Conference Presentation

Natural Language Processing for Swiss German Dialects

SCHERRER, Yves

Abstract

Most Natural Language Processing (NLP) applications focus on standardized, written language varieties. From a practical point of view, this focus is understandable: such systems are most likely to be used on written data of a standard variety, and this kind of data is also most easily available for training and parametrizing NLP systems. However, in many regions of the world, the linguistic reality is somewhat more complex: many speakers use some kind of non-standard language variety -- mostly in speech, but sometimes also in writing. Non-standard lects are subject to continuous variation along the dialectal and sociolinguistic level. From a methodological as well as a practical point of view, it is therefore interesting to include findings of variational linguistics in existing NLP methods. Our work focuses on Swiss German dialects. The German-speaking part of Switzerland has been subject to more than a century of dialectological research that has resulted in dialect atlases, grammars and lexicons. Today, the dialects represent the default variety of oral communication (Standard German is only used for writing). [...]
Natural Language Processing for Swiss German Dialects

Yves Scherrer

Université de Genève, Geneva, Switzerland
Columbia University, New York (2009-2010)

Joint work with Owen Rambow at Columbia University

4/16/2010
Outline

1. Introduction: Swiss German
2. Computational Linguistics and Dialectology
3. Three applications for Swiss German dialects
4. There’s a map for that!
5. Outlook
Languages in Switzerland

**Standard German / Swiss German**

**Diglossia**

**French**

**Italian**

**Romansh**

http://www.bfs.admin.ch/bfs/portal/de/index/regionen/thematische_karten/maps/bevoelkerung/sprachen_religionen.html
Variation in Swiss German dialects

Standard German:

Warst du auf dem Markt einkaufen?
Were you at the market shopping?
‘Have you been shopping at the market?’

Swiss German Dialects:

ZH  Bisch uf em Mèrt gsy go poschte?
SZ  Bisch ufem Märcht gsi go poschte?
SO  Besch ofem Märed gsy go boschte?
OW  Bischt ufm Märt gsi go ichoifä?
SG  Bisch uf dä Märt go poschtä?
BE  Bisch uf e Märit ga kömerle?
FR  Büschu z’Määret gsi ga iichufe?
WS  Bisch dü uf um Markt gsii ga ichöifu?

Differences between Standard German and Swiss German:

- Preterite tense replaced by present perfect
- go/ga + infinitive construction

Differences within Swiss German dialects:

- No standardized spelling
- Phonetics
- Lexicon
- Morphology
- Syntax

http://als.wikipedia.org/wiki/Alemannischer_Beispielsatz
Research questions

- How can we present the results of dialectological research in a more intuitive and dynamic way?
  - Research on Swiss German dialects since 100 years, but not easily accessible to laymen.
  - A computational approach can give new insights into interdialectal variation, dialect change, . . .

- How can we apply methods of computational linguistics / natural language processing to dialect data?
  - Few work on language variation (except in speech processing).
  - Reuse and adaptation of resources for Standard German.
  - Explicit dialectological data (rather than large corpora of raw text) calls for a knowledge-based approach.
Examples of computational linguistics applications:
Speech recognition systems, text-to-speech synthesizers, automated voice response systems, machine translation systems, parsers, part-of-speech taggers, web search engines... 

1. **Machine translation** from Standard German to a given Swiss German dialect
   - User provides a Standard German text and clicks on a Swiss map to select one target dialect, obtains a translation of the text.

2. **Identification** of Swiss German dialects
   - User provides a snippet of dialect text, obtains a shaded map.

3. **Parsing** of Swiss German dialect text
   - User provides a dialect text, obtains a structural analysis and a shaded map indicating where this analysis applies.

Written representations of dialect, as used in electronic media.
Existing resources for Standard German:
- Lemmatizers
- Morphological annotation
- Part-of-speech taggers
- Parsers
- Word-form lexicons

Rules to transform Standard German into Swiss German:
- Phonetic rules that adapt Standard German words phonetically.
- Lexicon correspondences that translate Standard German words which cannot be generated with phonetic rules.
- Syntactic rules that modify the sentence structure.

Most of these rules yield different results in different Swiss German dialects. Hence, the rules are associated with probability maps.
Target dialect coordinates

Standard German Text

Standard German lemmatization/tagging/parsing

Analyzed Standard German Text

Phonetic/lexical/syntactic transfer rules

Swiss German Text
Dialect identification

Morphologically annotated Standard German Lexicon

Phonetic/lexical transfer rules

Morphologically annotated Swiss German Lexicon

Each entry is associated to a map.

