Plural	guten	gute	guten

Inflection of predicative and copredicative adjectives

The inflected forms of Table (173) are used in attributive contexts, i.e. inside a complete noun phrase. In predicative, copredicative and adverbial contexts, Standard German uses the base form of the adjective (*gut* in our example). Most Swiss German dialects do so as well (*guet*), but some dialects use inflected forms of the strong series in one or another context.

In the Wallis and Freiburg dialects, adjectives in **predicative** position (i.e. as complements to verbs like *syy* 'to be' or *wärde/cho* 'to become') agree with the subject in gender and number, as do past participles in passive constructions:[29]

(174) a. er / sie / es ist alt – är / si / es isch altä / alti / alts 'he/she/it is old'
b. er / sie / es wird alt – är / si / es chunt altä / alti / alts 'he/she/it gets old'

(175) a. Das Haus ist verkauft worden. – S Hus isch verchaufts cho. 'The house has been sold.'
b. Das Haus ist verkauft. – S Hus isch verchaufts. 'The house is sold.'

Geographical data are available in SDS III/256 and SADS I/5, III/16. A comparison shows that the former agreement areas in Central Switzerland, Bern and Glarus now mostly use uninflected forms (Bucheli-Berger 2006).

A slightly different problem is the agreement of **copredicative** adjectives. According to Bucheli-Berger (2006, 92, footnote 2), a copredicative adjective relates as an attribute to a noun phrase, but also to the predicate of the sentence (see examples below). In contrast to (obligatory) predicative adjectives, copredicative adjectives are always optional (Berger and Glaser 2004). In alpine dialects (WS, FR, BEO, UW, GR), the copredicative adjective agrees in gender and number. In some Northeastern dialects (AP, SG), there is an invariable *e*-ending for all genders and numbers, which from a diachronic point of view corresponds to the frozen strong masculine singular form (SDS III/256; SADS I/12, I/17, II/13, II/17).

29 On the use of the verb *cho* as a passive auxiliary, see page 366. Both phenomena (adjective agreement and auxiliary verb type) have been attributed to the influence of neighboring Romance languages (Seiler 2004).

(176) Sie sollten die Milch / den Tee / das Bier warm trinken.[30]
- Sie sötted d Milch / de Thee / ds Bier warmi / warme / warms trinke. (WS, …)
- Sie sötted d Milch / de Thee / s Bier warme trinke. (AP, SG)

'You should drink the milk / tea / beer warm.'

Comparison forms

Like in Standard German, Swiss German adjectives have synthetic comparison forms with the suffixes *-er* for the comparative, and *-scht* or *-ischt* for the superlative (Lötscher 1983, pp. 95-96):

(177) a. lieb / lieber / am liebsten – lieb / lieber / am liebschte 'dear'
b. frisch / frischer / am frischesten – frisch / frischer / am frischischte 'fresh'

The stem vowel undergoes mutation (*umlaut*) more frequently than in Standard German. *Umlaut* is avoided only in adjectives with a transparent derivative origin.

(178) a. mager / magerer / am magersten
- mager / mägerer / am mägerschte 'skinny'

b. trocken / trockener / am trockensten
- troche / tröchner / am tröchnischte 'dry'

The examples above show the uninflected base forms; as in Standard German, comparatives and superlatives are inflected with the same endings as presented in Table (173) for positive forms.

Comparative particles

Comparison is usually expressed with an adjective in comparative degree and a base noun to which it relates. This base noun is attached with a comparative particle like *als* 'than' in Standard German. Swiss German uses the following variants (SADS III/22, 25, 28; Friedli 2006):

30 Warm beer is said to help against a cold.

(179) a. grösser *als* du
b. grösser *weder* du
c. grösser *wie* du (TG, SH)
d. grösser *wa* du (BEO, WS)
e. grösser *as* du 'taller than you'

The *als*-variant is preferred in nearly all regions; *weder* is used in all parts of Switzerland, but rarely preferred; *wie* and *wa* are used only in a small, determined area; *as* is a phonetic variant of *als* for which there is unfortunately no geographical data.

The comparison base (i.e., the element after the particle) is not necessarily a noun phrase. For instance, it can also be an infinitival phrase (180) or a subordinate clause (181). The distribution of the different particles seems to be quite independent of syntactic context. The only restriction applies to *weder*, which is very rarely used with subordinate clauses.

(180) Sie gehen lieber schwimmen *als* laufen.
– Si gönd lieber go schwimme *als/weder/wie/wa/as* go laufe.
'They prefer swimming to walking.'

(181) Er ist älter, *als* ich gedacht habe.
– Er isch älter, *als/wien/wan/as* ich tänkt han. 'He is older than I thought.'

A.4.5. Prepositions

In the preceding sections, we have already mentioned some problems related to prepositions. Prepositions and determiners can be contracted (page 327). The prepositions *a* and *i* are used as dative markers in some regions (page 329). In this section, we discuss one more feature, the selection of place-name prepositions.

Place-name prepositions

There is a small, but significant difference in the choice of prepositions in conjunction with place names. Standard German uses *nach* to convey direction, and *in* to convey place. In Swiss German, *uf* and *z* (formal equivalents of Standard German *auf* 'on' and *zu* 'to') are used respectively (Haas 2000, p. 97):

(182) a. nach Zürich (*auf Zürich) – uf Züri 'to Zurich'
b. in Basel (?zu Basel) – z Basel 'in Basel'

In spite of this general rule, the prepositions *nach* and *in* are increasingly being used in Swiss German and have a connotation of higher stylistic level (Wolfensberger 1967). We generate both variants, with a general preference for the traditional ones.

A.4.6. Verbs

This section starts with an outline of the inflectional paradigm of Standard German verbs and of the differences to be expected in Swiss German. We then detail and illustrate the inflectional paradigm of regular verbs and discuss some interdialectal differences. Later, different types of irregular verbs will be presented, and dialect-specific issues of morphology will be treated. We will then turn to periphrastic verb forms, and end with word order issues arising from complex verb phrases.

The inflectional paradigm in Standard German and in Swiss German

The Standard German verbal inflection consists of the following forms:

- the infinitive,
- the past participle (declined like an adjective when used attributively),
- the present participle (declined like an adjective when used attributively),

- six personal forms of present indicative tense (in regular verbs, the first and third person plural forms are identical, and the third person singular and the second person plural forms are identical),
- six personal forms of subjunctive I (in regular verbs, only the third person singular is distinct from the respective indicative form),
- six personal forms of preterit indicative tense (in regular verbs, the first and third person singular forms are identical, and the first and third person plural are identical),
- six personal forms of subjunctive II (in regular verbs, these forms are identical to the preterit indicative forms),
- two imperative forms.

Subjunctive usage is rather complicated. To mark indirect speech, subjunctive I (subjunctive present) is normally used. However, if its form is identical to indicative mood, and the speaker insists on conveying subjunctive mood, he can resort to the respective subjunctive II form (subjunctive past/conditional). If this form is still ambiguous, periphrastic forms with the auxiliary verb *werden* are used. To mark unreal or conditional statements, subjunctive II is normally used, with periphrastic forms used as backoff.

The most striking feature of Swiss German dialects is the absence of the preterit indicative paradigm; past tense is exclusively expressed periphrastically with auxiliary verbs and a past participle. Furthermore, the subjunctive I forms tend to be more distinct from present indicative forms, reducing the need for backoff solutions. Synthetic subjunctive II forms exist for regular and irregular verbs, but their current use is limited to the most frequent verbs. Finally, the present participle, already rare in Standard German, does not traditionally exist in Swiss German dialects, and is only used as a loan-translation. In other regards, Swiss German verb inflection follows the Standard German paradigms.

Regular verbs

In most dialects, the **infinitive** ends in *-e* [ə]. As with other final-*schwa* contexts (see pages 301 and 311), exceptions occur in BE-Haslital (*-en*) and in other alpine regions (*-ä, -a*; SDS III/1). In some Wallis dialects, the infinitive ending depends on the verb class (see page 363).

The Standard German **past participle** prefix *ge-* undergoes massive phonetic reduction. Before vowels (183a), semi-vowels (183b), nasal, liquid and affricate consonants (183c), the prefix is reduced to *g-* [k]. Before lenis plosives, the prefix is assimilated to the initial consonant, changing it to a fortis (183d). Before fortis plosives, it is assimilated to the initial consonant without leaving a trace in spelling (183e).

(183) a. geändert – gänderet 'changed'
b. gewettet – gwettet 'bet'
c. gemacht – gmacht 'made'
d. gedient – tienet (< diene) 'served'
e. getauscht – tuuschet (< tuusche) 'swapped'

In the Lower part of Wallis, the vowel is maintained in the prefix under the form *gi-* before plosives, and *ga-* before *r*. In the Upper part of Wallis, only the second part of the previous rule applies. Some GR-Walser dialects use *ga-* before plosives (SDS III/3; Bohnenberger 1913, §82):

(184) a. getrunken – gitrüüchu (WS-Lower), trüüchu (WS-Goms), gatruuche (GR-Walser) 'drunk'
b. gerechnet – garechnot (WS), grechnet (GR-Walser) 'calculated'

In Standard German, verbs with unseparable prefixes and foreign verbs ending in *-ieren* do not use the *ge*-prefix. This rule is less strict in Swiss German, and prefixed forms can be found depending on the lexeme and on the dialect.[31] To our knowledge, this phenomenon has not been studied in more detail.

31 Lötscher (1983, p. 97) claims that *ieren*-verbs require the prefix. We do not fully support his claim. For certain words, the presence or absence of the participle prefix even allows semantic differenciation:

(i) si hät schtudiert 'she has studied (at University)'

In addition to the prefix, past participle forms also have a distinct suffix. Strong (i.e. irregular) verbs have a *-e* suffix (or similar), while weak (i.e. regular) verbs have *-t* or *-et*. In Standard German, *-et* occurs always and only after *t, d, l, r, m, n, g*. In Swiss German, the situation is less clear. Verbs derived from a noun or containing another derivational suffix tend to use *-et*; some verbs ending in *d, t* use *-t*: *aazündt* < *aazünde* 'to ignite' (Lötscher 1983, pp. 97-98; Russ 1990, pp. 376-377).

Let us illustrate the **present indicative** paradigm with the regular verb *sueche* 'to seek':[32]

(185)

Person	*Singular*	*Plural*
1	sueche, suech	sueched (O), sueche (W), süeche (WS)
2	suechsch, suechscht	sueched (O), suechet (W), süechet (WS)
3	suecht	sueched (O), sueche (W), süechend (WS)

Christen (1998b) and Wolfensberger (1967) note some tendencies of ongoing language change. The first person singular is increasingly used without final *-e*, especially in Eastern dialects (SDS III/22-25). In the second person singular, the forms with final *-t* are still used in some Eastern and Central dialects (SDS III/26-27), but are being replaced by the *t*-less variant.

The plural endings mentioned in table (185) present the three inflection models in use in Swiss dialects (SDS III/31-34; Lötscher 1983, pp. 160-161):

- In Eastern dialects, all three persons have the same verb ending. The ending is *-ed* in Zurich and the Northeast, but *-id* in Central Switzerland and Appenzell, and *-end* in Graubünden.

(ii) si hät gschtudiert 'she has pondered'

32 The Standard German inflection paradigm is presented below:

(i)

Person	*Singular*	*Plural*
1	suche	suchen
2	suchst	sucht
3	sucht	suchen

- The Western two-form model closely follows the Standard German model with a common form for the first and third persons (*-e*) and a distinct form for the second person (*-et*). In the Basel region, the second person is *-ed*, and in BE-Haslital the first and third person suffix keeps the final *-n*.

- The Wallis dialects show three different plural forms. Hence, they make a formal distinction which is not made in Standard German (between the first and third person plural).

The **imperative** singular form is endingless (*suech!*), while the plural form is identical to the second person plural present tense (*suechet!*).

The **subjunctive I** forms are characterized by an *i* in all forms, except plural forms in *-e* in some Western dialects (Lötscher 1983, p. 98; Marti 1985, p. 143):[33]

(186)

Person	*Singular*	*Plural*
1	suechi	suechi, -e (W), suechid (O)
2	suechisch	suechid, -ed
3	suechi	suechi, -ed (W), suechid (O)

The regular **subjunctive II** forms are falling out of use and are replaced by periphrastic forms (see page 367). However, they can be constructed easily with the *-t(i)* suffix (Lötscher 1983, p. 98; Marti 1985, p. 145):[34]

33 The Standard German subjunctive I forms are listed below. Forms in italics are identical with the corresponding indicative forms.

(i)

Person	*Singular*	*Plural*
1	*suche*	*suchen*
2	suchest	suchet
3	suche	*suchen*

34 Standard German forms for comparison. All subjunctive II forms of this regular verb are identical with the preterit forms.

(187)

Person	Singular	Plural
1	suechti	suechti, -e (W), suechted (O)
2	suechtisch	suechtid, -ed (W), suechted (O)
3	suechti	suechti, -e (W), suechted (O)

Irregular verbs

In this section, we define several classes of irregular verbs and present some of their characteristics.

Strong verbs are lexical verbs which use some kind of vowel gradation pattern (*ablaut*) (188a), (188b) or simply have an irregular past participle (188c):

(188) a. singen / singt / sang / gesungen – singe / singt / — / gsunge 'to sing'
b. helfen / hilft / half / geholfen – hälfe / hilft / — / ghulfe 'to help'
c. bringen / bringt / brachte / gebracht
– bringe / bringt / — / proocht, prunge (WS) 'to bring'

Those strong verbs which show an alternation between *e* and *i* in the second and third persons singular extend the *i*-stem to the first person singular in many dialects (Russ 1990, p. 377):

(189) ich h*e*lfe / du hilfst / er hilft – i h*i*lf(e) / du hilfsch / er hilft

Short verbs are defined as verbs with a monosyllabic infinitive form (Nübling 1995a; Nübling 1995b). While Standard German only has two of them (*sein* 'to be' and *tun* 'to do'), there are many more in Swiss German:

(190) a. sein – syy (SDS III/40, 42, 50-51) 'to be'
b. haben – haa (SDS III/40, 43, 46-49) 'to have'

(i)

Person	Singular	Plural
1	suchte	suchten
2	suchtest	suchtet
3	suchte	suchten

c.	tun – tue	(SDS III/52-55)	'to do'
d.	gehen – gaa, goo	(SDS III/41, 56-58)	'to go'
e.	kommen – choo	(SDS III/100-102)	'to come'
f.	lassen – laa, loo	(SDS III/64-68, 120)	'to let'
g.	stehen – schtaa, schtoo	(SDS III/59, 119)	'to stand'
h.	sehen – gsee	(SDS III/96-98)	'to see'
i.	geben – gää, gee	(SDS III/90, 94)	'to give'
j.	nehmen – nää, nee	(SDS III/91-94, 116)	'to take'
k.	schlagen – schlaa, schloo	(SDS III/70-72, 121)	'to hit'
l.	ziehen – zie		'to pull'
m.	(an)fangen – (a)faa, (a)foo	(SDS III/74-78)	'to catch, to begin'

The preceding list of short verbs is complete, except for particle verbs like *aachoo* 'to arrive', *beschloo* 'to grow damp'. Note also that not all of these verbs are short in all dialects. For instance, Eastern dialects have bisyllabic *fange* instead of monosyllabic *foo*, and *züche* instead of *zie*.

Short verbs have some particularities in their inflection pattern:

- In Eastern dialects, short verbs show a present plural ending in *-nd*, unlike the regular verbs which end in *-ed*. The inflection pattern is fairly regular in the two-form and three-form (Bohnenberger 1913, 251ff.) dialects:

(191) a. wir sind / ihr seid / sie sind
   - mer / ir / si sind (O)
   - mer sy / dir syt / si sy (W)
   - mer sy / dir syt / schi sint (WS)

b. wir haben / ihr habt / sie haben
   - mer / ir / si händ (O)
   - mer hei / dir heit / si hei (W)
   - mer hei / ir heit / schi heint (WS)

c. wir gehen / ihr geht / sie gehen
   - mer / ir / si gönd (O)
   - mer gö / dir göt / si gö (W)
   - mer gee / ir geet / schi geent (WS)

- The present plural forms are formed with *umlaut* if phonetically possible:

(192) a. er geht / sie gehen – er goot / si gönd 'he goes / they go'
b. er tut / sie tun – er tuet / si tüend 'he does / they do'

- The subjunctive I and subjunctive II forms are unambiguous and also formed with *umlaut*.

Overall, short verbs are characterized by very irregular paradigms and high variability between different dialects. Their shortness also makes them prone to ambiguity: any short form starting with *g* is either of the verb 'to go' or 'to give'. However, given the limited number of short verbs, the context is usually sufficient to disambiguate these cases. Some short verbs (*syy, haa, tue*) are used as auxiliary verbs (see below).

Modal verbs are auxiliary verbs that indicate modality. Their inflection historically follows the preterit paradigm, which accounts for the identical first and third person singular forms (*ich kann, er kann* 'I can, he can'. In Standard German, one considers six verbs as modal verbs; the same list is used for Swiss German dialects (Russ 1990, p. 378):

(193)	a.	dürfen – töörfe, tööre	(SDS III/108, 109)	'to be allowed to'
	b.	können – chöne	(SDS III/104, 105, 109)	'to be able to'
	c.	mögen – möge	(SDS III/106, 107)	'may'
	d.	müssen – müese	(SDS III/86, 87)	'to have to'
	e.	sollen – söle		'ought to'
	f.	wollen – wöle	(SDS III/112-114)	'to want to'

Like short verbs, modal verbs are characterized by considerable interdialectal variation. They have distinct and widely used subjunctive II forms. Some modal verbs show an Eastern *nd*-plural, while others use the regular plural endings:

(194) a. wir wollen – mer wönd (O) 'we want to'
b. wir sollen – mer söled (O) 'we ought to'

In Swiss German, the past participles of the modal verbs are always identical to the infinitives. Contrarily to Standard German, it does not matter if the modal is used together with a lexical verb or not (Glaser 2003).

Verb classes in Highest Alemannic dialects

The Old High German verb classes are still effective in some Wallis dialects (Hotzenköcherle 1984, pp. 161-163; Lötscher 1983, p. 152).

The Lötschental dialect has the most complicated system, with three different inflection paradigms depending on the verb class. The infinitive and third person plural forms look as follows:

(195)

	Irregular	***OHG** -jan*	***OHG** -ôn*	***OHG** -ên*
Infinitive	schniidn	zeln	choru	losä
3rd plural	schniidnd	zelnd	chorund	losänd
	'to cut'	'to count'	'to choose'	'to listen'

In Central Wallis, there are only two classes: OHG *-ên* verbs have *-ä/-änd*, while all other verbs have *-u/-und*. In the westernmost part of German-speaking Wallis, all verbs have the *-u/-und* endings. Finally, the Eastern WS-Goms dialects are most similar to other Swiss German dialects and use *-e/-end* for all verbs.

An implementation of the Lötschental and Central Wallis systems would require all verbs to be annotated with their Old High German class. Although such verb lists may be available, we do not deem it worthwhile to implement these systems. Recall also that the situation described above draws on SDS data from the 1940s-1950s, and that further simplification may have taken place since. Our implementation will contain the *u*-system of Western Wallis and the *e*-system of Eastern Wallis.

Paraphrase of the preterit and pluperfect tense

As mentioned above, preterit tense has fallen out of use in most parts of Switzerland during the 16th century. Traces could still be found in the BE-Oberland dialects at the end of the 19th

century: *i was* 'I was',[35] *hatti* 'I had', *cham* 'I came' (Jörg 1976). Lötscher sees the loss of preterit tense as a consequence of the mainly spoken use of Swiss German:

> Auch der Verlust der einfachen Vergangenheitsformen des Verbs wird auf die rein mündliche Verwendung des Schweizerdeutschen zurückgeführt. Das zusammengesetzte Perfekt *es hat geschneit* ist ja noch heute die Form des lebhaften Erzählens, des spontanen Erlebens eines vergangenen Ereignisses aus der unmittelbaren Gegenwart. Die einfache Vergangenheit *es schneite* ist demgegenüber die Form, die im distanzierten Bericht über vergangenes Geschehen ohne direkten Bezug zur gegenwärtigen Situation verwendet wird. Diese Erzählweise ist viel eher jene des schriftlichen Berichts im Protokoll oder der objektivierenden Darstellung eines Geschichtsbuchs. Da man im mündlichen Gespräch viel eher um möglichst lebhaftes, unmittelbares Erzählen bemüht ist als um distanzierenden Bericht, ist es verständlich, daß so allmählich die distanzierende einfache Vergangenheitsform außer Gebrauch kam.[36] (Lötscher 1983, pp. 117-118)

As a consequence of the loss of preterit tense, Swiss German dialects use the (periphrastic) present perfect as the unique past tense. Hence, all occurrences of Standard German preterit tense have to be translated to present perfect (Lötscher 1983, pp. 96-97):

(196) a. Wir *aßen* und *gingen* dann ins Kino.
   – Mer *händ ggässe* und *sind* dänn is Kino *ggange*.
   'We ate and then went to the cinema.'

   b. Er *kam*, als wir gerade am Tisch *saßen*.
   – Er *isch cho*, wo mer grad am Tisch *gsässe sind*.
   'He arrived just when we were sitting at the table.'

35 It is remarkable that this Swiss German form has managed to stay much closer to modern English spelling than to the modern Standard German spelling *ich war*.

36 "The loss of the simple past forms of the verb is also attributed to the purely oral use of Swiss German. The compound perfect tense *it has snowed* still is the form of the lively narration, of the spontaneous experiencing of a past event from the immediate present. In contrast, the simple past *it snowed* is the form which is used in the detached report of past events without a direct relation to the current situation. This type of narration is rather the one of the written report in a transcript or the objectivized representation of a history book. Since in an oral conversation, one aims at the lively, immediate narration rather than at a detached report, it is understandable that the detaching simple past gradually fell into desuetude."