Swiss German Text

Lexicon lookup and map combination

Probability map
Dialect parsing

Morphologically annotated Standard German Lexicon

Phonetic/lexical transfer rules

Morphologically annotated Swiss German Lexicon

Standard German parsing model

Syntactic transfer rules

Swiss German parsing model

Swiss German Text

Parsing

Parse tree and probability map
Georeferenced rules

Example of a phonetic rule:

Variable  Standard German -nd in word-final position
e.g. Hund ‘dog’, Kind ‘child’

Variants  -nn [n], -ng [ŋ], -nt [nt], -nd [nd]

\[
\begin{align*}
\text{nd} & \rightarrow \text{nn} \\
\text{nd} & \rightarrow \text{ng} \\
\text{nd} & \rightarrow \text{nt} \\
\text{keep nd}
\end{align*}
\]

Black: \( p = 1 \)  White: \( p = 0 \)
Georeferenced rules

Example of a lexical rule:

Variable: Standard German *immer* ‘always’

Variants:
- *geng*
- *immer*
- *all*

Example of a syntactic rule:

Variable: Word order of auxiliary verb and past participle in subordinate clauses

Variants:
- **Aux + PP**
- **PP + Aux (like Standard German)**
Map sources and processing

- **SDS (Sprachatlas der deutschen Schweiz):**
  - Data collection 1939-1958, publication 1962-1997
  - 8 volumes, > 1000 hand-drawn maps
  - Need to scan and digitize maps
  - Phonetics, Morphology, Lexicon

- **SADS (Syntaktischer Atlas der deutschen Schweiz):**
  - Data collection recently completed, publication scheduled 2011
  - 118 items on questionnaire, syntax and morphosyntax only
  - Access to numerical raw data

- Some rules apply uniformly to all dialects and do not require maps.

- **Interpolation:**
  - Atlas maps only show values at the inquiry points.
  - To obtain values at the points in between, use interpolation.
Set of interpolated surface maps

\[ nd \rightarrow nn \]

\[ nd \rightarrow ng \]

\[ nd \rightarrow nt \]

\[ \text{keep } nd \]

Black: \( p = 1 \)  \quad \text{White: } p = 0
Map interpolation – Open questions

- The current interpolation algorithm merely looks at distance between inquiry points.
  - Influence of topographical, political, denominational borders?
  - Importance and population size of inquiry points?

- Continuous probability maps suggest continuous dialect change.
  - Rumpf et al. (2010) use polygons (Voronoi mosaics), i.e. clear-cut borders.
  - 19th century debate in dialectology.

- Atlas data is from the 1950s, but dialects have changed over time (Christen 1998):
  - Small-scale variants tend to disappear.
  - City dialects and suburb dialects influence each other.
  - Convergence in lexicon, divergence in grammar.

Different interpolation settings yield different results on the three tasks.
Models

Translation and dialect identification models working.
Parsing model to be built, depends on results on dialect identification.

Rule base
108 phonetic rules (fairly complete)
28 morphology rules (50% of target)
43 lexicon rules
Some syntactic rules for testing

Development and evaluation corpus
Text data from 5 dialects extracted from Swiss German Wikipedia.
Text data for evaluation

Need:
- Written dialect data
- Annotated with the author’s dialect
- For parsing: annotated with syntactic structure

Have:
- Blogs, Chat
  - No dialect annotation, huge differences in spelling
- Dialect literature
  - Only a few regions have literary traditions
  - Copyright restrictions
- Swiss German Wikipedia
  - Annotated with dialect (~6 categories with sufficient material)
  - Different spelling conventions in use, but fairly consistent
- Scientific data
  - Mostly non-digitized phonetic transcriptions
  - Dialect corpus under construction (University of Zurich)
Main characteristics
- Systems take advantage of existing Standard German data.
- Different applications use the same maps and rules.
- Knowledge-based methods are more adapted to the data situation.

Contributions to dialectology and computational linguistics
- New insights into dialectological questions by using methods from computational linguistics.
- Modelling of language variation in computational linguistics.

Perspectives
- Test different interpolation methods to model dialect change.
- Dialect speech synthesis to respect modality-related diglossia.
- Dialect corpora open new perspectives for model tuning and evaluation.
Danke!

Dank schöön!

Merci!

Vergelt’s Gott!