As usual in German syntax, the past participle takes its position in the right sentence bracket. In subordinate clauses, the whole verbal material is located in the right verb bracket. The ordering of auxiliary verb and past participle will be discussed on page 371.

With the absence of preterit tense, the traditional way of forming pluperfect tense also disappears. Instead, anteriority is expressed by the so-called double perfect:[37]

(197) Das Haus *war* schon *abgebrannt*, als die Feuerwehr kam.
– S Huus *isch* scho *abepränt gsii*, wo d Füürweer choo isch.
'The house had already burnt down when the fire brigade arrived.'

The past participle of the lexical verb is adjoined by the past participle of the corresponding auxiliary verb (*gsii* in the example). The main inflected verb is again the auxiliary verb (*isch*).

Standard German uses the auxiliary verbs *sein* and *haben* for perfective tenses. Their distribution partially depends on semantic criteria. For example, unaccusative verbs generally use *sein*, while transitive verbs (including reflexive verbs) use *haben*. Auxiliary verb usage is identical in Swiss German (bar a few exceptions, e.g. *sitzen, stehen, liegen* 'sit, stand, lie'), allowing us to draw on existing Standard German lexical data.

(198) a. Er *ist* gegangen. – Er *isch* ggange. 'He has gone.'
b. Er *hat* gegessen. – Er *hät* ggässe. 'He has eaten.'

Future tense

The future tense, expressed in Standard German with the auxiliary *werden* + infinitive, is not traditionally used in Swiss German (Baur 1983, p. 31). Present tense, sometimes combined with a temporal adverb, is sufficient (199). However, the *werden*-construction is widely used to convey an epistemic meaning of conjecture or speculation (200). Wolfensberger (1967) maintains that the epistemic reading is truly dialectal, while the temporal meaning has been borrowed recently from Standard German.

37 Double perfect, or *doppeltes Perfekt*, corresponds to the tense called *passé surcomposé* in French.

(199) a. Ich werde kommen. – I chume dänn. 'I will come.'
b. Ich werde morgen abreisen. – I reise morn ab. 'I am leaving tomorrow.'

(200) Er wird sich verfahren haben. – Er wird sech verfaare haa.
'He supposedly lost his way.'

There will not be any rules to deal with this phenomenon. Future tense is quite rare in Standard German, and its literal translation – although not traditional – is generally accepted in Swiss German.

Passive voice

Passive voice is constructed with the auxiliary verb *werden* and past participle in Standard German. This also holds for the Swiss German Plateau dialects. However, the alpine dialects of Graubünden, Wallis and Freiburg use the dialect variant of *kommen* (SDS III/266, SADS II/9). This divergence could be due to Romance substratum: the passive auxiliary is *venire* (or analogous) in Italian and Romansh (Seiler 2004).[38]

(201) Das Haus ist verkauft worden.
– S Hus isch verchauft *worde*.
– S Hus isch verchauft(s) *cho*. (GR, WS, FR) 'The house has been sold.'

Note that in WS and FR, the participle is inflected and agrees in gender and number with the subject (see page 352) – another feature of potential Romance influence.

In colloquial German, the so-called **dative passive** is rather popular: instead of the accusative object, the dative object becomes the (formal) subject of the passivized sentence. The auxiliary verb is *bekommen* or *kriegen* (202) (Zifonun et al. 1997, pp. 1824-1836).

38 Unfortunately, we could not find respective data from franco-provençal dialects, which would support this claim in the WS and FR regions. However, a glossary of Fribourg regional French seems to confirm this hypothesis with the following example (Grangier 1864, p. 211):

(i) cette maison viendra démolie, vendue, etc. 'this house will be demolished, sold, etc.'

(202) Er bekommt ein Buch geschenkt. 'He is given a book.'

This construction occurs only marginally in Swiss German (SADS III/10, III/26). Its absence might be due to lexical reasons: the lemma *kriegen* is restricted to GR dialects, whereas *bekommen* is translated as *übercho* in many dialects.

Periphrastic expression of conditional mood

As shown above, conditional mood can be expressed synthetically with all verbs using the subjunctive II forms. This possibility is widely used in the most frequent verbs, but is restricted to alpine regions (GR, WS) for regular verbs. Plateau dialects prefer the periphrastic expressions of conditional mood. Northeastern Switzerland uses the same pattern as in Standard German, i.e. *werden* in the subjunctive II form, plus the infinitive. In contrast, Western dialects prefer the auxiliary verb *tun* (SDS III/126; SADS IV/31):

(203) Dieser Hut würde ihr gefallen.
- Dä Huet *würd* ere gfalle. (ZH, TG, SG)
- Dä Huet *tät* ere gfalle. 'This hat would please her.'

Periphrastic indicative forms

Even present indicative verb forms can be expressed periphrastically, with the verb *tue* 'to do' + infinitive, in order to avoid awkward, infrequent, or complicated inflected forms (Lötscher 1983, pp. 107-109). Periphrastic forms are also used with intransitive imperatives, where the entire sentence would consist of only one (possibly monosyllabic) word. Moreover, semantic constraints on the verbs, for example those listed in Schönenberger and Penner (1995, pp. 318-319), seem to influence the acceptability of periphrastic forms. Wolfensberger (1967, p. 128) argues that periphrastic inflection is receding, but bases his claim on one test sentence only. We are not aware of more recent quantitative data about this phenomenon.

(204) a. ?er mälcht > er tuet mälche 'he is milking'
b. ?du verplämperlisch > du tuesch verplämperle 'you loiter away'

c. ?mer schpiled Klavier > mer tüend Klavier schpile 'we play piano'
d. ?Bätt! > Tue bätte! 'Pray!'

Gerund

Northeastern dialects (AP, SG-Rheintal; SDS III/1) have a special gerund form ending in *-id*. This form is used after the complementizer *z* 'to'. Other dialects use the infinitive in this context, as does Standard German.

(205) etwas zu essen – öppis z ess*e*, öppis z ess*id* (AP, SG) 'something to eat'

Continuous forms

The construction *sein* + *am* + infinitive is used as a periphrastic expression of continuous action. This construction exists in many spoken variants of German, but is not considered written standard.[39] In contrast, Swiss German dialects use this construction fairly frequently and even with objects. In that case, the particle *am* can be placed before the entire VP, between different objects, or immediately in front of the infinitive (Glaser 2003, p. 47). The different placements do not seem to depend on the dialect (SADS II/16, II/24, III/27). In Wallis dialects however, the construction *syy* + *dehinter* + z + infinitive is preferred.

(206) a. Er ist am Putzen.
– Er isch am butze.
– Er isch dehinter z butze. (WS) 'He is cleaning.'
b. ? Er ist am Schälen der Karotten (? Er ist die Karotten am schälen.)
– Er isch am d Rüebli rüschte.
– Er isch d Rüebli am rüschte.
– Er isch dehinter, d Rüebli z rüschte. (WS) 'He is peeling the carrots.'

39 In written Standard German grammar, the infinitive is considered to be nominalized. This prevents verb objects from being attached – they would have to be attached as genitive attributes, as in (206b). This interpretation also explains the upper case spelling of the infinitive.

Verb doubling

The verbs *goo* 'to go', *choo* 'to come', *laa* 'to let' and *aafaa* 'to begin' may govern an infinitival verb phrase. In this case, a short unstressed form of the corresponding verb (*go, cho, la, afa*) serves as a complementizer:

(207) a. Ich komme (um zu) essen. – I chume *cho* ässe. 'I come to eat.'
b. Ich gehe essen. – I goo *go* ässe. 'I go to eat.'
c. Er lässt die Uhr flicken. – Er laat d Uur *la* flicke. (W) 'He has his watch repaired.'
d. Ich fange an zu kochen. – I faa afa choche. (W) 'I begin to cook.'

Dialectal variation makes the situation more complex than stated above:

- In some regions, the verb *choo* 'to come' is combined with the complementizer *go* (instead of *cho*). We do not subscribe to Lötscher's thesis of semantic differenciation between the two variants (Lötscher 1983, p. 110); we merely see them as dialectal variants, as illustrated in SDS III/265.

- The particle *go* is used in all regions (SADS IV/5, 10). In Zürich dialect, it is sometimes doubled itself and yields *goge*: *Ich goo goge ässe.* Some Northeastern dialects use *gi*, along with the neighboring Alemannic dialects of Southern Germany and Vorarlberg (SDS III/265; Brandner and Salzmann 2009).

- The particles *la* and *afa* are only used in Western dialects (SDS III/263; SADS II/3, III/5, III/12, IV/30).[40]

- The verb *anfangen / aafaa* commands an infinitival clause with *zu* in Standard German, but a bare infinitival clause in Swiss German (see page 381). In regions with the *afa* particle, the latter can be considered as a replacement of Standard German *zu*.

40 In the latter case, the Eastern absence is probably linked to the fact that these dialects inflect *aafange* regularly, so that the form *aafaa* (which is at the origin of the phonetically reduced particle) never occurs at all in their paradigm.

The particles *go* and *cho* can be placed immediately before the main verb, or before the objects of the main verb. They can even be placed in both contexts at the same time (Brandner and Salzmann 2009):

(208) a. I goo e nöis Auto *go* chaufe.
b. I goo *go* e nöis Auto chaufe.
c. I goo *go* e nöis Auto *go* chaufe. 'I go to buy a new car.'

The doubling particle *la* represents additional material in present tense sentences, but replaces the past participle[41] in perfect tense sentences. Note also that the doubling particle precedes the infinitive of the lexical verb:

(209) Er hat die Uhr flicken *lassen*.
– Er het d Uur *la* flicke (*laa). (W). 'He had his watch repaired.'

According to the syntactic context, the doubling particle *afa* replaces the separable verb particle (which etymologically is *aa* in this case), the past participle, or serves as a separable particle in a context that usually doesn't allow separation:

(210) a. Ich fange *an* zu kochen.
– I faa *afa* choche. (W) 'I start cooking.'
b. Ich habe *angefangen* zu kochen.
– I ha *afa* choche. (W) 'I have started cooking.'
c. weil ich *anfange* (*fange an) zu kochen
– wil i faa *afa* choche. (W) 'because I start cooking.'

Ellipsis of movement verbs

If a movement verb is combined with a modal or auxiliary verb, and the sense of movement can be inferred from a directional adverb or a prepositional phrase, then the movement verb can be omitted (Lötscher 1983, p. 111):

41 Actually, the past participle of *lassen* is replaced by the infinitive in Standard German and non-doubling Swiss German dialects under the *Ersatzinfinitiv* rule (see page 380.

(211) a. Wir sind heimgegangen.
– Mer sind häi ggange. > Mer sind häi. 'We went home.'
b. Willst du ins Bett gehen?
– Wottsch is Bett goo? > Wottsch is Bett? 'Do you want to go to bed?'

Likewise, if a past participle of the verb *gehen* or *kommen* is used with an infinitival phrase, the corresponding doubling particle *go* or *cho* optionally replaces the past participle:

(212) Sie sind wählen gegangen.
– Si sind go wääle (ggange). 'They went to vote.'

Word order in modal and auxiliary verb clusters

Whenever two verbal elements cooccur in the right sentence bracket, the Standard German word order tends differ from the Swiss German word order (Lötscher 1978; Lötscher 1983, pp. 113-114; Haas 2000, pp. 97-98). In this section, we deal with syntactic configurations containing a modal or an auxiliary verb. Later on, the discussion will be extended to other verbs that subcategorize an infinitival complement. In particular, four configurations may trigger a reordering:

1. Compound tense main clauses, in which the left sentence bracket is occupied by the auxiliary verb, and the right sentence bracket by non-finite forms of a lexical and a modal verb.[42] In Standard German, the lexical verb precedes the modal, while the inverse order holds in most Swiss German dialects. In the absence of precise geographic information, we assume that inversion occurs in the same regions as in the second context.

   (213) Der Hund hat die Wurst nicht fressen sollen.
   – De Hund hät d Wuurscht nöd frässe söle. (GR, SGO)
   – De Hund hät d Wuurscht nöd söle frässe.
   'The dog shouldn't have eaten the sausage.'

2. Subordinate clauses, in which a finite modal verb and a non-finite lexical verb occur in

42 Note that in Standard German as well as in Swiss German, past participles of modal verbs are replaced by their infinitives (*Ersatzinfinitiv*, see page 362).

the right sentence bracket. Inversion occurs in all Swiss dialects except Graubünden and parts of St. Gallen (SADS I/6).

(214) dass du diese Geldbuße zahlen musst
- das du die Puess zale muesch (GR, SGO)
- das du die Puess muesch zale 'that you must pay this fine'

3. Compound tense subordinate clauses, in which the finite auxiliary verb and the non-finite lexical verb appear in the right sentence bracket. In this context, inversion only occurs in the Bern dialect. While the inverted variant was used exclusively in the 1950s (SDS III/261), the un-inverted variant now seems to be preferred by many speakers (SADS I/19).[43]

(215) als der Hund eine Wurst gefressen hat
- wo dr Hund e Wuurscht het gfrässe (BE)
- wo de Hund e Wuurscht gfrässe hät 'when the dog ate a sausage'

4. The second and the third point can be combined: in compound tense subordinate clauses with a modal verb, three verb forms occur in the right sentence bracket, which increases the combinatorics of word orders. While the same inversion pattern occurs in most Swiss German dialects, a second inversion pattern is restricted to parts of GR. Other parts of GR maintain the Standard German word order. Three other potential word orders do not appear in Swiss German dialects (SADS II/6, II/14).

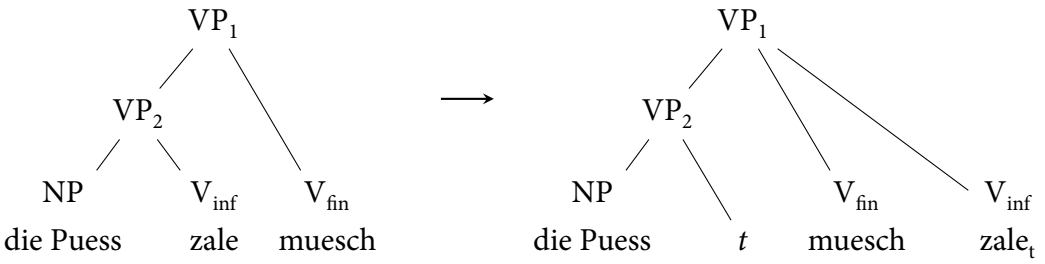
(216) dass du diese Geldbuße hast zahlen müssen
- das du die Puess häsch zale müese (GR)
- das du die Puess zale häsch müese (GR)
- das du die Puess häsch müese zale 'that you have had to pay this fine'

In all of these cases, the inversion takes place inside the verb cluster. Complements and adjuncts are not affected. Following Evers (1975), this phenomenon is known as **verb raising** (for a

43 Note that the SDS map refers to the sequence *gsi bi/bi gsi* 'has been', while the SADS data contains the sequence *zalt hät/hät zalt* 'has paid'. It remains to be investigated if the observed differences in geographical distribution are a result of lexical preferences or of language change.

discussion, see Schönenberger 1995). The idea behind this term is that the non-finite verb form is taken out of its VP and raised to the governing VP that contains the finite verb form:[44]

(217)



A slightly more complicated variant is **verb projection raising**, which raises the whole infinitival verb phrase (222). Negations do not move.

(218) Der Hund hat die Wurst nicht fressen sollen.
– De Hund hät nöd söle d Wuurscht frässe.

(219) dass du diese Geldbuße zahlen musst
– das du muesch die Puess zale

(220) als der Hund eine Wurst gefressen hat
– ?wo dr Hund het e Wuurscht gfrässe (BE)[45]

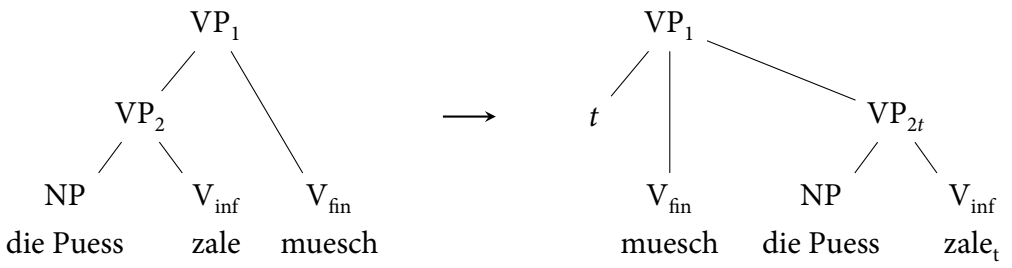
(221) dass du diese Geldbuße hast zahlen müssen
– das du häsch müese die Puess zale

44 The underlying assumption is that the basic word order of Swiss German is SVO, and that verbal complements are thus attached to the left of the head in deep structure.

45 Lötscher (1978, p. 24) suggests evidence for the verb projection raising variant in BE dialect. However, the SADS data does not mention this word order at all in a similar sentence (I/19). The Standard German sentence and its untested Bernese variant are as follows:

(i) Ich habe keine Ahnung, ob sie das Auto schon bezahlt hat.
– ?Ich ha kei Ahnig, ob si het das Auto scho zalt. 'I have no idea if she has already paid that car.'

(222)



Unfortunately, no data is available about the geographic distribution of verb projection raising. We suppose that it occurs in the same regions as the verb raising variants.

Raising and topological fields Let us briefly explain how these different word orders fit into the theory of topological fields. In the Standard German word order, all verbal material is accumulated in the right sentence bracket, the complements of the lexical verbs usually appear in the *Mittelfeld*, and the *Nachfeld* does not contain related material:

(223) [Der Hund]$_{VF}$ [hat]$_{LK}$ [die Wurst | nicht]$_{MF}$ [fressen | sollen]$_{RK}$ []$_{NF}$.

(224) []$_{VF}$ [dass]$_{LK}$ [du | diese Geldbuße]$_{MF}$ [zahlen | musst]$_{RK}$ []$_{NF}$.

(225) []$_{VF}$ [wie]$_{LK}$ [der Hund | eine Wurst]$_{MF}$ [gefressen | hat]$_{RK}$ []$_{NF}$.

(226) []$_{VF}$ [dass]$_{LK}$ [du | diese Geldbuße]$_{MF}$ [hast | zahlen | müssen]$_{RK}$ []$_{NF}$.

The Swiss German verb raising variants merely invert the elements inside the right verb brackets. In contrast, the analysis of verb projection raising is complicated by the restriction that the *Mittelfeld* may not contain verbal material. One possible solution is to maintain the first verbal element in the right sentence bracket, and to analyze the remaining complements and verbal elements as an embedded clause that goes into the *Nachfeld*:

(227) [De Hund]$_{VF}$ [hät]$_{LK}$ [nöd]$_{MF}$ [söle]$_{RK}$ [d Wuurscht frässe]$_{NF}$.

(228) []$_{VF}$ [das]$_{LK}$ [du]$_{MF}$ [muesch]$_{RK}$ [die Puess zale]$_{NF}$.

(229) []$_{VF}$ [das]$_{LK}$ [du]$_{MF}$ [häsch müese]$_{RK}$ [die Puess zale]$_{NF}$.

This analysis may look awkward, but it corresponds exactly to what is proposed for Standard German infinitivals with *zu*. Example (230a) is structurally parallel to the Standard German word order with modals, example (230b) reflects the Swiss German verb projection variant, and example (230c) is parallel to the Swiss German verb projection raising variant. The following section will discuss these types of verbs in more detail.

(230) a. Hans hat das Auto zu reparieren versucht.
b. Hans hat das Auto versucht zu reparieren.
c. Hans hat versucht, das Auto zu reparieren. 'Hans has tried to repair the car.'

(231) a. $[\text{Hans}]_{\text{VF}}$ $[\text{hat}]_{\text{LK}}$ $[\text{das Auto}]_{\text{MF}}$ $[\text{zu reparieren} \mid \text{versucht}]_{\text{RK}}$ $[]_{\text{NF}}$.
b. $[\text{Hans}]_{\text{VF}}$ $[\text{hat}]_{\text{LK}}$ $[\text{das Auto}]_{\text{MF}}$ $[\text{versucht} \mid \text{zu reparieren}]_{\text{RK}}$ $[]_{\text{NF}}$.
c. $[\text{Hans}]_{\text{VF}}$ $[\text{hat}]_{\text{LK}}$ $[]_{\text{MF}}$ $[\text{versucht}]_{\text{RK}}$, $[\text{das Auto zu reparieren}]_{\text{NF}}$.

Lötscher (1978) remarks that the Standard German word order with the sentence-final inflected verb is usually maintained in Swiss German in simple verb clusters, while more inversions take place with growing complexity of the verb cluster. He takes this as evidence that the Swiss German inversions occur for performance reasons because they reduce the amount of left-branching structures which are cognitively hard to process. Reducing the cognitive load would be particularly important in spoken language varieties like Swiss German, but less so in written Standard German. However, it is unclear how this hypothesis relates to his findings that the Swiss German word order was more widespread in Middle High German.

Word order in serial verb clusters

Besides modal and auxiliary verbs, verb clusters (and the resulting word order changes) appear in conjunction with lexical verbs that subcategorize an infinitival complement. These verbs include perception verbs (*sehen* 'to see', *hören* 'to hear'), causative verbs (*lassen* 'to let', *helfen* 'to help'), verbs of beginning and ending (*anfangen* 'to begin', *beginnen* 'to begin', *aufhören* 'to stop'), and others (*versuchen* 'to try', *vergessen* 'to forget', *lernen* 'to learn'). However, not all of these verbs share the same syntactic properties.

In this section, we start by relating the different word orders triggered by lexical verbs with

those triggered by modal and auxiliary verbs discussed above. Then, we discuss the theoretical implications of these word orders on the theory of syntactic complexity. Afterwards, we deal with lexical and geographical variation patterns of word order phenomena. We end this section by clarifying some formal differences in the lexical forms of the relevant verbs and by an explanation of the *Ersatzinfinitiv* phenomenon.

Coherent, incoherent and third constructions Recall the example used above to illustrate three possible word orders in Standard German (we swap the second and third variants for the sake of presentation logic):

(232) a. Hans hat das Auto zu reparieren versucht.
b. Hans hat versucht, das Auto zu reparieren.
c. Hans hat das Auto versucht zu reparieren.

The first sentence represents a so-called **coherent** construction (for a discussion, see Rambow 2003). In this case, the main verb and the infinitival are adjacent to each other, and the objects of both verbs occur in the same topological field, the *Mittelfeld*. In contrast, the second sentence represents an **incoherent** construction. The sentence is split into two clauses: a matrix clause with the main verb and its objects, and an embedded clause with the infinitival and its objects. In the third sentence, the word order of the coherent construction is inverted inside the right sentence bracket; for lack of a better term, this construction has been called **third construction** (den Besten and Rutten 1989).

The distinction between mono-clausal (coherent, third) and bi-clausal (incoherent) constructions is most apparent in subordinate clauses. The remainder of the presentation will therefore be based on another set of examples:

(233) a. $[]_{VF}$ [weil]$_{LK}$ [er | dir | das Geschirr]$_{MF}$ [abwaschen | hilft]$_{RK}$ $[]_{NF}$
b. $[]_{VF}$ [weil]$_{LK}$ [er | dir]$_{MF}$ [hilft]$_{RK}$, [das Geschirr abzuwaschen]$_{NF}$
c. ? $[]_{VF}$ [weil]$_{LK}$ [er | dir | das Geschirr]$_{MF}$ [hilft | abzuwaschen]$_{RK}$ $[]_{NF}$
'because he helps you wash up the dishes'

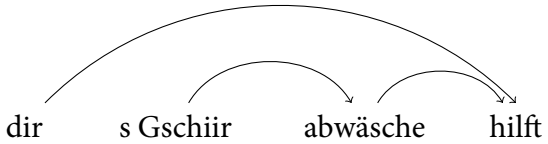
Swiss German dialects also show these three construction types:[46]

(234) a. $[]_{VF}$ [wil]$_{LK}$ [er | dir | s Gschiir]$_{MF}$ [abwäsche | hilft]$_{RK}$ $[]_{NF}$
b. $[]_{VF}$ [wil]$_{LK}$ [er | dir]$_{MF}$ [hilft]$_{RK}$ [s Gschiir abwäsche]$_{NF}$
c. $[]_{VF}$ [wil]$_{LK}$ [er | dir | s Gschiir]$_{MF}$ [hilft | abwäsche]$_{RK}$ $[]_{NF}$
'because he helps you wash up the dishes'

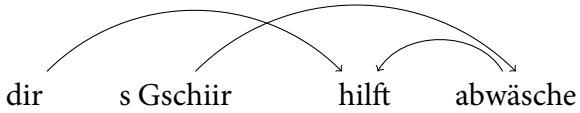
As said in the previous section, these three configurations can be compared with the three modal cluster configurations. If (234a) corresponds to the original structure, then (234b) is obtained by verb projection raising, and (234c) is obtained by verb raising (or, for that matter, by verb projection raising and scrambling, see Schönenberger 1995, section 5.3).

In the surface structure of (234a), the complements in the *Mittelfeld* and the verbal material in the right bracket form a dependency tree without crossing branches. In contrast, the third construction example (234c) yields a tree with crossing branches:

(235) a.



b.



Note that when the embedded clause does not have any complements, it is not possible to distinguish a "third" from an incoherent structure:

(236) a. $[]_{VF}$ [wil]$_{LK}$ [er | dir]$_{MF}$ [hilft | abwäsche]$_{RK}$ $[]_{NF}$
b. $[]_{VF}$ [wil]$_{LK}$ [er | dir]$_{MF}$ [hilft]$_{RK}$ [abwäsche]$_{NF}$

Third constructions and non-context-freeness As discussed in Section 2.3, the Swiss German third constructions have contributed to the debate on the computational complexity of human language. Shieber (1985) has shown that third constructions, containing cross-serial dependen-

46 The alternance of infinitive forms with and without *zu* will be discussed below.

cies, cannot be analyzed by a context-free grammar. While similar examples from Dutch could be generated by a context-free grammar (albeit not with the linguistically desirable constituent structure; this fact refuted the claim of the so-called *strong* context-freeness of natural language), the Swiss German constructions could not even be analyzed *weakly* by a context-free grammar, because the cases of the arguments depended on the crossed verbs.

Let us illustrate this phenomenon with some examples. Each example shows three acceptable Swiss German variants: a coherent variant, an incoherent variant, and a third construction variant. The indices in the examples show that the case of *Hans* depends on the verb *laa* or *hälfe* (A), while the case of *Huus* depends on *aaschtriiche* (B). Only the third construction variants with the crossed dependencies *ABAB* are evidence for non-context-freeness.

(237) a. dass wir Hans ein Haus anstreichen lassen
- das mer [de Hans]$_A$ [es Huus]$_B$ [aaschtriiche]$_B$ [lönd]$_A$
- das mer [de Hans]$_A$ [lönd]$_A$ [es Huus]$_B$ [aaschtriiche]$_B$
- das mer [de Hans]$_A$ [es Huus]$_B$ [lönd]$_A$ [aaschtriiche]$_B$

'that we let Hans paint a house'

b. dass wir Hans helfen, ein Haus anzustreichen
- das mer [em Hans]$_A$ [es Huus]$_B$ [aaschtriiche]$_B$ [hälfed]$_A$
- das mer [em Hans]$_A$ [hälfed]$_A$ [es Huus]$_B$ [aaschtriiche]$_B$
- das mer [em Hans]$_A$ [es Huus]$_B$ [hälfed]$_A$ [aaschtriiche]$_B$

'that we help Hans to paint a house'

The use of transformational grammars and traces overcome the limitation of pure context-free grammars, albeit at substantial cost in terms of computational complexity. In the field of dependency grammar, so-called **non-projective** structures allow crossing arcs. Since non-projective structures already occur in Standard German dependency analyses, the syntactic machinery required to transform Standard German dependency structures into Swiss German dependency structures does not need to expanded further to account for the Swiss German word orders.

While Shieber's findings do not immediately threaten our perspective of machine translation, he pointed out a linguistically interesting feature that our system needs to cover. Cross-serial dependencies will therefore be implemented, although there is no data available about the geographic distribution of the different variants.

Lexical variation The general picture of word oder in serial verb clusters is complicated by several factors of lexical and geographical variation:

- the types of constructions (incoherent, coherent, third) allowed with a particular verb;
- the presence or absence of the infinitive particle *zu* (*z* in Swiss German);
- the presence of reduced verb particles for *lassen* and *anfangen*.

The following table lists the variation patterns for several verbs in Standard German and in Swiss German. ***I*** stands for incoherent, ***C*** for coherent, and ***3*** for third constructions. ✓ denotes the presence of a syntactic construction without additional particle; [+zu] stands for the presence of a construction with the *zu/z* particle, whereas [+la] and [+afa] marks the use of a reduced verb particle. — denotes the absence of a construction.

(238)

	Standard German			*Swiss German*		
Verb	***I***	***C***	***3***	***I***	***C***	***3***
Perception verbs	—	✓	—	✓	—	✓
lassen	—	✓	—	[+la] (W)	✓(O)	[+la] (W)
helfen	[+zu]	✓	[+zu]	✓	✓	✓
anfangen	[+zu]	[+zu]	[+zu]	[+afa] (W) / ✓(O)	—	[+afa] (W) / ✓(O)
aufhören	[+zu]	[+zu]	[+zu]	✓	—	—
lernen	[+zu]	✓	[+zu]	✓	✓	✓
vergessen	[+zu]	[+zu]	[+zu]	✓/ [+zu] (NO)	—	✓

With the verbs presented above, Standard German third constructions are at the limit of acceptability. In any case, they necessarily occur with a *zu*-infinitive, while Swiss German third constructions occur with bare infinitives. Standard German third constructions appear more readily with verbs of thinking (*glauben* 'believe', *denken* 'think', *meinen* 'mean'). Such constructions only occur marginally in Swiss German, as a result of syntactic borrowing from Standard German. Rather, they would be used with a verb-second subordinate clause:

(239) weil er einen Regenbogen glaubt zu sehen

– wil er glaubt, er gsäch en Rägeboge 'because he thinks that he sees a rainbow.'

The table above also shows that the *zu*-particle is less frequent in Swiss German than in Standard German. As a result, the coherent constructions with *zu* are ruled out in Swiss German:

(240) Er hat zu rauchen aufgehört. – *Er hät rauche ufghört. 'He stopped smoking.'

Lexical forms In order to clarify the examples in the following section, two idiosyncrasies in the lexical realization of the mentioned verbs need to be presented here.

In most Swiss German dialects (except in Basel and the Northeastern region around SH, TG, SG, AP, see SDS III/96-98), perception verbs have initial *g* in the entire inflection paradigm:

(241) a. hören – ghööre, hööre (BS, NO) 'to hear'
b. sehen – gsee, see (BS, NO) 'to see'
c. spüren, fühlen – gschpüüre, schpüüre (BS, NO) 'to feel'

The lexemes *lernen* 'to learn' and *lehren* 'to teach' collapse onto one lexeme *leere* in most dialects. Northeastern dialects use *lerne* in the sense 'to learn', and a different lexematic type (i.e. *biibringe* or similar) for the sense 'to teach'. In Uri, both senses collapse in the lexeme *lerne* (SDS V/15).

Ersatzinfinitiv As with modal verbs, the past participle of Standard German *lassen* is replaced by an infinitive:

(242) er hat das Haus anstreichen *gelasst / lassen 'he had the house painted'

This phenomenon is called *Ersatzinfinitiv*. In Swiss German, it interferes with the presence of the verb doubling particles *la* and *afa*, which replace the past participle in Western dialects (see example (209) on page 370).

Perception verbs also use *Ersatzinfinitiv* in Standard German and most Swiss German dialects, except those on the Northwestern fringe of Switzerland:

(243) Du hast mich kommen hören.

- Du häsch mi ghört cho.
- Du häsch mi ghöre cho. (BA, SO, BE-Seeland, FR) 'You heard me come.'

The verbs *helfen* and *lernen* use infinitive in the same Swiss regions, but not in Standard German:

(244) Er hat geholfen abzuwaschen.
- Er hät ghulfe abwäsche.
- Er hät hälfe abwäsche. (BA, SO, BE-Seeland, FR) 'He helped washing up.'

(245) Er hat kochen gelernt.
- Er hät gleert choche.
- Er hät leere choche. (BA, SO, BE-Seeland, FR) 'He learned to cook.'

A.4.7. Subordinate clauses

In this section, we present some phenomena that are not directly linked to a specific category of words. Rather, they all refer to the way subordinate clauses are used in Swiss German.

Infinitival subordinate clauses

For purposive subordinate clauses, the construction *um…zu* + infinitive is used in Standard German. In Western Swiss German, it is substituted by *für…z*. In Eastern Swiss German, *zum* is used. Some mixed forms also occur marginally (SADS I/1, I/6, I/11, IV/14).

(246) Ich habe mich gesetzt, *um* ein Buch *zu* lesen.
- I bi aneghocket *für* es Buech *z* läse. (W)
- I bi aneghocket *zum* es Buech läse. (O) 'I sat down to read a book.'

In *Vorfeld* subordinate clauses, (simple) *zu* is often omitted (Lötscher 1983, pp. 105-106):

(247) Deswegen den Kopf *zu* verlieren nützt nichts.
- Wäge dem de Chopf verlüüre nützt nüt.

'To lose one's mind because of that is not worth it.'

Argument clauses that are introduced by *zu* in Standard German tend to be rendered with a finite subordinate clause:

(248) Es freut uns, Sie wieder einmal zu sehen.
– S fröit is, das mer sy wider emool gseend. 'We are pleased to see you again.'

In contrast, adjunct clauses like the following keep the *zu*:

(249) Er ist weggegangen, ohne etwas zu sagen.
– Er isch furtggange, ooni öppis z säge. 'He left without saying anything.'

As noted on page 368, in sentences of the type *something/nothing/a lot to …*, the gerund form *-id* is used in AP, SG (SDS III/1, SADS II/4).

Double complementizer in indirect questions

Indirect questions are introduced by an interrogative adverb. In Swiss German, the conjunction *das* may be added immediately after the interrogative adverb (Penner and Bader 1995):

(250) a. Ich will wissen, wo du wohnst.
– Ich wott wüsse, wo *(das)* d wohnsch. 'I want to know where you live.'
b. Ich frage mich, wem diese Uhr gehört.
– ich froog mi, wäm *(das)* die Uur ghöört. 'I ask myself whose watch this is.'
c. Ich weiss, wer gekommen ist.
– Ich weiss, wer *(das)* cho isch. 'I know who came.'

Question SADS II/26 provides geographical information about this phenomenon. The *das* variant is accepted nearly everywhere, but the variant without *das* is preferred in most regions. As an interesting small-scale variant, the second person singular clitic is omitted in Solothurn dialect: *ich wott wüsse, wo wohnsch*. It is unclear if this omission also occurs after other complementizers (*das, wil*).

Additional *das* can also appear after some temporal conjunctions (Lötscher 1983, pp. 112-113).

(251) Wir warten, bis (dass) der Vater heimkommt.
– Mir warted, bis *(das)* de Vatter hai chunt. 'We wait until father comes home.'

The use of *dass* is possible in Standard German, but considered obsolete. The variant with additional *das* is frequent, albeit not preferred, in BE-Oberland (SADS IV/12).

Embedded main clause structures

Embedded clauses that follow main clause syntax and lack an overt complementizer (*uneingeleitete Nebensätze*) exist in Standard German, but are more popular in Swiss German (Lötscher 1983, p. 115; Penner and Bader 1995), especially in two contexts.

First, subordinate clauses with the complementizer *dass* 'that' in Standard German are rendered as subordinate clauses with the finite verb in the first position (empty *Vorfeld*):[47]

(252) Ich bin froh, dass ich fertig bin.
– I bi froh, bini fertig. 'I am glad I have finished.'

Second, conditional clauses are transformed to dependent clauses with the finite verb in the second position:

(253) Ich wäre froh, wenn ich bald fertig wäre.
– I wäär froh, i wäär bald fertig. 'I would be glad if I finished soon.'

These transformations are optional, and subordinate clauses with conjunctions are generally preferred (SADS II/25). In that case, the conjunction *das* is pronounced *as* in AP, GL, BA.

47 We introduce commas for readability; Dieth (1986) does not propose specific punctuation rules for Swiss German.

Long *wh*-extractions *(Satzverschränkung)*

Swiss German dialects allow long extractions out of complement clauses. For example, relative pronouns may be put at the front of the main clause, but remain bound to the verb of the subordinate clause. Long extractions are restricted to main clauses with a verb of reporting or belief. This phenomenon is called *Satzverschränkung* in German grammar (Penner and Bader 1995). It is fairly frequent in colloquial German, but rare in modern written Standard German (Andersson and Kvam 1984).

(254) a. Wo hast du gesagt, dass ein Abfalleimer ist?
– Wo häsch gseit das en Abfallchübel isch? 'Where did you say is a trash can?'
b. ?Wer hast du gesagt, dass gestern gekommen ist?
– Wer häsch gsäit dass geschter cho isch? 'Who did you say came yesterday?'

SADS data suggest that long extractions are common in all dialects (IV/24, IV/26). They can also be combined with the coordinated type of dependent clauses, as discussed above:[48]

(255) a. Wo hast du gesagt, ist ein Abfalleimer?
– Wo häsch gseit isch en Abfallchübel?
b. Wer hast du gesagt, ist/sei gestern gekommen?
– Wer häsch gsäit isch/seg geschter cho?

Relative clauses

We have already presented the specificities of relative clauses in the section on relative pronouns (p. 337). We have pointed out that most Swiss German dialects use an invariable relative particle *wo*, while in some smaller dialect regions, inflected relative pronouns are used. The invariable particle complicates relative resumption for dative and prepositional objects. Otherwise, Swiss German relative clauses follow the syntactic structure of Standard German relative clauses.

48 Since this example is a case of indirect speech, the use of subjunctive is fairly common (255b).

A.4.8. Characteristics of spoken language

When referring to Standard German, one often implicitly refers to its written norm rather than to its spoken usage. Hence, differences between Swiss German and Standard German are not necessarily interdialectal differences, but may simply arise of the different contexts of use. Several examples illustrate this fact.

In general, sentences tend to be shorter and have a simpler internal structure. Lötscher (1983, p. 114) also notes that the *Nachfeld* position is occupied more often in Swiss German than in Standard German. This would reduce cognitive load by taking material out of the *Mittelfeld* and anticipating the verbal material of the right sentence bracket.

Baur (1983, p. 32) and Lötscher (1983, pp. 115-116) mention that dialects tend to use parataxis more frequently than hypotaxis, especially in contrastive and concessive contexts:

(256) a. Ich arbeite, *während* du spazieren gehst.
   – Ich schaffe *und* du gaasch go schpaziere. 'I work while you go for a walk.'
   b. *Obwohl* Fritz vier Omelette gegessen hat, hat er immer noch nicht genug.
   – De Fritz het vier Amelette ggässe *und* hät immer nonig gnue.
   'Although Fritz has eaten four omelettes, he still hasn't had enough.'

Another common feature of spoken language is the reduced use of nominalizations, although nominalization tendencies have begun to appear with increasing dialect use in news broadcast. Nevertheless, complex noun phrases are still frowned upon by dialect purists.

We believe that these features should not be handled in a machine translation system. First, they are tendencies rather than categoric rules, and their real frequency in everyday language has not been studied. Second, they are features of a particular modality, not of a particular dialect or dialect area. Transcribed Standard German speech will contain many of the features mentioned above. Such a source text will result in a relatively natural Swiss German translation, while Standard German material produced for written use will lead to less natural Swiss German translations.

Finally, the distinction of cliticized and stressed pronouns, the insertion of *sandhi* consonants

to ensure smooth pronunciation, as well as the loss of the preterit tense and the word order differences in verb clusters, have commonly been attributed to the spoken nature of Swiss German, and are also found in more standard-like varieties of spoken German. As we have pointed out in the preceding sections, these phenomena will be accounted for.

A.5. Lexical resources

A description of a language usually consists of two parts, a grammar and a lexicon. The grammar describes the morphological and syntactic processes needed to create sentences. The lexicon, further separated in closed word classes and open word classes, lists the lexical elements used as building blocks by the grammar. Closed word classes contain a relatively small and stable number of items. Typically, prepositions, conjunctions, determiners and pronouns form closed classes; special types of verbs like modals or auxiliaries also form a closed class. In contrast, open word classes are of potentially infinite size because they are augmented constantly by derivation or borrowing. Nouns, lexical verbs and adjectives typically form open classes.

How do these distinctions fit into the the picture that we have been drawing in the preceding sections? The core of Swiss German grammar has been exposed in the morphosyntax section: noun, adjective and verb inflection have been presented, and various issues of periphrastic constructions and word order have been discussed. Inevitably, this description included some closed-class words required by morphological or syntactic processes: pronouns for possessive dative constructions, prepositions for prepositional dative marking, auxiliary verbs for periphrastic verb forms.

Furthermore, we have presented two approaches for generating open-class words. Swiss German words can be generated from known Standard German words with the help of phonetic rules, or from other Swiss German words by means of derivation. The time and space limitations of this work do not allow us to present the Swiss German lexicon in more detail. However, we are aware that many Swiss German words cannot be generated correctly with these two techniques. For instance, some historic accidents have made that certain concepts are expressed completely differently in the respective language varieties. The following paragraphs will thus present some

relevant lexicographic work and discuss the potential issues of integrating it into a machine translation system.

The most complete collection of Swiss German lexical material is the *Wörterbuch der schweizerdeutschen Sprache*, also commonly called *Schweizerisches Idiotikon* (Staub et al. 1881-). It is based on various sources from the 19th and 20th century. Started in 1881, 15 ⅔ volumes have been published, 1 ⅓ will follow until 2020. The published volumes have been made available for full text search through optical character recognition.[49] However, the search results are still displayed in the form of page scans. This makes the data difficult to utilize in a computational setting.

All lexical entries in the *Idiotikon* come with geographical references, but the latter are not systematic and difficult to extract reliably. While an electronic, georeferenced dialect dictionary of Swiss German dialects unfortunately will not be realized soon, exactly this goal is being achieved for the Austrian German dialects. The Austrian counterpart of the *Idiotikon*, the *Wörterbuch der bairischen Mundarten in Österreich*, is not only being digitized, but the geographical references of each lexical item are systematized, so that they can easily be drawn on a map (Wandl-Vogt 2006a; Wandl-Vogt 2008).

In the preceding sections, we have already referred to several grammars of specific Swiss dialects, which have been published in the second half of the 20th century. Some of these grammars were accompanied by a dictionary volume: Suter (1984) for Basel dialect (recently re-edited as Christoph Merian Stiftung (2010)), Greyerz and Bietenhard (1976) for Bern dialect, and A. Weber and Bächtold (1961) for Zürich dialects. However, these collections have been published too early to be available electronically, and they only cover selected regions of the Swiss dialect landscape.

The volumes IV to VIII of the SDS also deal with lexical issues. However, its editors chose a set of words that very much depended on the choice of informants and on the time of data collection. The criteria for the selection of lexical phenomena described in the SDS are detailed in Section 3.3.4.

49 http://www.idiotikon.ch/Register, accessed 19.4.2012.

Another potential resource for Swiss German lexical material are word lists posted on the internet by committed dialect speakers.[50] Laymen often focus on lexical differences because they perceive these as the most salient type of interdialectal differences. While many items in these lists are surely characteristic of the given dialect, most of them are quite rare in everyday speech. In consequence, the usefulness of these lists may be limited.

To sum up, Swiss German dictionaries either lack digital availability, precise georeferencing, or both. Despite these drawbacks, we believe that explicit lexical information enhances the performance of a dialect machine translation system. In order to keep the workload at acceptable levels, we propose to include lexical information restricted to high frequency words which are not correctly generated with the phonetic rules. This selective approach guarantees that the most frequent words will be translated correctly without wasting time with rare lexical irregularities.

A.6. Conclusion

The aim of this chapter was to present the most salient linguistic differences between Standard German and Swiss German dialects, and among the various Swiss German dialects. We have started with a presentation of phonetic differences, with the goal to use phonetic transformations to generate dialect word forms from Standard German word forms at run time. We have further presented phenomena of derivational and inflectional morphology as well as syntax, with the aim of incorporating these phenomena as transformation rules from a Standard German source.

We believe that the data presented in this chapter further motivates our choice of integrating

50 For example:

- http://www.syndicate.ch/zueriduetsch/ (Zürich dialect),
- http://gaby-paul.freepage.de/woerterbu.htm (Zürich dialect),
- http://www.berndeutsch.ch (Bern dialect),
- http://www.edimuster.ch/baernduetsch/woerterbuechli.htm (Bern dialect),
- http://www.dialektwoerter.ch/ (supraregional),
- http://www.drs1.ch/www/de/drs1/themen/mundart.html (supraregional).

All links accessed on 11.2.2010.

dialectological maps to parametrize the transformation rules. Indeed, the interdialectal variation is considerable, and different linguistic phenomena vary across different isoglosses, such that we find it inadequate to conceive the Swiss German dialects as discrete, homogeneous entities.

B. List of digitized SDS maps

General remarks

This appendix lists the SDS maps that have been digitized. Every section refers to one digital map, which may have been compiled from several original SDS maps, or which may contain only part of the information contained in the original SDS map. Hence, the SDS map reference is indicated in the section title. For phonetic maps, the phoneme/sound in question is marked with square brackets.

Variants are shown as list items. The number of inquiry points in which a variant is valid is printed in angle brackets. A star means that the corresponding variant is handled as a "default" variant, i.e., that the number of inquiry points is subtracted from all other variants. If the variant name used in the *ArcGIS* files is not identical with the string shown in the list items, the variant name is indicated at the right edge of the column.

The list is generated automatically from the database that has been populated during the digitization process. All annotations were made in German; hence the German comments in the list.

1/035: [e]ng

- ä ⟨347⟩
- e (3 Symbole) ⟨211⟩

Nicht berücksichtigte Varianten:

- i ⟨6⟩

1/045: D[o]nnstag

- o (Rest) ⟨372⟩
- u (2 geschlossenste Varianten) ⟨160⟩

Nicht berücksichtigte Varianten:

- a

Palatalisierung und Brechung nicht berück-

sichtigt.

1/048: Schl[i]tten

- i (2 geschlossenste Varianten) ⟨303⟩
- e (Rest) ⟨262⟩

Brechung nicht berücksichtigt.

1/050: Ch[u]chi

- o (Rest) ⟨293⟩
- u (2 geschlossenste Varianten) ⟨283⟩

Palatalisierung nicht berücksichtigt.

1/052: F[ü]chse

- ü (2 geschlossenste Varianten) ⟨267⟩
- ö (Rest) ⟨217⟩
- i (alle palatalisierten Varianten) ⟨75⟩

1/060: [ü]ber

- über, iber ⟨398⟩ über
- uber ⟨184⟩

ü/i siehe 1/156

1/060: d[u]rch

- dur ⟨409⟩
- dür, dir ⟨230⟩ dür

ü/i siehe 1/156

1/061: [A]bend

- o (2 Symbole), ou ⟨329⟩ o
- a (4 Symbole) ⟨261∗⟩

Mischwerte bei beiden Klassen verzeichnet.

1/095: Schn[ee]

- ee ⟨552∗⟩ mono
- ei ⟨29⟩ di

Vokalqualität und Brechung (eä) nicht berücksichtigt.

1/100: L[oh]n

- o ⟨526∗⟩
- u ⟨25⟩

Brechung (oe) und Diphthongierung (ou) nicht berücksichtigt.

1/105: [Ei]s

- ii ⟨553∗⟩
- ei ⟨7⟩
- i ⟨4⟩

ii angenommen an Punkten, wo das Wort nicht vorkommt.

1/106: M[au]s

- uu, üü ⟨535∗⟩ uu
- ui ⟨18⟩

- u ⟨6⟩
- ou ⟨2⟩

Nicht berücksichtigte Varianten:

- öi ⟨1⟩
- üi ⟨1⟩
- ü ⟨1⟩

u-Qualität (BA) nicht berücksichtigt.

Entrundung siehe Karte 1/152.

1/107: M[äu]se

- üü, ii ⟨555∗⟩ üü
- ü ⟨4⟩
- ei ⟨3⟩
- öü ⟨2⟩

Nicht berücksichtigte Varianten:

- ui ⟨1⟩

i-Qualität nicht berücksichtigt.

Entrundung siehe Karte 1/156.

1/109: G[ei]ss

- ei (alle e-Werte), ee, ii ⟨235⟩ ei
- äi ⟨212⟩
- ai ⟨54⟩
- aa ⟨29⟩
-ää ⟨18⟩
- oa ⟨6⟩

Nicht berücksichtigte Varianten:

- ìi (Diphthong) ⟨4⟩
- oo ⟨3⟩
- ae ⟨1⟩

Monophthongierung BE/FR siehe 1/121. Monophthongierung O integriert, da 1/124 nicht geeignet.

1/112: S[eil]

- eil ⟨485∗⟩ L
- eu (alle e-Varianten mit Vokalisierung) ⟨87⟩
- öü (alle ö-Varianten mit Vokalisierung) ⟨30⟩

Nicht berücksichtigt: Vokallänge, Vokalqualität (i/e/è), Vokal der nicht vokalisierten Varianten (siehe 1/109).

1/121: [Au]gen

- ou, oo, uu, öi ⟨303⟩ ou
- au ⟨251⟩
- ai ⟨8⟩
- oi ⟨4⟩

Palatalisierung siehe separate Karte.

Monophthongierung siehe separate Karte.

1/121: [Au]gen (Monophthong)

- ou, au ⟨501⟩ di
- o, ù ⟨64⟩ mono

Vokallänge nicht berücksichtigt.

1/124: B[au]m

- ou, au ⟨405⟩ ou

- o, ö ⟨102⟩ o
- oo, öö ⟨58⟩ oo

Diphthongqualität siehe 1/121.

Palatalisierung siehe separate Karte.

n-Endung nicht berücksichtigt (siehe heim/hein).

1/129: r[äu]chern

- öi ⟨292⟩
- oi ⟨247⟩
- ai ⟨27⟩

Kein Scan! Von 1/124 adaptiert.

Entrundung und Monophthongisierung siehe andere Karten.

1/134: t[ie]f

- öi, öü, äu, öö, üü, ei ⟨281⟩ öi
- üü ⟨178⟩
- ie ⟨74⟩
- äi ⟨36⟩

Nicht berücksichtigte Varianten:

- ii ⟨2⟩

Entrundung und Monophthongierung siehe andere Karten.

1/138: Kn[ie]

- öi, ei ⟨277⟩ öi
- üü, ii ⟨257⟩ üü
- äu ⟨19⟩
- äi ⟨9⟩

Nicht berücksichtigte Varianten:

- öö ⟨1⟩

Entrundung siehe andere Karten.

1/148: schn[ei]en

- ei ⟨287⟩
- ii ⟨186⟩
- ai ⟨106⟩

1/152: Entrundung u/ü

- uu(w) ⟨543∗⟩ u
- üü(w) ⟨22⟩ ü

Gilt für Entrundung u → ü, o → ö.

1/152: b[au]en

- ou (alle senkrechten Zeichen) ⟨285⟩
- uu(w), üü(w) ⟨143⟩ uu
- au (alle waagerechten Zeichen) ⟨120⟩
- ui(w) ⟨15⟩ ui

Nicht berücksichtigte Varianten:

- öi ⟨1⟩
- iib ⟨2⟩

Palatalisierung (senkrechte Schraffierung) nicht berücksichtigt.

w im Hiatus siehe separate Karte.

Entrundung siehe 1/152.

1/152: bau(w)en

- uu, üü, ui ⟨458∗⟩ now
- uuw, üüw, uiw, üüb ⟨107⟩ w

Unterscheidung w/b nicht berücksichtigt (b 2 Nennungen WS).

Gilt für Hiatus bei Phänomenen 1/152 und 1/156 (nicht aber 1/152i).

1/156: Entrundung ü/i

- öi, oi ⟨486∗⟩ ü
- ei, ii ⟨79⟩ i

Gilt für Entrundungen ü → i, ö → e, öi → ei, ou → öi.

1/156: n[eu]er

- öi, ei (senkrechte Zeichen) ⟨238⟩ öi
- oi, öi (waagerechte Zeichen) ⟨152⟩ oi
- üü(w), ii(w) ⟨148⟩ üü
- üüi, üi ⟨46⟩ üi

Nicht berücksichtigte Varianten:

- ui ⟨1⟩
- ew ⟨1⟩
- ei ⟨1⟩
- euw ⟨3⟩
- ööw ⟨4⟩
- öüw ⟨3⟩

Entrundung (Schraffur) siehe separate Karte.

1/160: [A]pfel

- ö ⟨476⟩
- e ⟨95⟩

Beide Karten von Vollkarte subtrahiert.

1/161: Schw[e]ster

- ö ⟨303⟩
- e ⟨273⟩

Beide Karten von Vollkarte subtrahiert.

1/164: Br[i]lle

- ü ⟨379⟩
- i ⟨180⟩

Beide Karten von Vollkarte subtrahiert.

l-Vokalisierung nicht berücksichtigt.

1/165: M[i]lch

- i, ì ⟨544∗⟩ i
- ü, ö ⟨76⟩ ü

Nicht berücksichtigte Varianten:

- u ⟨1⟩

l-Vokalisierung nicht berücksichtigt.

2/003ff: Schn[a]bel

- a, ä ⟨511∗⟩ K
- aa, ää ⟨280⟩ L

Jede der Karten 2/003 (Laden), 2/004 (weben), 2/008 (sagen), 2/026 (Schnabel), 2/044 (Gabel) ergibt einen Parameter, die dann alle zusammengefasst werden. allfeats.shp manuell erstellt.

Halblange und lange Varianten zusammengefasst.

Vokalqualität nicht berücksichtigt.

2/022: f[ah]ren

- aa ⟨541∗⟩
- a ⟨30⟩

Nicht berücksichtigte Varianten:

- oa ⟨*IT*⟩

Halblang und lang zusammengefasst.

2/044: s[a]gen

- ä ⟨555⟩
- a ⟨10⟩

Nur Vokalqualität berücksichtigt. Länge siehe 2/003ff.

2/069: Gesch[irr]

- iir ⟨342⟩
- ier ⟨107⟩
- ir ⟨103⟩
- ür ⟨10⟩

Halblange und lange Varianten zusammengefasst.

Separate Karte mit i → ü FR???

Schwa-Qualität bei Diphthong nicht berücksichtigt.

2/071, 2/072: schr[ei]ben

- ii, ei ⟨558∗⟩ I
- i ⟨252⟩ k

Die Karten ergeben 4+3 Parameter, die gezählt werden.

2/073: m[a]len

- oo ⟨291⟩
- aa ⟨214⟩
- a ⟨50⟩

2/074: m[ah]len

- aa ⟨317⟩
- a ⟨222⟩

2/077: r[ei]ten

- i ⟨290⟩
- ii ⟨261⟩
- ei ⟨7⟩

Halblang und lang zusammengefasst.

2/084: Pfl[aum]e

- uum ⟨398⟩
- uumm ⟨57⟩
- um ⟨50⟩
- umm ⟨42⟩
- uim ⟨14⟩
- uimm ⟨2⟩

Nicht berücksichtigte Varianten:

- öi/oi ⟨1⟩

In 2 separate Kartensets aufteilen: Vokallänge + Konsonantenlänge.

Vergleichsmaterial beachten.

2/087: s[ä]gen

- aa ⟨395∗⟩
- ä ⟨75⟩
- a ⟨56⟩
- oo, o ⟨22⟩ oo
- ää ⟨21⟩

Nicht berücksichtigte Varianten:

- oa ⟨*IT*⟩
- eä ⟨1⟩

2/088: tr[a]gen

- ää ⟨389∗⟩
- aa ⟨137⟩
- ä ⟨37⟩

Nicht berücksichtigte Varianten:

- a ⟨3⟩
- oo ⟨1⟩
- oa ⟨*IT*⟩

2/093: Gr[a]s, sp[a]ren, ...

- a ⟨549⟩
- ä ⟨65⟩

Vokallänge und Qualität nicht berücksichtigt.

Für alle vier Wörter je ein Shapefile mit ä-Werten erstellt, dann zusammengefasst und Werte summiert. allfeats.shp erhält ein Feld cnt_ä[0..4] und ein Feld cnt_a [4..0].

2/094: [K]ind

- ch (verschiedene Lauttypen) ⟨547∗⟩
- kch, kh ⟨14⟩ k

2/095: -li[ch]

- -lich (ck + gg Ostschweiz) ⟨549∗⟩
- -lig (gg Basel) ⟨16⟩

Von 2/095 abgeleitet, mangels besserem Material.

2/095: drü[ck]en

- ck ([kx], [gx]) ⟨485∗⟩
- gg ([k], [g]) ⟨82⟩

2/097: tr[ink]en

- ink ⟨378∗⟩ nk

- ingg ⟨74⟩ ngg
- iich, eech ⟨61⟩ Lch
- iih, eeh ⟨44⟩ Lh
- eich ⟨24⟩ Dch

Nicht berücksichtigte Varianten:

- eih ⟨1⟩

Vokalqualität nicht berücksichtigt.

2/113: wa[chs]t

- chs ⟨428∗⟩
- ggs ⟨132⟩

Nicht berücksichtigte Varianten:

- gsch ⟨*IT*⟩

2/118: la[ng]

- ng, nng ⟨513∗⟩ ng
- ngg, ngk ⟨50⟩ ngg

2/119: fi[nd]en

- nd ⟨422∗⟩
- ng, nng ⟨92⟩ ng
- nn ⟨41⟩
- n ⟨13⟩

nn/nd-Distribution ist nicht mit 2/120 kongruent.

2/120: Hu[nd]

- nd ⟨413∗⟩
- ng, nng ⟨94⟩ ng
- nt ⟨47⟩
- n, nn ⟨27⟩ nn

Zwischenwerte wurden in Übergangsgebieten (BEO, WS, GR) als nd und als nt annotiert, ansonsten als nd.

Unterscheidung n/nn wird idR nicht verschriftlicht, daher zusammengefasst.

2/138: ge[rn]

- rn (2 Symbole) ⟨397∗⟩
- re ⟨153⟩
- r ⟨11⟩
- ren ⟨11⟩

Schwa-Qualität nicht berücksichtigt.

2/143: Da[rm]

- rm ⟨502∗⟩
- re ⟨41⟩
- rem ⟨24⟩

Vokalkürzung, Vokalbrechung und Umlaut nicht berücksichtigt.

2/144: Läu[s]e

- s (bei allen Wörtern ausser Eis) ⟨506∗⟩
- sch (mindestens bei Läuse, Mäuse) ⟨59⟩

2/147: fo[l]gen

- l (normales oder velarisiertes l) ⟨434∗⟩
- u (vokalisiertes l) ⟨130⟩

2/161: m[ähe]n

- äje (alle Varianten mit j) ⟨463⟩
- äe (Diphthong) ⟨100⟩
- ää (Monophthong) ⟨14⟩

Qualität des Vollvokals und des Schwa nicht berücksichtigt.

2/162: br[ühe]n

- üeje (alle Varianten mit j) ⟨418⟩
- üe ⟨153⟩

Qualität des Vollvokals und des Schwa nicht berücksichtigt.

2/164: [D]eckel

- t (inkl. Zwischenwerte) ⟨411∗⟩
- d ⟨126⟩

Lücken Ostschweiz gefüllt.

2/165: [T]ag

- t (inkl. Zwischenwerte) ⟨487∗⟩
- d ⟨90⟩

Nicht identisch mit 2/164 (WS, GR).

2/170: Gar[b]e

- b ⟨509∗⟩
- p (inkl. Zwischenwerte) ⟨74⟩

Mischwerte bei beiden Varianten verzeichnet.

2/171: ho[b]len

- b ⟨517∗⟩
- p (inkl. Zwischenwerte) ⟨52⟩

Konsistent mit 2/170, obwohl Zwischenwerte subjektiv eher b zuzuordnen wären.

Phonetischer Kontext hängt von Regel el → le ab.

2/172: Ra[d]

- d (inkl. Zwischenwerte) ⟨526∗⟩
- t ⟨40⟩

Zwischenwerte nicht konsistent mit 2/170, 2/171.

2/186: Ta[nn]e

- nn (geminiert oder lang) ⟨391⟩
- n ⟨179∗⟩

2/192: Sa[m]en

- m ⟨491∗⟩
- mm (geminiert oder lang) ⟨76⟩

2/198: Te[ll]er

- ll (geminiert oder lang) ⟨377⟩
- u (aus geminiertem, langem oder kurzem l) ⟨114⟩
- l ⟨112⟩

Nicht berücksichtigte Varianten:

- ld

Velarisierung nicht berücksichtigt.

In FR Vergleichsmaterial ”Keller”benutzt.

3/001: Endung Gerundium

- e ⟨428∗⟩
- ä ⟨59⟩
- id, ed (nicht unterschieden) ⟨37⟩ id
- u (mindestens 1 Klasse) ⟨27⟩
- a ⟨18⟩

Unterscheidet sich von 3/001 Infinitiv nur durch die id/ed-Variante.

3/001: Infinitiv

- n vor Vokal, 0 vor Konsonant ⟨545∗⟩ 0
- n in allen Kontexten ⟨20⟩ n

3/001: Schwa 2 Varianten

- e ⟨508⟩
- ä ⟨59⟩

Varianten u/a/e zusammengefasst zu e.

Für unbetonte -el, -er, -en.

3/001: Schwa 3 Varianten

- e ⟨464⟩
- ä ⟨87⟩
- a ⟨18⟩

Varianten u/ä zusammengefasst zu ä.

Für Diphthongkontexte ie, ue, üe.

3/001: Schwa Infinitiv

- -e ⟨464∗⟩
- -ä ⟨59⟩
- -u (mindestens 1 Klasse) ⟨28⟩
- -a ⟨18⟩

3/003: Präfix PP

- p-, t-, k-/gg- (trunke) ⟨440∗⟩ fort
- g-, d-, g- (drunke) ⟨93⟩ len
- gi- (gitrüüche) ⟨25⟩
- ga- (gatroucha) ⟨11⟩

Gilt nur für Plosive. Sonst überall g-.

3/003: Schwa PP

- e ⟨456∗⟩
- ä ⟨65⟩
- u ⟨26⟩
- a ⟨18⟩

Karte fotografiert.

Praktisch identisch mit 3/001.

3/007: PP bringen

- pracht, procht, gabracht ⟨467∗⟩ pracht
- prunge, gibrunge ⟨125⟩ prunge

3/008: PP laufen

- gloffe ⟨305⟩
- glüffe, gliffe ⟨140⟩ glüffe
- glaufe ⟨121⟩
- gluffe ⟨60⟩

3/009: Inf tragen

- träääge (rote Symbole) ⟨425∗⟩
- traage, draage (schwarze Symbole) ⟨140⟩ traage

3/009: PP tragen

- treit, träit, dräit, trait, treet, traat ⟨460⟩ treit
- traage, gitraage ⟨78⟩ traage
- träääge ⟨31⟩

3/016: sch[ie]ssen

- ie ⟨345∗⟩
- üü ⟨224⟩

3/016: verl[ie]ren

- üü ⟨392⟩
- ie ⟨172∗⟩

3/018: fl[ie]gen, l[ü]gen

- üü ⟨450∗⟩
- öü ⟨89⟩
- ie ⟨44⟩

3/019: ziehen

- zie ⟨478∗⟩
- züche ⟨48⟩
- zühe ⟨31⟩
- züe ⟨17⟩

3/022: 1Sg Normalverben

- -e ⟨395∗⟩
- 0 ⟨72⟩
- -ä ⟨59⟩
- -u ⟨27⟩
- -a ⟨18⟩

Alle Varianten (ausser 0) haben kontextabhängiges -n.

Vokalqualität von 3/001 kopiert.

Syntaktischer und phonetischer Kontext nicht berücksichtigt (siehe darauffolgende Karten).

Gilt gemäss Legende nur im Satzauslaut!

3/031, 3/034: 1Pl Normalverben

- -e, -en, -ä, -u, -i ⟨286⟩ e
- -ed, -et ⟨168⟩ ed
- -id ⟨91⟩
- -en, -end, -ent ⟨39⟩ end

Karte fotografiert.

Varianten -ä, -u, -i je weniger als 5 Punkte.

3/032: 2Pl Normalverben

- -ed, -et, -ät, -ut ⟨402⟩ ed
- -id, -it ⟨155⟩ id
- -end, -en, -ent ⟨20⟩ end

- -e ⟨16⟩

Varianten -ät, -ut je weniger als 5 Punkte.

3/033: 3Pl Normalverben

- -e, -en, -ä, -u, -i ⟨255⟩ e
- -ed, -et ⟨168⟩ ed
- -id ⟨91⟩
- -end, -en, -ent ⟨61⟩ end
- -und, -unt ⟨24⟩ und

3/040, 3/041: 1Sg Kurzverben

- -n nur vor Vokal (grosser oder kleiner Punkt) ⟨538∗⟩ 0
- -ne (3/041) ⟨139⟩
- -n in allen Kontexten (kleiner Punkt oder kein Punkt) ⟨89⟩ n

Zusatzkarte -ne zur Zeit nicht benutzt.

3/044: 1Pl Kurzverben

- -nd (wie 3/031 -ed + -id + -end) ⟨344⟩
- 0 (wie 3/031 0 + -und) ⟨243⟩
- -n ⟨11⟩
- -nt ⟨3⟩

Teilweise von 3/031 abgeleitet.

3/044: 2Pl Kurzverben

- -nd ⟨334⟩
- -t ⟨245⟩
- -d ⟨22⟩
- -n ⟨2⟩

3/044: 3Pl Kurzverben

- -nd (wie 3/033 -ed + -id + -end) ⟨346⟩
- 0 (wie 3/033 -e) ⟨210⟩
- -nt (wie 3/033 -und) ⟨34⟩
- -n ⟨12⟩

Teilweise von 3/033 abgeleitet.

3/046: 1Sg haben

- ha, han ⟨501∗⟩ ha
- hä, hän ⟨64⟩ hä

3/046: Inf haben

- haa ⟨530∗⟩
- hää ⟨35⟩

3/046: PP haben

- gha ⟨513∗⟩
- ghäbe ⟨54⟩
- ghä ⟨25⟩
- ghabe ⟨15⟩

Nicht berücksichtigte Varianten:

- ghabet ⟨6⟩
- ghäbet ⟨5⟩
- gcha (nicht auf Karte)

3/047: Pl haben

- he- ⟨141⟩

- hei- ⟨133⟩
- hä- ⟨132⟩
- häi- ⟨65⟩
- hi- ⟨38⟩
- ha- ⟨19⟩

Nicht berücksichtigte Varianten:

- ho- ⟨8⟩
- hö- ⟨5⟩
- Mischformen hä-/häi- ⟨6⟩
- häid (GL)

3/048: 23Sg haben

- he- (Region I) ⟨382⟩
- hä- (Region II) ⟨202⟩

3/050: Inf/PP sein

- (g)sii ⟨511⟩ si
- (g)see ⟨49⟩ se

ö-Varianten zu e-Varianten zusammengefasst.

3/050: Pl sein

- si- ⟨511⟩
- se- (Fläche Nordost + Punktsymbole LU) ⟨37⟩
- sö- ⟨12⟩

Nicht berücksichtigte Varianten:

- sei- ⟨5⟩
- sie- ⟨3⟩

3/054: PP tun

- taa, gitaa, daa ⟨437⟩ taa
- tue ⟨125⟩

3/056: 1Sg gehen

- goo(n), gaa(n), ... ⟨470⟩ goo
- gang ⟨63⟩
- gange ⟨49⟩

Vokalqualität siehe separate Karte.

3/056ff: 1Sg Vokalqualität Kurzverben

- o (2 Symbole) ⟨306⟩
- a ⟨240⟩
- u ⟨29⟩
- ou ⟨7⟩

Nicht berücksichtigte Varianten:

- ä (schlää, lää, fää)
- ee, ei (schlee, schlei, fee)

Aus Karten 3/056 (gehen), 3/064 (lassen), 3/070 (schlagen), 3/074 (anfangen). Maximale Ausdehnung jeder Variante.

3/056ff: Inf/PP kommen

- cho (wie 1Sg -a + -o) ⟨534⟩ o
- chu ⟨29⟩ u
- chou ⟨7⟩ ou

Von 1Sg abgeleitet.

3/057: 23Sg gehen

- gaa-, goo- ⟨372⟩ goo
- gei- ⟨115⟩
- gee-, gii- ⟨60⟩ gii
- gäi- ⟨26⟩

Vokalqualität siehe separate Karte.

Mischformen bei beiden Klassen verzeichnet.

3/057ff: 23Sg Vokalqualität Kurzverben

- o (2 Symbole) ⟨315⟩
- a ⟨257⟩
- ou ⟨13⟩

Nicht berücksichtigte Varianten:

- ä (schlääsch)
- ee, ei (schleesch, schleisch, geisch, schteisch)

Aus Karten 3/057 (gehen), 3/071 (schlagen), 3/075 (anfangen). Maximale Ausdehnung jeder Variante.

3/058: Pl gehen

- gö-, gä-, ga-, go-, ... ⟨511⟩ gö
- gange- ⟨36⟩ gang
- gönge- ⟨22⟩ göng

Vokalqualität (und -länge) siehe separate Karte.

Mischwerte bei beiden Klassen verzeichnet.

3/058ff: Pl Vokalqualität Kurzverben

- ö (inkl. entrundet zu e) ⟨333⟩
- a ⟨115⟩
- e (3 Symbole) ⟨60⟩
- ä ⟨44⟩
- o (2 Symbole) ⟨39⟩
- öi ⟨22⟩
- ü ⟨17⟩
- u ⟨9⟩
- ai ⟨5⟩

Nicht berücksichtigte Varianten:

- ie (schliend)

Aus Karten 3/058 (gehen), 3/059 (stehen), 3/066 (lassen), 3/072 (schlagen), 3/076 (anfangen). Maximale Ausdehnung jeder Variante.

3/059: Pl stehen

- stö-, stä-, sta-, ... ⟨503⟩ stö
- stande, stange, stane ⟨30⟩ stand
- stönde ⟨25⟩ stönd

Vokalqualität siehe separate Karte.

Mischwerte bei beiden Klassen verzeichnet.

3/064: 1Sg lassen

- laa(n), loo(n), luu(n) ⟨552∗⟩ loo
- loss ⟨13⟩

Nicht berücksichtigte Varianten:

- laasse ⟨3⟩

Vokalqualität siehe separate Karte.

3/066: Pl lassen

- lö-, lä-, la-, löi- ⟨528⟩ lö
- löit, löie, löied ⟨18⟩ löie

Nicht berücksichtigte Varianten:

- laasse ⟨5⟩

Vokalqualität siehe separate Karte.

3/070: 1Sg schlagen

- schlaa(n) (2 Symbole), schloo(n), schluu(n), ... ⟨512⟩ schloo
- schlag ⟨28⟩
- schlee, schlää, schlei ⟨17⟩ schlee

Nicht berücksichtigte Varianten:

- schlach(e) ⟨6⟩
- schlah(e) ⟨8⟩

Vokalqualität siehe separate Karte.

3/070: Inf schlagen

- schloo ⟨512⟩
- schlage ⟨28⟩ schlag
- schlee ⟨17⟩

Unterschied zu 1Sg??

3/071: 23Sg schlagen

- schlaa- (2 Symbole), schloo-, schlou- ⟨511⟩ schloo
- schlee-, schlää-, schlei- ⟨45⟩ schlee

Nicht berücksichtigte Varianten:

- schlach- ⟨11⟩ (verteilt)
- schlag- ⟨3⟩
- schleisch ⟨3⟩
- schlechsch ⟨2⟩

Vokalqualität siehe separate Karte.

3/072: Pl schlagen

- schlö-, schlä-, ... ⟨514⟩ schlö
- schlage- (3 Symbole) ⟨21⟩ schlag
- schlöie- ⟨19⟩ Xlöie

Nicht berücksichtigte Varianten:

- schlach ⟨5⟩
- schlah ⟨8⟩
- schlie ⟨1⟩

Vokalqualität siehe separate Karte.

3/074: 1Sg (an)fangen

- faa(n), foo(n) ⟨327⟩ foo
- fang, fange ⟨169⟩ fang
- fee(n), fää(n) ⟨57⟩ fee

Nicht berücksichtigte Varianten:

- fäe ⟨2⟩
- foch ⟨7⟩
- fohe ⟨7⟩
- föö ⟨6⟩
- andere kleinere Varianten

Vokalqualität siehe separate Karte.

3/075: 23Sg (an)fangen

- faa-, foo- ⟨328⟩ foo
- fang- ⟨161⟩
- fee-, fää-, fääh- ⟨85⟩ fee
- foch-, foh- ⟨13⟩ foch

Nicht berücksichtigte Varianten:

- fie ⟨2⟩
- föö ⟨6⟩

3/076: Pl (an)fangen

- fö-, fä-, fe-, föi- ⟨338⟩ fö
- fange- ⟨201⟩ fang
- fönge- ⟨11⟩ föng

Nicht berücksichtigte Varianten:

- föied ⟨1⟩
- fooh- ⟨5⟩
- fooch- ⟨1⟩

Vokalqualität siehe separate Karte.

Karte fotografiert. Legende auf 3/077.

3/078: Inf (an)fangen

- foo, faa, fuu ⟨284⟩ foo
- fange ⟨123⟩
- fee, fää ⟨84⟩ fee

Nicht berücksichtigte Varianten:

- foohe ⟨4⟩
- foche ⟨2⟩

Vokalqualität siehe separate Karte.

3/080: Pl liegen

- ligg- ⟨413∗⟩
- lig- ⟨186⟩

3/081: 23Sg liegen

- liisch, liit ⟨477∗⟩ lii
- liggsch, liggt ⟨103⟩ ligg
- liggosch, liggot ⟨24⟩ liggo

3/086: 13Sg müssen

- mues ⟨512∗⟩
- mue ⟨53⟩

3/087: Pl müssen

- müe- ⟨292⟩
- müesse- ⟨244∗⟩
- mue- ⟨31⟩

Nicht berücksichtigte Varianten:

- muesse ⟨3⟩
- müi ⟨6⟩

3/090: Pl geben

- gä-, ge-, gi- ⟨511⟩ gä
- gäbe-, gebe-, gibe- ⟨89⟩ gäbe
- gäi- ⟨11⟩

Vokalqualität siehe 3/094.

3/091: Pl nehmen

- näme-, neme-, nime- ⟨334⟩ näm

- nä-, ne-, ni- ⟨258⟩ nä
- näi- ⟨13⟩

Nicht berücksichtigte Varianten:

- nämd ⟨5⟩

Vokalqualität siehe 3/094.

3/092: 1Sg nehmen

- nim (nim+nin+nii) ⟨493⟩
- nüm (nüm+nün) ⟨45⟩
- nem (nem+nen+nee) ⟨29⟩

Von 23Sg abgeleitet.

3/092: 23Sg nehmen

- nim- ⟨439∗⟩
- nin- ⟨50⟩
- nüm- ⟨36⟩
- nen- ⟨17⟩
- nii- ⟨14⟩
- nee- ⟨12⟩
- nün- ⟨11⟩

Nicht berücksichtigte Varianten:

- nem ⟨2⟩

3/094: Vokalqualität geben/nehmen

- ä (2 Symbole) ⟨454⟩
- e (2 Symbole) ⟨89⟩
- i ⟨34⟩

Nicht berücksichtigte Varianten:

- ea ⟨6⟩

3/095: PP nehmen

- gnoo (2 Symbole) ⟨370⟩
- gnuu ⟨131⟩
- gnaa ⟨29⟩
- gnoe ⟨28⟩

Nicht berücksichtigte Varianten:

- gnou ⟨5⟩

3/096: Präfix sehen

- g- ⟨502⟩
- 0 ⟨84⟩

3/096: Sg sehen

- gsee-, see- ⟨449∗⟩ see
- gsie-, sie- ⟨90⟩ sie

Nicht berücksichtigte Varianten:

- sech- ⟨9⟩
- siech- ⟨7⟩

3/097: Pl sehen

- see-, gsee- ⟨429∗⟩ see
- sie-, gsie- ⟨59⟩ sie
- sech-, gsech- ⟨30⟩ sech
- gsej- ⟨24⟩ sej
- gsei- ⟨17⟩ sei
- gseä- ⟨16⟩ seä

Nicht berücksichtigte Varianten:

- sieh- ⟨8⟩
- siech- ⟨2⟩
- sea- ⟨3⟩

ei/ej: Legende unklar, je 2 Varianten.

3/100: 1Sg kommen

- chum (alle u-Varianten in 23Sg) ⟨512⟩
- chom (alle o-Varianten in 23Sg) ⟨40⟩
- chüm (alle ü-Varianten in 23Sg) ⟨15⟩

Von 23Sg abgeleitet.

3/100: 23Sg kommen

- chun- ⟨441∗⟩
- chu- ⟨37⟩
- chuu- ⟨29⟩
- chon- ⟨28⟩
- chüü- ⟨15⟩
- chui- ⟨14⟩
- choo- ⟨12⟩

3/102: Pl kommen

- chöme- ⟨339∗⟩ chöm
- chome- ⟨75⟩ chom
- cheme- ⟨63⟩ chem
- chö- ⟨45⟩
- cho- ⟨34⟩
- chäme- ⟨25⟩ chäm
- chä- ⟨15⟩

Mischformen bei beiden Karten verzeichnet.

3/105: Pl können

- chö-, che- ⟨226⟩ chö
- chöne-, chene- ⟨221⟩ chöne
- chöi- ⟨65⟩
- chöö- ⟨31⟩
- chä- ⟨19⟩

Nicht berücksichtigte Varianten:

- chäne- ⟨4⟩
- chü- ⟨5⟩
- chu- ⟨2⟩
- chune- ⟨1⟩
- chüne- ⟨2⟩

Entrundung ö → e nicht berücksichtigt.

3/106: Sg mögen

- mag ⟨452∗⟩
- ma ⟨115⟩

Nicht berücksichtigte Varianten:

- mage ⟨4⟩

3/107: 13Pl mögen

- möge-, mege- ⟨502⟩ möge
- mö-, möö-, mee- ⟨53⟩ mö
- möi- ⟨10⟩

Von 2Pl abgeleitet: möge-, mege-, mög-, meg-, möch-, mech- zusammengefasst.

3/107: 2Pl mögen

- möge-, mege- ⟨438∗⟩ möge
- mö-, möö-, mee- ⟨53⟩ mö

- mögt, megt ⟨50⟩ mög
- möcht, mecht ⟨27⟩ möch
- möit ⟨10⟩ möi

Nicht berücksichtigte Varianten:

- alle mo-Varianten

3/108: Pl dürfen

- töörfe- (alle f-haltigen Varianten in Sg) ⟨513⟩ töörf
- tööre- (alle f-losen Varianten in Sg) ⟨50⟩ töör

Von Sg abgeleitet.

3/108: Sg dürfen

- taarf ⟨421⟩
- töörf ⟨87⟩
- taar ⟨43⟩

Nicht berücksichtigte Varianten:

- terf ⟨8⟩
- tärf ⟨1⟩
- töör ⟨4⟩
- täär ⟨4⟩

3/112: 2Sg wollen

- wottsch ⟨247⟩ wotsch
- witt ⟨201⟩
- wosch ⟨110⟩
- wilt ⟨50⟩

Nicht berücksichtigte Varianten:

- wilsch ⟨10⟩ (verteilt)
- wottisch ⟨9⟩
- woltisch ⟨9⟩
- wöttsch, wettsch (wenige Nennungen, siehe Legende)

3/113: 3Sg wollen

- will ⟨344⟩
- wott ⟨342⟩
- wolt ⟨23⟩

Nicht berücksichtigte Varianten:

- wilt ⟨*IT*⟩

wolt ist fast überall mit Alternativvariante verzeichnet; wahrscheinlich veraltet, aber trotzdem berücksichtigt.

3/114: Pl wollen

- wä-, we- ⟨268⟩ wä
- wei-, wäi-, wai- ⟨142⟩ wei
- wee-, wii- ⟨51⟩ wee
- welle- ⟨38⟩ well
- wö- ⟨23⟩

Nicht berücksichtigte Varianten:

- wäint ⟨2⟩
- wäid ⟨8⟩
- wellend ⟨5⟩
- well ⟨1⟩
- weue ⟨4⟩
- wolle (häufig, aber veraltet und fast überall mit Alternativvariante)
- wolte ⟨5⟩ (aber veraltet und fast überall mit Alternativvariante)

Vokalqualität ä/e siehe 1/035. Vokalqualität ei siehe 1/109.

3/131: das

- s ⟨368⟩
- ds ⟨247∗⟩

Schraffur nicht berücksichtigt.

Überlappungen bei beiden Varianten verzeichnet.

3/132: die (Fem Sg / Pl)

- d ⟨513∗⟩
- di ⟨69⟩

Überlappungen bei beiden Varianten verzeichnet.

Phonetischer Kontext nicht berücksichtigt, siehe 3/133.

3/134: der

- de ⟨362⟩
- dr ⟨350∗⟩

Überlappungen doppelt verzeichnet. Phonetischer Kontext nicht berücksichtigt.

Kann auch für Dat Fem Sg benutzt werden, aber 3/135 wäre genauer.

3/142: ein (Neut)

- es ⟨447∗⟩
- e ⟨237⟩

Schraffierte Punkte und Überlappungen doppelt verzeichnet. Syntaktischer Kontext nicht berücksichtigt, siehe 3/143.

3/144: eine

- e ⟨532∗⟩
- en ⟨105⟩

Nicht berücksichtigte Varianten:

- ne

Schraffur nicht berücksichtigt.

Überlappungen doppelt verzeichnet.

3/145: ein (Masc)

- e ⟨370⟩
- en ⟨284∗⟩

Nicht berücksichtigte Varianten:

- ne

Schraffur nicht berücksichtigt.

Überlappungen doppelt verzeichnet.

3/152: Diminutiv

- Umlaut + -li (2 Symbole) ⟨506⟩ U_li
- -ji ⟨26⟩

Nicht berücksichtigte Varianten:

- -li
- Umlaut + -i (2 Varianten)
- -i
- Umlaut + -ji

3/154: Diminutiv nach -el

- Umlaut + -eli (2 Symbole) ⟨515⟩ U_eli
- -ulti, -elti, -lti, -ilti ⟨34⟩ ulti
- Umlaut + -i ⟨27⟩ U_i
- Umlaut + -elti ⟨11⟩ U_elti

3/159: Fem Derivation -erin

- -eri ⟨403∗⟩ i
- -ere ⟨135⟩ e
- -era ⟨32⟩ a

Nicht berücksichtigte Varianten:

- -erne ⟨3⟩

3/162: Fem Derivation -in

- -i ⟨327∗⟩
- -in ⟨115⟩
- -ene, -enä ⟨92⟩ ene
- -eni ⟨42⟩

3/163: Derivation -ung

- -ig (Strich oder Kreuz in Karte) ⟨540∗⟩
- -ung (grosser Kreis in Karte) ⟨60⟩
- -ing ⟨13⟩

Nicht berücksichtigte Varianten:

- ug ⟨1⟩

3/164, 3/165, 3/166: M Pl -e/0

- 0 ⟨468⟩
- -e ⟨381⟩

Schwa-Qualität nicht berücksichtigt.

Zusamengefasst aus Karten 3/164, 3/165, 3/166.

3/187: F Pl -i

- -ene (alle Symbole ohne Punkt) ⟨419⟩
- -ine (alle Symbole mit Punkt) ⟨95⟩
- -eni ⟨48⟩

Nicht berücksichtigte Varianten:

- -ne ⟨4⟩
- -ini ⟨5⟩
- -i ⟨10⟩ (verteilt)

Schwa-Qualität nicht berücksichtigt.

3/195, 3/196: ich (voll, proklitisch)

- i, ii ⟨381⟩ i
- ich, iich, iech ⟨333⟩ ich
- ig, iig ⟨63⟩ ig

Varianten i und ich: jeweils maximale Ausdehnung der Karten 3/195 und 3/196.

3/195, 3/196: mich, dich

- mi, mii ⟨370⟩ mi
- mich, miich, miech ⟨333⟩ mich
- mir ⟨11⟩

Analog zu ich/i.

Variante mir gemäss Legende.

Gleiche Karte gilt auch für dich/di/dir.

3/201: sie (Fem Sg)

- si(i), schi(i) ⟨523∗⟩ si
- sie (Diphthong, 2 Symbole) ⟨35⟩
- sei (2 Symbole) ⟨10⟩

Nicht berücksichtigte Varianten:

- seie ⟨3⟩

3/201: sie (Pl)

- si(i), schi(i) ⟨528∗⟩ si
- sü ⟨16⟩
- sei (2 Symbole) ⟨10⟩

Nicht berücksichtigte Varianten:

- söü ⟨4⟩
- sie ⟨5⟩
- si/siu ⟨3⟩

3/203, 3/204: Vokalqualität -i(h)r

- -ir, -iir, -ihr ⟨352⟩ i
- -ier ⟨218⟩ ie

Nicht berücksichtigte Varianten:

- öö ⟨4⟩

Gilt für mir, dir, wir, ihr.

3/203: wir

- mir, miir, mier, möör ⟨494⟩ mir
- wir, wier ⟨73⟩ wir

3/204: ihr (2Pl)

- ir, iir, ier, öör ⟨391⟩ ir
- dir, diir, dier ⟨180⟩ dir

3/207, 3/217: uns(-er)

- üs-,isch- ⟨338⟩ üs
- öis- ⟨201⟩
- üns-, ünsch-, insch- ⟨36⟩ üns
- uns- (BS) ⟨4⟩

Jeweils maximale Ausdehnung der Varianten in den Karten 3/207 und 3/217.

3/209: euch

- öich ⟨212⟩
- öi ⟨200⟩
- üüch ⟨45⟩
- ööch ⟨41⟩
- euw ⟨35⟩

Nicht berücksichtigte Varianten:

- üich ⟨7⟩
- üü ⟨8⟩
- iuch ⟨4⟩
- iuw ⟨4⟩
- euwch ⟨7⟩
- eib ⟨2⟩

3/211: M mein/dein/sein

- mi ⟨328⟩
- min ⟨200⟩
- mine ⟨36⟩

Vokallänge nicht berücksichtigt.

3/212: F mein/dein/sein

- mi ⟨432∗⟩
- mini ⟨134⟩

Nicht berücksichtigte Varianten:

- min ⟨2⟩

Vokallänge nicht berücksichtigt.

3/213: N mein/dein/sein

- mis ⟨320∗⟩
- mi ⟨59⟩

Vokallänge nicht berücksichtigt.

Lückenhaft; Phänomen ist ähnlich mit 3/252, aber selbiges ist weniger verbreitet.

3/214: Pl mein/dein/sein

- mini ⟨393∗⟩
- miner ⟨92⟩
- mi ⟨58⟩
- min ⟨22⟩

Vokallänge nicht berücksichtigt.

3/215: M ihr

- ire (2 Symbole), ira, irä ⟨421∗⟩ ire
- iren ⟨68⟩
- iru ⟨10⟩

Nicht berücksichtigte Varianten:

- ine ⟨1⟩
- irne ⟨3⟩
- ir ⟨4⟩

Vokalqualität ire/ere/iere nicht berücksichtigt.

3/216: F ihr

- iri (2 Symbole) ⟨417∗⟩
- irni ⟨58⟩
- iru (2 Symbole) ⟨25⟩
- ire ⟨18⟩

Nicht berücksichtigte Varianten:

- ine ⟨6⟩
- irne ⟨8⟩
- irje ⟨3⟩

3/216: Pl ihr

- iri (wie 3/216 iri ohne irer) ⟨358⟩
- irer (wie 3/214 miner) ⟨92⟩
- irni (wie 3/216 irni ohne irer) ⟨58⟩
- iru (wie 3/216 iru ohne irer) ⟨18⟩

Nicht berücksichtigte Varianten:

- Varianten von 3/216
- ire ⟨3⟩ (nach Subtraktion von irer)

Keine Originalkarte, Varianten gemäss Legende rekonstruiert.

3/217: M unser/euer

- -e ⟨396∗⟩
- -ere ⟨104⟩
- -en ⟨54⟩
- -er ⟨13⟩

Nicht berücksichtigte Varianten:

- -em ⟨3⟩

3/217: MN Dat unser/euer

- -em (wie 3/217 e + en) ⟨449⟩
- -erem (wie 3/217 ere + er) ⟨113⟩

Keine Originalkarten; Varianten von Nominativformen abgeleitet.

3/218: F unser/euer

- -i ⟨380∗⟩
- -eri ⟨177⟩

Nicht berücksichtigte Varianten:

- -er ⟨4⟩
- -e ⟨4⟩

3/218: Pl unser/euer

- -i (wie 3/218 -i ohne -er) ⟨288⟩
- -eri (wie 3/218 -eri ohne -er) ⟨177⟩
- -er (wie 3/214 miner) ⟨92⟩

Keine Originalkarte, Varianten gemäss Legende rekonstruiert.

3/218: euer (Stamm)

- öi- ⟨419⟩
- euw- ⟨69⟩
- üü- ⟨39⟩
- iiw- ⟨19⟩

Nicht berücksichtigte Varianten:

- eib-/äib- ⟨4⟩
- euch-/öüch-/üüch- ⟨5⟩

Entrundung öi → ei nicht berücksichtigt.

3/219: N unser/euer

- -ers ⟨302∗⟩
- -es ⟨184⟩
- -er ⟨57⟩
- -e ⟨29⟩

3/220: F Dat meiner

- mir ⟨54⟩ r
- miner ⟨43⟩ ner
- mire ⟨15⟩ re
- minere ⟨5⟩ nere

Daten gemäss Legende in Karte III/220 rekonstruiert (kein Scan).

Allerdings fehlen zu viele Punkte, um die Karte sinnvoll einsetzen zu können.

3/223: was

- was ⟨438∗⟩
- wa(n) ⟨138⟩ wa

3/252: Adj -es

- -s ⟨513∗⟩
- 0 ⟨68⟩

3/253: Adj -e (Pl)

- -i ⟨439∗⟩
- 0 ⟨324⟩

3/254: Adj -e (F Sg)

- 0 ⟨400∗⟩
- -i ⟨163⟩

Nicht berücksichtigte Varianten:

- -u ⟨3⟩

3/261: Wortstellung Nebensatz

- PP Aux (bin gewesen) ⟨446⟩ pp_aux
- Aux PP (gewesen bin) ⟨120⟩ aux_pp

4/149: alles

- alls, allts ⟨278⟩ alls
- ales, alles ⟨143⟩ ales
- aus, auts ⟨141⟩ aus

Nicht berücksichtigte Varianten:

- aues, auwes ⟨2⟩

4/150: als

- wo(n) ⟨429⟩ wo
- wa(n) ⟨41⟩ wa
- wenn ⟨40⟩
- wie(n) ⟨17⟩ wie

Nicht berücksichtigte Varianten:

- wil ⟨8⟩
- as, als ⟨8⟩ (verteilt)

Abwesenheit des Hiatus-n (GR) nicht berücksichtigt.

4/151: auch

- au, äu ⟨222⟩ au
- oo, o ⟨131⟩ o
- ou ⟨125⟩
- oi, öi, ai ⟨62⟩ oi
- ooch,och ⟨7⟩ ooch
- oog ⟨5⟩

Nicht berücksichtigte Varianten:

- ä, ää ⟨5⟩ (verteilt)
- oich ⟨2⟩

4/155, 4/164: ein wenig

- e chli, chlii, chlee ⟨418⟩ chli
- e bitzli ⟨93⟩ bitzli
- en bitz ⟨90⟩ bitz
- e chlei ⟨34⟩ chlei
- e chlai ⟨15⟩ chlai
- e chläi ⟨15⟩ chläi
- e weeni ⟨13⟩ weeni
- e weng ⟨13⟩ weng

Nicht berücksichtigte Varianten:

- e weenig ⟨3⟩
- e wengeli ⟨2⟩
- e wee ⟨2⟩

ei an 1 oder 2 Hauptstellen → Variante ei

ei an Nebenstellen und 1 Hauptstelle → Variante i

4/159: gar

- gar ⟨437⟩
- gär ⟨77⟩

Vokalqualität nicht berücksichtigt.

4/167: nicht

- nid, nit ⟨396⟩ nid
- nöd, nöt ⟨91⟩ nöd
- nüd, nüt (2 Symbole) ⟨45⟩ nüd
- id, it ⟨13⟩ id
- net ⟨9⟩

Nicht berücksichtigte Varianten:

- öd ⟨2⟩

- ni ⟨1⟩

4/168: nicht mehr

- nüme, nümme ⟨305⟩ nüme
- nime, nimme ⟨55⟩ nimme
- nüm, nümm ⟨42⟩ nümm
- nimee, nimmee ⟨31⟩ nimee
- nüpmee/nütmee ⟨28⟩ nütmee
- nipmee/nitmee ⟨26⟩ nitmee
- nume, numme ⟨25⟩ numme
- nümee, nümmee ⟨21⟩ nümee
- nimi, nimmi ⟨15⟩ nimmi
- nome, nomme ⟨11⟩ nomme

Nicht berücksichtigte Varianten:

- alle dreisilbigen Varianten (nümmemee, ...)
- nimm ⟨1⟩
- nüüme ⟨1⟩
- nümi ⟨1⟩
- niime ⟨5⟩ (verteilt)

o-Variante gemäss Legende.

Entrundung ü/i integriert, da unregelmässige Ubereinstimmung mit Karte 1/156.

4/171: nichts

- nüüt, nüüd, niit, niid ⟨365⟩ nüüt
- nüt, nüd, nit, nid ⟨105⟩ nüt
- nünt ⟨47⟩
- nix ⟨33⟩
- nütz ⟨6⟩

Nicht berücksichtigte Varianten:

- neid ⟨3⟩
- nöüd ⟨3⟩
- nuid ⟨1⟩

Halblang und kurz zusammengefasst.

Entrundung nicht berücksichtigt.

4/172: noch

- no ⟨442⟩
- na ⟨45⟩
- noch ⟨28⟩
- noh ⟨22⟩
- nuch ⟨18⟩
- nu ⟨14⟩

Nicht berücksichtigte Varianten:

- nog ⟨1⟩
- nug ⟨2⟩
- nuh ⟨1⟩

4/176: nur

- nume, numme ⟨186⟩ nume
- nu ⟨108⟩
- nur ⟨88⟩
- ume, umme ⟨51⟩ ume
- gad, grad ⟨14⟩ gad
- nun ⟨12⟩
- blooss, blööss ⟨11⟩ blooss

Nicht berücksichtigte Varianten:

- äbä ⟨2⟩
- nuwa ⟨*IT*⟩

4/180: und

- und ⟨423⟩
- u(n) (2 Symbole) ⟨117⟩ u_N
- un ⟨24⟩

Nicht alle u(n) haben ein Hiatus-n, aber die Legende ist zu wenig explizit.

4/181: sonst

- süsch(t) ⟨210⟩ süsch
- susch(t) ⟨206⟩ susch
- sus(t) ⟨99⟩ sus
- süs ⟨27⟩
- sundsch(t) ⟨17⟩ xdsch
- schüsch(t) ⟨7⟩ schx
- suschter ⟨6⟩ xter

Nicht berücksichtigte Varianten:

- schusch(t) ⟨2⟩
- sündsch(t) ⟨5⟩ (verteilt)

Unterscheidung +/- t nicht in Karte (aber in Legende).

4/183: wie (Fragewort)

- wie ⟨413⟩
- wi ⟨227⟩
- we ⟨53⟩

4/184: wieder

- wider ⟨409⟩
- umhi ⟨37⟩
- mum, amum, num, ... ⟨26⟩ mum
- ume, umme, widerume ⟨22⟩ ume
- widerum, widerem ⟨19⟩ wxum
- umi, umi ⟨16⟩ umi

Nicht berücksichtigte Varianten:

- namal, nochemal ⟨4⟩
- aber ⟨6⟩

5/112: Gruss

- grüezi ⟨216⟩ grxzi
- guete Tag ⟨200⟩ gxetag
- grüessech ⟨135⟩ grxch
- guet Tag ⟨109⟩ gxtag
- grüessi ⟨20⟩ grxssi

Nicht berücksichtigte Varianten:

- -gwünscht ⟨3⟩
- -Gott (viele Nennungen, aber nirgends einzige Variante)
- guete Morge ⟨3⟩
- guete Aabe ⟨4⟩
- guete ⟨1⟩
- Grussfragen ⟨2⟩

Unterscheidung guet Tag/guete Tag gemäss Legende.

Unterscheidung -ech/-i/-di gemäss Legende.

Tagwoll u.ä. nicht in Karte.

5/115: danke

- danke, i danke ⟨373⟩ danke
- merci ⟨172⟩
- vergelts Gott ⟨122⟩ vxgx
- dank-schöön ⟨57⟩ dxschx

Nicht berücksichtigte Varianten:

- dank gar schöön ⟨3⟩
- Dank heigischt ⟨5⟩

Mehrfachnennungen berücksichtigt, aber nicht Häufigkeiten.

5/118: schau

- lueg, log, lug ⟨397⟩ lueg
- lue ⟨86⟩
- gugg, gug ⟨44⟩ gugg
- lüeg ⟨29⟩
- lotz ⟨25⟩
- gschou ⟨11⟩

Nicht berücksichtigte Varianten:

- liog ⟨2⟩
- löög ⟨2⟩
- lio ⟨1⟩
- gsich ⟨1⟩

Diese Formen gelten fürs ganze Verbparadigma (ausg. lue/lueg-Unterscheidung).

Vokalqualität-Regeln nachher: ue/uä/...

Auch implementieren: los/lose.

5/119: ja

- jo, joo ⟨332⟩ jo
- ja, jaa ⟨288⟩ ja
- jä, jää ⟨40⟩ jä

Nicht berücksichtigte Varianten:

- jou ⟨2⟩
- Sonderfälle (häufig in BA)

6/020: gestern

- geschter ⟨512∗⟩ gXter
- geschtert ⟨30⟩ gXtert
- gescht ⟨23⟩ gXt
- göschter ⟨9⟩ göXter

Nicht berücksichtigte Varianten:

- geschte ⟨4⟩
- geschtere ⟨4⟩
- geschtet ⟨1⟩
- gerscht ⟨1⟩

6/021: heute

- hüt, hit ⟨539∗⟩ hüt
- hütte, hitte ⟨57⟩ hütte

Entrundung nicht berücksichtigt.

| und || bei beiden Varianten verzeichnet (ohne Häufigkeitsunterscheidung).

6/022, 6/023, 6/024: oft

- mängisch(t) [6/023] ⟨349⟩ mXsch
- mängsmal [6/023] ⟨191⟩ mXsmal
- mängmal [6/023] ⟨167⟩ mXmal
- öppe, eppe, äppe, appe [6/024] ⟨104⟩ öppe
- vil [6/022] ⟨85⟩ vil
- öppenemal, öppemal, eppenemal, äppenemal [6/024] ⟨57⟩ öXmal
- öppedie, öppenedie [6/024] ⟨29⟩ öXdie
- vilmal [6/022] ⟨19⟩ vilmal
- ette, ätte [6/024] ⟨12⟩ ette

Nicht berücksichtigte Varianten:

- dick, dickmal (6/022)
- vilfärt (6/022)
- oft, öftersch (6/022)
- allpott (6/022)
- allewiil (6/022)
- esie (6/022)
- schuppemal, e Schuppe mal (6/022)
- fliissig (6/022)
- mangmol, mangsmol (6/023)
- manchmol, manchsmol (6/023)
- mängimal, mängemal (6/023)
- ötsche (6/024)
- ötschenemal (6/024)
- ettenemal (6/024)
- öpperemal (6/024)
- öppesie (6/024)
- ettesie (6/024)
- friäppe (6/024)

Vokalqualität (mäng/meng, mal/mol) nicht berücksichtigt.

6/026: immer

- immer ⟨226⟩
- allewiil, allwiil ⟨109⟩ alewil
- all ⟨82⟩
- gäng ⟨54⟩
- geng ⟨52⟩
- albig, aldig ⟨28⟩ albig
- ging ⟨27⟩
- eisder, eisdert ⟨16⟩ eisder
- gi ⟨7⟩

Nicht berücksichtigte Varianten:

- ail
- albis, albigs

- amig
- albed
- algu
- allpott, allipott
- allziit, allzig, allzi
- aade
- eisset, eissert
- eischter
- eisdig
- eissig
- eischtig
- eisderig
- ummeder, under
- tuschuur
- ständig, bständig
- zue
- di ganz Ziit

Alle Orte mit kleinem Punkt: ausschliesslich Variante immer.

Alle Orte ohne kleinen Punkt: andere Varianten.

6/027: jeweils

- albe ⟨156⟩
- esie, sie, esienig ⟨88⟩ esie
- ame ⟨64⟩
- amel, ämel ⟨58⟩ amel
- amig ⟨53⟩
- allig ⟨42⟩
- alle ⟨31⟩
- albis, albigs ⟨24⟩ albis
- amed, amid ⟨20⟩ amed
- amigs ⟨20⟩
- albig ⟨19⟩

Nicht berücksichtigte Varianten:

- alme
- abe
- albed, albets
- ames
- albes
- ami
- almig
- almi
- alli
- alligs
- allemal

6/029: früher

- früener ⟨312⟩ fxner
- früecher ⟨181⟩ fxcher
- früejer ⟨150⟩ fxjer
- früer ⟨103⟩

Nicht berücksichtigte Varianten:

- früeher ⟨14⟩ (verteilt)

Kleine und grosse Symbole wurden nicht unterschieden. Karten sind ev. auch für Adjektiv früh nutzbar.

Vokalqualität (üe/ie) nicht berücksichtigt.

6/030: bald

- bald ⟨250⟩
- glii, glei ⟨153⟩ glii
- baud ⟨36⟩
- bau ⟨29⟩

- bold ⟨21⟩
- boll ⟨11⟩
- angänds ⟨10⟩ angx

Nicht berücksichtigte Varianten:

- ball ⟨< 6⟩
- bou ⟨< 6⟩
- blötzli ⟨< 6⟩
- flugs ⟨< 6⟩
- bhänd ⟨< 6⟩
- gschwind, gschwindhaft ⟨< 6⟩

6/098: hier

- daa ⟨273⟩
- hie ⟨190⟩

Nicht berücksichtigte Varianten:

- daahie ⟨11⟩ (verteilt)

Vokalqualität (daa/doo) nicht berücksichtigt.

6/100: heim, nach Hause

- hei, häi, hai ⟨368⟩ hei
- heim, häim ⟨37⟩ heim
- hiim, heem ⟨37⟩ hiim
- häin, hein ⟨32⟩ hein
- haa, hoo, hoa ⟨28⟩ haa
- haam, ham, hom ⟨24⟩ ham
- hii, hee ⟨21⟩ hii
- hää, hèè ⟨13⟩ hää

Nicht berücksichtigte Varianten:

- hoam ⟨4⟩
- hem ⟨1⟩
- hìim (Diphthong) ⟨2⟩
- haim ⟨2⟩
- hiime ⟨2⟩
- heme ⟨*IT*⟩

6/101: dort

- det ⟨229⟩
- dert ⟨116⟩
- dört ⟨89⟩
- daa ⟨79⟩
- döt ⟨66⟩
- dei ⟨45⟩

Nicht berücksichtigte Varianten:

- sälb (nur in Spontanmaterial, immer mit Alternativvariante) ⟨11⟩
- düert ⟨2⟩

Keine Unterscheidung zwischen Erst- und Spontanbelegen. Vokalquantität (ö/öö) nicht berücksichtigt.

6/105: heraus

- use ⟨346⟩
- uuse ⟨95⟩
- usse ⟨30⟩
- usa ⟨25⟩
- uusa ⟨19⟩
- uusser ⟨16⟩
- ussa ⟨12⟩
- uusse ⟨10⟩

Nicht berücksichtigte Varianten:

- Varianten mit < 4 Nennungen

- em-Varianten

Vokalquantität-Regeln nachher anwenden: uu/ui (UW), schwa

6/106: hinaus

- use ⟨304⟩
- uuse ⟨88⟩
- usi ⟨76⟩
- uus ⟨34⟩
- usse ⟨17⟩
- uusi ⟨16⟩
- uss ⟨15⟩
- us ⟨8⟩
- uuss ⟨6⟩

Nicht berücksichtigte Varianten:

- Varianten mit < 4 Nennungen
- em-Varianten
-

Vokalquantität-Regeln nachher anwenden: uu/ui (UW), schwa

6/107: herein

- ine ⟨182⟩
- ie ⟨169⟩
- iine ⟨53⟩
- iche ⟨35⟩
- inha ⟨29⟩
- ihe ⟨21⟩
- icha ⟨15⟩
- iha ⟨12⟩
- iner ⟨12⟩
- inne ⟨12⟩
- iehe ⟨5⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

6/108: hinein

- ie ⟨171⟩
- ine ⟨155⟩
- iine ⟨52⟩
- i ⟨33⟩
- iche ⟨32⟩
- inhi ⟨29⟩
- ihi ⟨24⟩
- ini ⟨19⟩
- iin ⟨17⟩
- ii ⟨16⟩
- inni ⟨9⟩
- inne ⟨7⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

6/109: herauf

- ue ⟨191⟩
- ufe ⟨133⟩
- uufe ⟨78⟩
- uehe ⟨34⟩
- ueche ⟨32⟩
- ueha ⟨18⟩
- uche ⟨15⟩

- wuha ⟨9⟩
- ueher ⟨7⟩
- uer ⟨7⟩
- embrueche ⟨5⟩ emxche
- embruer ⟨5⟩ emxr
- embruuf ⟨5⟩ emxuuf
- uhe ⟨5⟩
- uuf ⟨5⟩
- embruehe ⟨4⟩ emxhe
- embruf ⟨4⟩ emxuf
- uf ⟨3⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

über-Varianten nicht unterschieden.

Die Situation im WS ist sehr kompliziert, daher viele Varianten.

6/110: hinauf

- ue ⟨180⟩
- ufe ⟨113⟩
- uufe ⟨67⟩
- uehi ⟨35⟩
- ueche ⟨34⟩
- uuf ⟨29⟩
- ufi ⟨24⟩
- embruf ⟨16⟩ emxuf
- embruuf ⟨16⟩ emxuuf
- uche ⟨13⟩
- ui ⟨10⟩
- uufi ⟨8⟩
- wuhi ⟨7⟩
- uehe ⟨5⟩
- uf ⟨5⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

über-Varianten nicht unterschieden.

6/111: herab

- abe ⟨305⟩
- aabe ⟨60⟩
- ahe ⟨49⟩
- ache ⟨39⟩
- appe ⟨34⟩
- aha ⟨16⟩
- embriche ⟨16⟩ emxe
- apper ⟨13⟩
- apfa ⟨5⟩
- embricher ⟨5⟩ emxer
- apphe ⟨4⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

6/112: hinab

- abe ⟨288⟩
- aabe ⟨58⟩
- ahi ⟨49⟩
- ache ⟨34⟩
- appe ⟨31⟩
- embrii ⟨28⟩
- aab ⟨23⟩

- abi ⟨19⟩
- apphi ⟨7⟩
- aahi ⟨6⟩

Nicht berücksichtigte Varianten:

- diverse Varianten mit < 4 Nennungen

Bibliography

Aepli, Noëmi and Nora Hollenstein (2012). *Part-of-Speech Tagging für Schweizerdeutsch*. Term paper, Institute for Computational Linguistics, University of Zurich.

Aho, Alfred V. (1990). "Algorithms for finding patterns in strings". In: *Handbook of Theoretical Computer Science*. Ed. by Jan van Leeuwen. Vol. A, Algorithms and Complexity. MIT Press, Pages 255–300.

Allauzen, Cyril et al. (2007). "OpenFst: a general and efficient weighted finite-state transducer library". In: *Proceedings of the 12th International Conference on Implementation and Application of Automata (CIAA 2007)*. Ed. by Springer. Vol. 4783. Lecture Notes in Computer Science, Pages 11–23.

Andersson, Sven-Gunnar and Sigmund Kvam (1984). *Satzverschränkung im heutigen Deutsch: eine syntaktische und funktionale Studie unter Berücksichtigung alternativer Konstruktionen*. Vol. 24. Studien zur deutschen Grammatik. Tübingen: Gunter Narr.

Auer, Peter, Frans Hinskens, and Paul Kerswill, eds. (2005). *Dialect Change*. Cambridge: Cambridge University Press.

Auroux, Sylvain et al., eds. (2000). *History of the language sciences*. Berlin, New York: De Gruyter.

Bailey, Charles-James (1973). *Variation and Linguistic Theory*. Arlington: Center for Applied Linguistics.

Barbiers, Sjef, Leonie Cornips, and Susanne van der Kleij, eds. (2002). *Syntactic Microvariation*. Vol. II. Amsterdam: Meertens Institute Electronic Publications in Linguistics.

Barbiers, Sjef, Leonie Cornips, and Jan Pieter Kunst (2007). "The Syntactic Atlas of the Dutch Dialects (SAND): A Corpus of Elicited Speech and Text as an Online Dynamic Atlas". In: *Creating and digitizing language corpora*. Ed. by J. C. Beal, K. P. Corrigan, and H. L. Moisl. Vol. 1. Hampshire: Palgrave Macmillan, Pages 54–90.

Baron, Alistair, Paul Rayson, and Dawn Archer (2009). "Automatic Standardization of Spelling for Historical Text Mining". In: *Proceedings of Digital Humanities 2009*. University of Maryland.

Baur, Arthur (1983). *Was ist eigentlich Schweizerdeutsch?* Winterthur: Gemsberg.

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

Beesley, Kenneth R. and Lauri Karttunen (2003). *Finite State Morphology*. Stanford: CSLI Publications.

Benajiba, Yassine and Mona Diab (2010). "A web application for dialectal Arabic text annotation". In: *Proceedings of the LREC 2010 Workshop on Semitic Language Processing*. La Valletta, Malta, Pages 91–98.

Benincà, Paola and Cecilia Poletto (2007). "The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects". In: *Nordlyd* 34(1). Ed. by Kristine Bentzen and Øystein Alexander Vangsnes, Pages 35–52.

Berger, Claudia Bucheli and Elvira Glaser (2004). "Zur Morphologie des (ko)prädikativen Adjektivs und Partizips II im Alemannischen und Bairischen". In: *Morphologie und Syntax deutscher Dialekte und Historische Dialektologie des Deutschen*. Ed. by Franz Patocka and Peter Wiesinger. Wien: Edition Praesens.

Biadsy, Fadi, Julia Hirschberg, and Nizar Habash (2009). "Spoken Arabic Dialect Identification Using Phonotactic Modeling". In: *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. Athens, Greece, Pages 53–61.

Bickel, Hans and Robert Schläpfer, eds. (2000). *Die viersprachige Schweiz*. 2nd ed. Aarau: Sauerländer.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bohnenberger, Karl (1913). *Die Mundart der deutschen Walliser im Heimattal und in den Aussenorten*. Beiträge zur Schweizerdeutschen Grammatik 6. Frauenfeld: Huber.

— (1953). *Die alemannische Mundart: Umgrenzung, Innengliederung und Kennzeichnung*. Tübingen: J. C. B. Mohr.

Boitet, Christian (1997). "(Human-Aided) Machine Translation: A Better Future?" In: *Survey of the State of the Art in Human Language Technology*. Ed. by Ronald Cole et al. Pisa: Giardini Editori e Stampatori. Chapter 8.3.

Bouchard-Côté, Alexandre et al. (2007). "A Probabilistic Approach to Diachronic Phonology". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL 2007)*. Prague, Pages 887–896.

Brandner, Ellen and Martin Salzmann (2009). "Crossing the lake: Motion verb constructions in Bodensee-Alemannic and Swiss German." In: *Groninger Arbeiten zur germanistischen Linguistik* 48, Pages 81–113.

Brants, Sabine et al. (2002). "The TIGER Treebank". In: *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*. Sozopol, Bulgaria, Pages 24–41.

Brants, Thorsten (2000). "TnT – A Statistical Part-of-Speech Tagger". In: *Proceedings of the Sixth Applied Natural Language Processing Conference and First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL 2000)*. Seattle, USA, Pages 224–231.

Bresnan, Joan and Ronald Kaplan (1982). "The Mental Representation of Grammatical Relations". In: Cambridge: MIT Press. Chapter Lexical Functional Grammar: A Formal System for Grammatical Representation.

Brown, Peter F. et al. (1993). "The mathematics of statistical machine translation: parameter estimation". In: *Computational Linguistics* 19(2), Pages 263–311.

Bucheli, Claudia and Elvira Glaser (2002). "The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems". In: *Syntactic Microvariation*. Ed. by Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij. Vol. II. Amsterdam: Meertens Institute Electronic Publications in Linguistics.

Bucheli-Berger, Claudia (2006). "Syntaktische Raumbilder im Höchstalemannischen". In: *Raumstrukturen im Alemannischen*. Ed. by Hubert Klausmann. Graz/Feldkirch: Neugebauer, Pages 91–96.

Buchholz, Sabine and Erwin Marsi (2006). "CoNLL-X shared task on Multilingual Dependency Parsing". In: *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)*. New York City, Pages 149–164.

Burrough, Peter A. and Rachael A. McDonnell (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, et al. (2012). "Findings of the 2012 Workshop on Statistical Machine Translation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, Pages 10–51.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan (2011). "Findings of the 2011 Workshop on Statistical Machine Translation". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, Pages 22–64.

Carl, Michael and Andy Way, eds. (2003). *Recent Advances in Example-based Machine Translation*. Dordrecht: Kluwer Academic Publishers.

Cavnar, W. B. and J. M. Trenkle (1994). "N-gram based text categorization". In: *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.

Chambers, J. K. and Peter Trudgill (1980). *Dialectology*. 1st ed. Cambridge: Cambridge University Press.

— (1998). *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.

Chiang, David (2010). "Learning to translate with source and target syntax". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Uppsala, Sweden, Pages 1443–1452.

Chiang, David, Mona Diab, et al. (2006). "Parsing Arabic Dialects". In: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Compuational Linguistics (EACL 2006)*. Trento, Pages 369–376.

Chiang, David and Owen Rambow (2006). "The hidden TAG model: Synchronous grammars for parsing resource-poor languages". In: *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+8)*. Sydney, Australia, Pages 1–8.

Christen, Helen (1997). "Koiné-Tendenzen im Schweizerdeutschen?" In: *Varietäten des Deutschen – Regional- und Umgangssprachen*. Ed. by Gerhard Stickel. Berlin, New York: De Gruyter, Pages 346–363.

— (1998a). "Convergence and Divergence in the Swiss German Dialects". In: *Folia Linguistica* 32(1-2), Pages 53–67.

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

Christen, Helen, Elvira Glaser, and Matthias Friedli, eds. (2010). *Kleiner Sprachatlas der deutschen Schweiz*. Frauenfeld: Huber.

Christoph Merian Stiftung, ed. (2010). *Neues Baseldeutsch Wörterbuch*. Basel: Christoph Merian Verlag.

Clark, Alexander, Chris Fox, and Shalom Lappin, eds. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell.

Clopper, Cynthia G. and John C. Paolillo (2006). "North American English Vowels: A factor-analytic perspective". In: *Literary and Linguistic Computing, Special Issue on Progress in Dialectometry* 21(4), Pages 445–462.

Cole, Ronald et al., eds. (1997). *Survey of the State of the Art in Human Language Technology*. Pisa: Giardini Editori e Stampatori.

Collins, Michael (1997). "Three Generative, Lexicalised Models for Statistical Parsing". In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*. Madrid, Spain, Pages 16–23.

Collins, Michael, Philipp Koehn, and Ivona Kučerová (2005). "Clause restructuring for statistical machine translation". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor, USA, Pages 531–540.

Cooper, Kathrin (1995). "Null Subjects and Clitics in Zurich German". In: *Topics in Swiss German Syntax*. Ed. by Zvi Penner. Bern: Peter Lang, Pages 59–72.

Cox, Trevor F. and Michael A.A. Cox (1994). *Multidimensional scaling*. London, Glasgow: Chapman & Hall.

den Besten, Hans and Jean Rutten (1989). "On verb raising, extraposition and free word order in Dutch". In: *Sentential complementation and the lexicon*. Ed. by Dany Jaspers. Dordrecht: Foris, Pages 41–56.

Diab, Mona et al. (2010). "COLABA: Arabic dialect annotation and processing". In: *Proceedings of the LREC 2010 Workshop on Semitic Language Processing*. La Valletta, Malta, Pages 66–74.

Dieth, Eugen (1986). *Schwyzertütschi Dialäktschrift*. Ed. by Christian Schmid-Cadalbert. 2nd ed. Aarau: Sauerländer.

Dubey, Amit (2005). "What to do when lexicalization fails: parsing German with suffix analysis and smoothing". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor, USA, Pages 314–321.

Eisenstein, Jacob et al. (2010). "A latent variable model for geographic lexical variation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. Cambridge, MA, USA, Pages 1277–1287.

Embleton, Sheila (1993). "Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project". In: *Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics (QUALICO)*. Ed. by Reinhard Köhler and Burghard B. Rieger. Dordrecht/Boston/London: Kluwer Academic Publishers, Pages 267–276.

Engel, Ulrich (1998). *Deutsche Grammatik*. Heidelberg: J. Groos.

Evers, Arnold (1975). *The transformational cycle in Dutch and German*. Bloomington, Indiana: Indiana University Linguistics Club.

Ferguson, Charles A. (1959). "Diglossia". In: *Word* (15), Pages 325–340.

Fischer, Ludwig (1989). *Luzerndeutsche Grammatik. Ein Wegweiser zur guten Mundart*. Ed. by Walter Haas. 2nd ed. Hitzkirch: Comenius.

Forst, Martin (2002). "La traduction automatique dans le cadre formel de la LFG – Un système de traduction entre l'allemand standard et le zurichois". In: *Publications du Centre de traduction littéraire de Lausanne (CTL)*. Vol. 41. Université de Lausanne, Switzerland.

Foth, Kilian, Michael Daum, and Wolfgang Menzel (2004). "A broad-coverage parser for German based on defeasible constraints". In: *Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2004)*. Vienna, Austria, Pages 45–52.

Fraser, Alexander (2009). "Experiments in morphosyntactic processing for translating to and from German". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, Pages 115–119.

Friedli, Matthias (2006). "Der Komparativanschluss im Schweizerdeutschen – ein raumbildendes Phänomen". In: *Raumstrukturen im Alemannischen*. Ed. by Hubert Klausmann. Graz/Feldkirch: Neugebauer, Pages 103–108.

Glaser, Elvira (2003). "Schweizerdeutsche Syntax. Phänomene und Entwicklungen". In: *Gömmer MiGro? Veränderungen und Entwicklungen im heutigen SchweizerDeutschen*. Germanistica Friburgensia 18. Freiburg, Schweiz: Universitätsverlag Freiburg, Pages 39–66.

Gleßgen, Martin-D. and Achim Stein (2005). "Resources and tools for analyzing Old French texts". In: *Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft / Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Ed. by Claus D. Pusch, Johannes Kabatek, and Wolfgang Raible. Tübingen: Gunter Narr, Pages 135–145.

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

— (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. 3 vol. Tübingen: Max Niemeyer.

Goebl, Hans and Guillaume Schiltz (2000). "Neuere Entwicklungen in der europäischen Dialektologie (1950-2000)". In: *History of the language sciences*. Ed. by Sylvain Auroux et al. Vol. 3. Berlin, New York: De Gruyter, Pages 2352–2365.

Gooskens, Charlotte and Wilbert Heeringa (2004). "Perceptual evaluation of Levenshtein dialect distance measurements using Norwegian dialect data". In: *Language Variation and Change* 16(3), Pages 189–207.

Grangier, Louis (1864). *Glossaire fribourgeois, ou Recueil des locutions vicieuses usitées dans le canton de Fribourg*. Fribourg: C. Clerc.

Greyerz, Otto von and Ruth Bietenhard (1976). *Berndeutsches Wörterbuch für die heutige Mundart zwischen Burgdorf, Lyss und Thun*. Bern: Francke.

Grieve, Jack (to appear). "A comparison of statistical methods for the aggregation of regional lingustic variation". In: *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages*. Ed. by Benedikt Szmrecsanyi and Bernhard Wälchli.

Haas, Walter (2000). "Die deutschsprachige Schweiz". In: *Die viersprachige Schweiz*. Ed. by Hans Bickel and Robert Schläpfer. 2nd ed. Aarau: Sauerländer, Pages 57–138.

Hägi, Sara and Joachim Scharloth (2005). "Ist Standarddeutsch für Deutschschweizer eine Fremdsprache? Untersuchungen zu einem Topos des sprachreflexiven Diskurses". In: *Linguistik online* 24. http://www.linguistik-online.de/24_05/haegiScharloth.html.

Haiman, John and Paola Benincà (1992). *The Rhaetoromance Languages*. Routledge, London, New York.

Hajic, Jan (2000). "Morphological Tagging: Data vs. Dictionaries". In: *Proceedings of the Sixth Applied Natural Language Processing Conference and First Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL 2000)*. Seattle, USA, Pages 94–101.

Hall, Johan and Joakim Nivre (2008). "A dependency-driven parser for German dependency and constituency representations". In: *Proceedings of the ACL 2008 Workshop on Parsing German (PaGe 2008)*. Columbus, Ohio, USA, Pages 47–54.

Hammarström, Harald and Lars Borin (2011). "Unsupervised Learning of Morphology". In: *Computational Linguistics* 37(2). to appear.

Heeb, Dominik. *Chochichästli-Orakel*. http://dialects.from.ch.

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

Heeringa, Wilbert, Peter Kleiweg, et al. (2006). "Evaluation of String Distance Algorithms for Dialectology". In: *Proceedings of the ACL 2006 Workshop on Linguistic Distances*. Sydney, Australia, Pages 51–62.

Heeringa, Wilbert and John Nerbonne (2001). "Dialect Areas and Dialect Continua". In: *Language Variation and Change* (13), Pages 375–400.

Herrgen, Joachim (2000). "Die Dialektologie des Deutschen". In: *History of the language sciences*. Ed. by Sylvain Auroux et al. Vol. 2. Berlin, New York: De Gruyter, Pages 1513–1535.

Hinrichs, Erhard W., Julia Bartels, et al. (2000). "The Tübingen Treebanks for Spoken German, English, and Japanese". In: *Verbmobil: Foundations of Speech-to-Speech Translation*. Ed. by Wolfgang Wahlster. Berlin: Springer.

Hinrichs, Erhard W., Sandra Kübler, et al. (2004). "Recent developments in linguistic annotations of the TüBa-D/Z treebank". In: *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany, Pages 51–62.

Höfler, Stefan and Michael Piotrowski (2011). "Building Corpora for the Philological Study of Swiss Legal Texts". In: *Journal for Language Technology and Computational Linguistics* 26(2), Pages 77–88.

Hopcroft, John E., Rajeev Motwani, and Jefrey D. Ullman (2006). *Introduction to Automata Theory, Languages and Computation*. 3rd ed. Boston: Pearson Addison-Wesley.

Hotzenköcherle, Rudolf (1962). *Einführung in den Sprachatlas der deutschen Schweiz. A: Zur Methodologie der Kleinraumatlanten*. Bern: Francke.

— (1984). *Die Sprachlandschaften der deutschen Schweiz*. Ed. by Niklaus Bigler and Robert Schläpfer. Aarau: Sauerländer.

Hotzenköcherle, Rudolf and Rudolf Brunner, eds. (1972-1976). *SDS Phonogramme: Begleittexte zu den Tonaufnahmen für den Sprachatlas der deutschen Schweiz*. 4 vols. Bern.

Hotzenköcherle, Rudolf, Robert Schläpfer, et al., eds. (1962-1997). *Sprachatlas der deutschen Schweiz*. 8 vols. Bern: Francke.

Hughes, Baden et al. (2006). "Reconsidering Language Identification for Written Language Resources". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Hulden, Mans (2009). "Foma: a finite-state toolkit and library". In: *Proceedings of EACL 2009*. Athens, Greece.

Ingle, N. (1980). "A Language Identification Table". In: *Technical Translation International*.

Institut für Deutsche Sprache, Programmbereich Korpuslinguistik (2007). *Korpusbasierte Wortformenliste (bzw. Grundformenliste) DEREWO, v-30000g-2007-12-31-0.1, mit Benutzerdokumentation*. Mannheim. http://www.ids-mannheim.de/kl/derewo.

Jaberg, Karl and Jakob Jud, eds. (1928-1960). *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen: Ringier.

Jain, Anil K. and Richard C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Jansche, Martin (2001). "Re-Engineering Letter-to-Sound Rules". In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*. Pittsburgh, PA, USA, Pages 111–117.

Jiampojamarn, Sittichai, Grzegorz Kondrak, and Tarek Sherif (2007). "Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion". In: *Human Language Technologies: Proceedings of the Sixth Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*. Rochester, NY, USA, Pages 372–379.

Jones, William J. (2000). "Early dialectology, etymology and language history in German-speaking countries". In: *History of the language sciences*. Ed. by Sylvain Auroux et al. Vol. 2. Berlin, New York: De Gruyter, Pages 1105–1115.

Jörg, Ruth (1976). *Untersuchungen zum Schwund des Präteritums im Schweizerdeutschen*. Basler Studien zur deutschen Sprache und Literatur 52. Bern: Francke.

Kelle, Bernhard (2001). "Zur Typologie der Dialekte in der deutschsprachigen Schweiz: Ein dialektometrischer Versuch". In: *Dialectologia et Geolinguistica* 9, Pages 9–34.

Kessler, Brett (1995). "Computational Dialectology in Irish Gaelic". In: *Proceedings of the Seventh Conference of the European Chapter of the Association for Compuational Linguistics (EACL 1995)*. Dublin, Ireland, Pages 60–66.

Klausmann, Hubert, ed. (2006). *Raumstrukturen im Alemannischen*. Graz/Feldkirch: Neugebauer.

Knecht, Pierre (2000). "Die französischsprachige Schweiz". In: *Die viersprachige Schweiz*. Ed. by Hans Bickel and Robert Schläpfer. 2nd ed. Aarau: Sauerländer, Pages 139–176.

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

Koehn, Philipp and Kevin Knight (2002). "Learning a Translation Lexicon from Monolingual Corpora". In: *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*. Philadelphia, PA, Pages 9–16.

— (2003). "Empirical Methods for Compound Splitting". In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*. Budapest, Hungary, Pages 187–193.

Kondrak, Grzegorz and Tarek Sherif (2006). "Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification". In: *Proceedings of the ACL 2006 Workshop on Linguistic Distances*. Sydney, Australia, Pages 43–50.

König, Werner (2007). *dtv-Atlas Deutsche Sprache*. 15th ed. München: Deutscher Taschenbuch Verlag.

Kristol, Andres M. (2005). "Introduction à la géographie linguistique". Lecture notes. University of Neuchâtel.

Kruskal, J. (1999). "An overview of sequence comparison: Time warps, string edits and macromolecules". In: *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Ed. by David Sankoff and Joseph Kruskal. Stanford: CSLI Press, Pages 1–44.

Kübler, Sandra (2008). "The PaGe 2008 Shared Task on Parsing German". In: *Proceedings of the ACL 2008 Workshop on Parsing German (PaGe 2008)*. Columbus, Ohio, Pages 55–63.

Kübler, Sandra, Erhard W. Hinrichs, and Wolfgang Maier (2006). "Is it really that difficult to parse German?" In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*. Sydney, Australia, Pages 111–119.

Labov, William (1966). *The social stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

Landolt, Christoph (2007). "Neuere Entwicklungen in der historischen Dialektlexikographie des Deutschen". In: *Lexicographica* 23, Pages 151–172.

Lavie, Alon and Abhaya Agarwal (2007). "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments". In: *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*. Prague, Czech Republic.

Lee, Jay and William A. Kretzschmar Jr. (1993). "Spatial analysis of linguistic data with GIS functions". In: *International Journal of Geographical Information Systems* 7(6), Pages 541–560.

Lehmann, Winfred P. (1992). *Historical Linguistics*. 3rd ed. London, New York: Routledge.

Léon, Jacqueline (2000a). "La traduction automatique I: les premières tentatives jusqu'au rapport ALPAC". In: *History of the language sciences*. Ed. by Sylvain Auroux et al. Vol. 3. Berlin, New York: De Gruyter, Pages 2767–2774.

— (2000b). "La traduction automatique II: développements récents". In: *History of the language sciences*. Ed. by Sylvain Auroux et al. Vol. 3. Berlin, New York: De Gruyter, Pages 2774–2780.

Levenshtein, Vladimir I. (1966). "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals". In: *Soviet Physics Doklady* 10(8), Pages 707–710.

Lindén, Krister, Miikka Silfverberg, and Tommi Pirinen (2009). "HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers". In: *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009)*. Zurich, Switzerland.

Liver, Ricarda (1999). *Rätoromanisch: Eine Einführung in das Bündnerromanische*. Gunter Narr, Tübingen.

Lötscher, Andreas (1978). "Zur Verbstellung im Zürichdeutschen und in anderen Varianten des Deutschen". In: *Zeitschrift für Dialektologie und Linguistik* 45, Pages 1–29.

— (1983). *Schweizerdeutsch. Geschichte, Dialekte, Gebrauch*. Frauenfeld: Huber.

Lüdi, Georges et al. (1997). *Le paysage linguistique de la Suisse. Recensement fédéral de la population 1990*. Bern: Office fédéral de la statistique.

Lurati, Ottavio (2000). "Die sprachliche Situation in der Südschweiz". In: *Die viersprachige Schweiz*. Ed. by Hans Bickel and Robert Schläpfer. 2nd ed. Aarau: Sauerländer, Pages 177–210.

Mackay, Wesley and Grzegorz Kondrak (2005). "Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005)*. Ann Arbor, Pages 40–47.

Mahlow, Cerstin and Michael Piotrowski (2009a). "A Target-Driven Evaluation of Morphological Components for German". In: *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th Birthday*. Münster: Simon Clematide, Manfred Klenner, and Martin Volk, Pages 85–99.

— eds. (2009b). *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009)*. Vol. 41. Communications in Computer and Information Science. Zurich, Switzerland: Springer.

Mann, Gideon S. and David Yarowsky (2001). "Multipath translation lexicon induction via bridge languages". In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*. Pittsburgh, PA, USA, Pages 151–158.

Marti, Werner (1972). *Bärndütschi Schrybwys*. Bern, Switzerland: Francke.

— (1985). *Berndeutsch-Grammatik für die heutige Mundart zwischen Thun und Jura*. Bern, Switzerland: Francke.

McEnery, Tony and Richard Xiao (2008). "Parallel and Comparable Corpora: What is Happening?" In: *Incorporating Corpora: The Linguist and the Translator*. Ed. by Gunilla Anderman and Margaret Rogers. Multilingual Matters. Chapter 2, Pages 18–31.

Mitchell, Tom (1997). *Machine learning*. McGraw Hill.

Moeschler, Jacques (2007). "Introduction to Semantics". In: *Speech and Language Engineering*. Ed. by Martin Rajman. Lausanne: EPFL Press. Chapter 3.

Müller, Martin et al. (2009). *Chunsch druus? Schweizerdeutsch verstehen – die Deutschschweiz verstehen*. Bern: Schulverlag blmv.

Nerbonne, John, Rinke Colen, et al. (2011). "Gabmap – A web application for dialectology". In: *Dialectologia, Special Issue* (II), Pages 65–89.

Nerbonne, John and Wilbert Heeringa (2001). "Computational Comparison and Classification of Dialects". In: *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*. Ed. by Wolfgang Viereck and Hedwig Gwosdek. Vol. 9. Alessandria: Edizioni dell'Orso, Pages 69–83.

Nerbonne, John, Wilbert Heeringa, and Peter Kleiweg (1999). "Edit Distance and Dialect Proximity". In: *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Ed. by David Sankoff and Joseph Kruskal. Stanford: CSLI Press, Pages v–xv.

Nerbonne, John and Peter Kleiweg (2005). "Toward a Dialectological Yardstick". In: *Journal of Quantitative Linguistics* (5).

Nerbonne, John, Peter Kleiweg, et al. (2008). "Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering". In: *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*. Ed. by Christine Preisach et al. Berlin: Springer, Pages 647–654.

Nerbonne, John and Christine Siedle (2005). "Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede". In: *Zeitschrift für Dialektologie und Linguistik* 72(5), Pages 129–147.

Niebaum, Hermann and Jürgen Macha (2006). *Einführung in die Dialektologie des Deutschen*. 2nd ed. Tübingen: Niemeyer.

Nießen, Sonja and Hermann Ney (2004). "Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information". In: *Computational Linguistics* 30(2), Pages 181–204.

Nübling, Damaris (1992). *Klitika im Deutschen*. Tübingen, Germany: Gunter Narr.

— (1995a). "Die Kurzverben im Schweizerdeutschen. In der Kürze liegt die Würze oder: Im Spannungsfeld zwischen Reduktion und Differenzierung". In: *Alemannische Dialektforschung.*

Bilanz und Perspektiven. Ed. by Heinrich Löffler. Vol. 68. Basler Studien zur deutschen Sprache und Literatur. Francke, Pages 165–180.

— (1995b). "Kurzverben in germanischen Sprachen: Unterschiedliche Wege – gleiche Ziele". In: *Zeitschrift für Dialektologie und Linguistik (ZDL)* LXII(2), Pages 127–154.

Osenova, Petya, Wilbert Heeringa, and John Nerbonne (2009). "A quantitative analysis of Bulgarian dialect pronunciation". In: *Zeitschrift für Slavische Philologie* 66(2), Pages 425–458.

Papineni, Kishore et al. (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. Philadelphia, PA, Pages 311–318.

Paul, Hermann (2007). *Mittelhochdeutsche Grammatik*. Ed. by Heinz Peter Prell. 25th ed. Tübingen: Niemeyer.

Penner, Zvi, ed. (1995). *Topics in Swiss German Syntax*. Bern: Peter Lang.

Penner, Zvi and Thomas Bader (1995). "Issues in the Syntax of Subordination: A Comparative Study of the Complementizer System in Germanic, Romance, and Semitic Languages with Special Reference to Bernese Swiss German". In: *Topics in Swiss German Syntax*. Ed. by Zvi Penner. Bern: Peter Lang, Pages 73–290.

Petrov, Slav and Dan Klein (2008). "Parsing German with language agnostic latent variable grammars". In: *Proceedings of the ACL 2008 Workshop on Parsing German (PaGe 2008)*. Columbus, Ohio, USA, Pages 33–39.

Phonogrammarchiv der Universität Zürich (1952). *Der sprechende Atlas: Plattentext in verschiedenen schweizerdeutschen Dialekten*. Zürich: Phonogrammarchiv.

Piotrowski, Michael and Cathrin Senn (2012). "Harvesting Indices to Grow a Controlled Vocabulary: Towards Improved Access to Historical Legal Texts". In: *Proceedings of the EACL 2012 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012)*. Ed. by Kalliopi Zervanou and Antal van den Bosch. Avignon, France, Pages 24–29.

Pollard, Carl and Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Popovic, Maja, Daniel Stein, and Hermann Ney (2006). "Statistical Machine Translation of German Compound Words". In: *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL)*. Turku, Finland, Pages 616–624.

Pucher, Michael et al. (2010). "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis". In: *Speech Communication* 52(2), Pages 164–179.

Rafferty, Anna and Christopher Manning (2008). "Parsing three German treebanks: Lexicalized and unlexicalized baselines". In: *Proceedings of the ACL 2008 Workshop on Parsing German (PaGe 2008)*. Columbus, Ohio, USA, Pages 40–46.

Rambow, Owen (2003). "Coherent constructions in German: Lexicon or syntax?" In: *Verb Constructions in German and Dutch*. Ed. by Pieter A. M. Seuren and Gerard Kempen. Benjamins.

Rambow, Owen et al. (2006). *Parsing Arabic Dialects*. Tech. rep. John Hopkins Summer Workshop.

Rash, Felicity (1998). *The German Language in Switzerland. Multilingualism, Diglossia, and Variation*. Bern: Peter Lang.

Řehůřek, Radim and Milan Kolkus (2009). "Language Identification on the Web: Extending the Dictionary Method". In: *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2009)*. Springer, Pages 357–368.

Rumpf, Jonas et al. (2009). "Structural analysis of dialect maps using methods from spatial statistics". In: *Zeitschrift für Dialektologie und Linguistik* 76(3), Pages 280–308.

— (2010). "Quantification and statistical analysis of structural similarities in dialectological area-class maps." In: *Dialectologia et Geolinguistica* 18, Pages 73–98.

Ruoff, Arno (1981). *Häufigkeitswörterbuch gesprochener Sprache*. Tübingen: Niemeyer.

Russ, Charles V. J. (1990). "High Alemannic". In: *The dialects of modern German: a linguistic survey*. Ed. by Charles V. J. Russ. London: Routledge. Chapter 12, Pages 364–393.

Salzmann, Martin and Guido Seiler (2011). "Variation as the exception or the rule? Swiss relatives, revisited". In: *Empirical Issues in Syntax and Semantics* (8), Pages 453–485.

Sankoff, David and Joseph Kruskal, eds. (1999). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Stanford: CSLI Press.

Saussure, Ferdinand de (1916). *Cours de Linguistique Générale*. Paris: Payot.

Scharloth, Joachim (2004). "Zwischen Fremdsprache und nationaler Varietät. Untersuchungen zum Plurizentrizitätsbewusstsein der Deutschschweizer". In: *TRANS. Internet-Zeitschrift für Kulturwissenschaften* 15. http://www.inst.at/trans/15Nr/06_1/scharloth15.htm.

Scherrer, Yves (2007a). "Adaptive String Distance Measures for Bilingual Dialect Lexicon Induction". In: *Proceedings of the ACL 2007 Student Research Workshop*. Prague, Pages 55–60.

— (2007b). "Phonetic Distance Measures for the Induction of a Translation Lexicon for Dialects". Mémoire de DEA.

— (2008a). "Part-of-Speech Tagging with a Symbolic Full Parser: Using the TIGER Treebank to Evaluate Fips". In: *Proceedings of the ACL 2008 Workshop on Parsing German (PaGe 2008)*. Columbus, Ohio, USA, Pages 16–23.
— (2008b). "Transducteurs à fenêtre glissante pour l'induction lexicale". In: *Actes de la 12e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues (RECITAL 2008)*. Avignon, France.
— (2010). "Des cartes dialectologiques numérisées pour le TALN". In: *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*. Montréal.
— (2011a). "Morphology generation for Swiss German Dialects". In: *Systems and Frameworks for Computational Morphology, Second International Workshop (SFCM 2011)*. Ed. by Cerstin Mahlow and Michael Piotrowski. Berlin, Heidelberg: Springer, Pages 130–140.
— (2011b). "Syntactic transformations for Swiss German dialects". In: *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Edinburgh, Scotland, Pages 30–38.
— (2012). "Recovering dialect geography from an unaligned comparable corpus". In: *Proceedings of the EACL 2012 Workshop on Visualization of Language Patters and Uncovering Language History from Multilingual Resources*. Avignon.

Scherrer, Yves and Bruno Cartoni (2012). "The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.

Scherrer, Yves and Owen Rambow (2010). "Word-based dialect identification with georeferenced rules". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. Cambridge, MA, USA.

Schmid, Helmut (1995). "Improvements in Part-of-Speech Tagging with an Application to German". In: *Proceedings of the EACL 1995 SIGDAT Workshop: From Text to Tags*. Dublin, Ireland.
— (2004). "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors". In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland, Pages 162–168.
— (2005). "A Programming Language for Finite State Transducers". In: *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005)*. Helsinki, Finland, Pages 308–309.

Schmid, Helmut (2006). "Trace prediction and recovery with unlexicalized PCFGs and slash features". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*. Sydney, Australia, Pages 177–184.

Schmid, Helmut, Arne Fitschen, and Ulrich Heid (2004). "SMOR: A German computational morphology covering derivation, composition, and inflection". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal, Pages 1263–1266.

Schönenberger, Manuela (1995). "Constituent Order in the VP: Verb Raising and Verb Projection Raising". In: *Topics in Swiss German Syntax*. Ed. by Zvi Penner. Bern: Peter Lang, Pages 347–411.

Schönenberger, Manuela and Zvi Penner (1995). "Probing Swiss-German Clause Structure by means of the Placement of Verbal Expletives: *tun* "do" Insertion and Verb Doubling". In: *Topics in Swiss German Syntax*. Ed. by Zvi Penner. Bern: Peter Lang.

Séguy, Jean (1973a). "La dialectométrie dans l'Atlas linguistique de la Gascogne". In: *Revue de linguistique romane* 37, Pages 1–24.

— (1973b). "La fonction minimale du dialecte". In: *Les dialectes romans de France à la lumière des atlas régionaux*. Ed. by G. Straka and P. Gardette. Paris, Pages 27–37.

Seiler, Guido (2002). "Prepositional dative marking in Upper German: a case of syntactic microvariation". In: *Syntactic Microvariation*. Ed. by Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij. Vol. II. Amsterdam: Meertens Institute Electronic Publications in Linguistics.

— (2004). "Gibt es einen Alpensprachbund?" In: *Alemannisch im Sprachvergleich*. Ed. by Elvira Glaser, Peter Ott, and Rudolf Schwarzenbach. Wiesbaden: Steiner, Pages 485–493.

Sennrich, Rico et al. (2009). "A New Hybrid Dependency Parser for German". In: *Von der Form zur Bedeutung: Texte automatisch verarbeiten – From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*. Ed. by C. Chiarcos, R. E. de Castilho, and M. Stede. Tübingen: Narr, Pages 115–124.

Shackleton Jr., Robert G. (2007). "Phonetic Variation in the Traditional English Dialects". In: *Journal of English Linguistics* 35(1), Pages 30–102.

Sherif, Tarek and Grzegorz Kondrak (2007). "Substring-Based Transliteration". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*. Prague, Czech Republic, Pages 944–951.

Shieber, Stuart M. (1985). "Evidence against the context-freeness of natural language". In: *Linguistics and Philosophy* 8, Pages 333–343.

Siebenhaar, Beat (2005). "Die dialektale Verankerung regionaler Chats in der deutschsprachigen Schweiz". In: *Moderne Dialekte – Neue Dialektologie*. Ed. by Eckhard Eggers, Jürgen Erich Schmidt, and Dieter Stellmacher. Stuttgart: Steiner, Pages 691–717.

Siebenhaar, Beat, Martin Forst, and Eric Keller (2006). "Speech synthesis of dialectal variants as a method for research on prosody". In: *Topics in Dialectal Variation. Joensuu: Joensuun yliopistopaino (Studies in Languages 40)*. Ed. by Markku Filippula et al., Pages 145–162.

Siebenhaar, Beat and Walter Voegeli (1997). "Mundart und Hochdeutsch im Vergleich". http://www.uni-leipzig.de/~siebenh/pdf/Siebenhaar_Voegeli_iPr.pdf.

Simard, Michel, George F. Foster, and Pierre Isabelle (1992). "Using Cognates to Align Sentences in Bilingual Corpora". In: *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*. Montréal, Canada, Pages 67–81.

Skut, Wojciech et al. (1997). "An annotation scheme for free word order languages". In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP 1997)*. Washington DC, Pages 88–95.

Snover, Matthew et al. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*. Boston, MA, Pages 223–231.

Sokal, Robert R. and F. James Rohlf (1995). *Biometry: the principles and practice of statistics in biological research*. 3rd ed. New York: W.H. Freeman.

Stähli, Adrian, Christa Dürscheid, and Marie-José Béguelin (2011). "sms4science: Korpusdaten, Literaturüberblick und Forschungsfragen". In: *Linguistik online* 48(4).

Staub, Friedrich (1877). "Ein schweizerisch-alemannisches Lautgesetz". In: *Frommanns Deutsche Mundarten* 7 (NF 1), Pages 18–35, 191–207, 333–389.

Staub, Friedrich et al., eds. (1881-). *Schweizerisches Idiotikon : Wörterbuch der schweizerdeutschen Sprache*. Frauenfeld: Huber.

Steiner, Janine (2006). "Syntaktische Variation in der Nominalphrase – ein Fall für die Dialektgeographin oder den Soziolinguisten?" In: *Raumstrukturen im Alemannischen*. Ed. by Hubert Klausmann. Graz/Feldkirch: Neugebauer, Pages 109–115.

Stolcke, Andreas (2002). "SRILM – an extensible language modeling toolkit". In: *Proceedings of the 7th International Conference on Spoken Language Processing (Interspeech 2002)*, Pages 901–904.

Suter, Rudolf (1976). *Baseldeutsch-Grammatik*. 2nd ed. Basel: C. Merian.

— (1984). *Baseldeutsch-Wörterbuch*. Basel: C. Merian.

Thielen, C. et al. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Tech. rep. University of Stuttgart and University of Tübingen.

Tiedemann, Jörg (2009). "Character-based PSMT for Closely Related Languages". In: *Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*. Barcelona, Pages 12–19.

Trüb, Rudolf (2003). *Sprachatlas der deutschen Schweiz, Abschlussband*. Bern/Basel: Francke.

Ulmann, Martin (1995). "Decomposing German Compound Nouns". In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 1995)*. Tzigov Chark, Bulgaria, Pages 265–270.

Uszkoreit, Hans and Annie Zaenen (1997). "Grammar Formalisms". In: *Survey of the State of the Art in Human Language Technology*. Ed. by Ronald Cole et al. Pisa: Giardini Editori e Stampatori. Chapter 3.3.

Vaillant, Pascal (2008a). "A Layered Grammar Model: Using Tree-Adjoining Grammars to Build a Common Syntactic Kernel for Related Dialects". In: *Proceedings of the Ninth International Workshop on Tree Adjoining Grammar and Related Formalisms (TAG+9)*. Tübingen, Pages 157–164.

— (2008b). "Grammaires factorisées pour des dialectes apparentés". In: *Actes de la 15e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*. Avignon, Pages 159–168.

Vauquois, Bernard (1976). "Automatic Translation – A Survey of Different Approaches". In: *Proceedings of the 6th International Conference on Computational Linguistics (COLING 1976)*.

Veith, Werner H. (1970). *-Explikative +Applikative +Komputative Dialektkartographie*. Vol. 4. Germanistische Linguistik. Olms.

— (1982). "Theorieansätze einer generativen Dialektologie". In: *Dialektologie – Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Ed. by Werner Besch et al. Berlin, New York: De Gruyter, Pages 277–295.

Vijay-Shanker, K. and Aravind Joshi (1988). "Feature-structure based Tree Adjoining Grammars". In: *Proceedings of the 12th International Conference on Computational Linguistics (COLING 1988)*. Budapest, Pages 714–719.

Vilar, David, Jan-Thorsten Peter, and Hermann Ney (2007). "Can We Translate Letters?" In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Pages 33–39.

Volk, Martin et al. (2010). "Challenges in Building a Multilingual Alpine Heritage Corpus". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. La Valletta, Malta.

Wandl-Vogt, Eveline (2006a). "Mapping Dialects – Die Karte als primäre Zugriffsstruktur für Dialektwörterbücher". In: *Wiener Schriften zur Geographie und Kartographie* 17.

— (2006b). "Von der Karte zum Wörterbuch – Überlegungen zu einer räumlichen Zugriffsstruktur für Dialektwörterbücher". In: *Proceedings of the 12th Congress of the European Association for Lexicography (EURALEX 2006)*. Torino, Italia, Pages 721–732.

— (2008). "An der Schnittstelle von Dialektwörterbuch und Sprachatlas: Das Projekt "Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)"". In: *Germanistische Linguistik 190-191. Sprachgeographie digital. Die neue Generation der Sprachatlanten*. Ed. by Stephan Elspaß and Werner König. Hildesheim: Olms, Pages 197–212.

Weber, Albert (1948). *Zürichdeutsche Grammatik. Ein Wegweiser zur guten Mundart*. Zürich: Schweizer Spiegel.

Weber, Albert and Jacques M. Bächtold (1961). *Zürichdeutsches Wörterbuch für Schule und Haus*. Zürich: Schweizer Spiegel.

Weber, David J. and William C. Mann (1979). "Prospects for Computer-assisted dialect adaption". In: *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics (ACL 1979)*, Pages 109–110.

Weddige, Hilkert (2007). *Mittelhochdeutsch: Eine Einführung*. 7th ed. München: C. H. Beck.

Wehrli, Eric (1997). *L'analyse syntaxique des langues naturelles: problèmes et méthodes*. Paris: Masson.

— (2007). "Fips, a "Deep" Linguistic Multilingual Parser". In: *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*. Prague, Pages 120–127.

Wehrli, Eric, Luka Nerima, and Yves Scherrer (2009). "Deep linguistic multilingual translation and bilingual dictionaries". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece, Pages 90–94.

Weinreich, Uriel (1954). "Is a structural dialectology possible?" In: *Word* 10, Pages 388–400.

Wolfensberger, Heinz (1967). *Mundartwandel im 20. Jahrhundert. Dargestellt an Ausschnitten aus dem Sprachleben der Gemeinde Stäfa*. Beiträge zur Schweizerdeutschen Mundartforschung 14. Frauenfeld: Huber.

Yzaguirre, Lluís de et al. (2000). "Some Technical Aspects about Aligning Near Languages". In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*. Athens, Greece, Pages 545–548.

Zielinski, Andrea and Christian Simon (2009). "Morphisto – An Open-source Morphological Analyzer for German". In: *Proceedings of the Seventh International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2008)*. Ispra, Italy, Pages 224–231.

Zifonun, Gisela et al. (1997). *Grammatik der deutschen Sprache*. Berlin, New York: De Gruyter.

Zue, Victor and Ronald Cole (1997). "Spoken Language Input". In: *Survey of the State of the Art in Human Language Technology*. Ed. by Ronald Cole et al. Pisa: Giardini Editori e Stampatori. Chapter 1.