Universität Zürich[UZH]

Institute of Computational Linguistics

Compilation of a Swiss German Dialect Corpus and its Application to Part-of-Speech Tagging

Nora Hollenstein, Noëmi Aepli
Institute of Computational Linguistics, University of Zurich

Introduction

- Compilation of *NOAH's Corpus of Swiss German Dialects* consisting of various text genres, manually annotated with Part-of-Speech tags
- Training and evaluation of a statistical Part-of-Speech tagger, achieving an accuracy of 90.62%

Swiss German

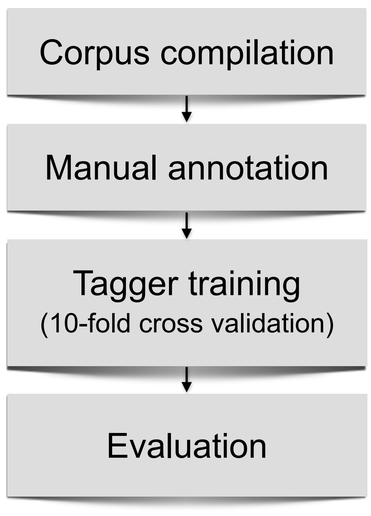
- Swiss German is a low-resourced language and belongs to the Alemannic group of dialects.
- Swiss German is a dialect continuum whose dialects are very different from Standard German.
- It is used in spoken language and in informal written texts (emails, blogs, text messages, etc.).

Differences to Standard German

- Vocabulary: different genus for the same word
  - Standard German: *das Radio*
  - Swiss German: *der Radio*
- Verb tenses: no preterite form in Swiss German
  - Standard German: *Ich las ein Buch.*
  - Swiss German: *Ich ha es buech gläse.*
- Use of auxiliary verbs:
  - Standard German: *Mir ist kalt.*
  - Swiss German: *Ich ha chalt.*
- Verb order is more flexible in Swiss German
  - Standard German: *Sie lies inn gehen.*
  - Swiss German: *Sie hät ihn gah lah.*
- Merged words in Swiss German
  - Standard German: *gehen wir*
  - Swiss German: *gömmer*

Method

The procedure of our work can be summarised as shown in the graphic below:



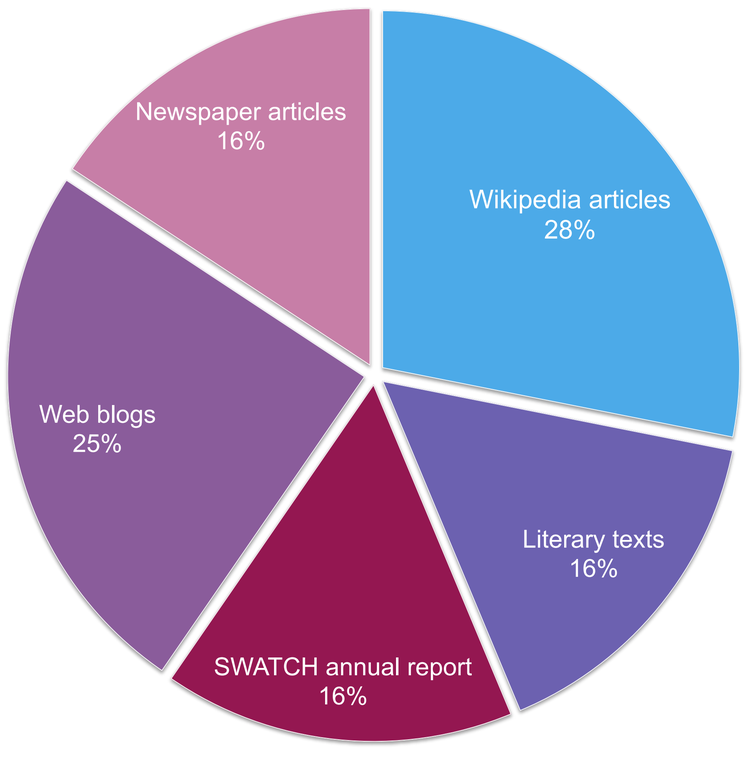
Tagset

- **Basic tagset**
  Stuttgart-Tübingen-Tagset (STTS), the standard for German.
- **Additional attributes**
  - Introduction of the tag *PTKINF* for infinitive particles:
    Standard German: *Ich gehe einkaufen.*
    Swiss German: *Ich go go (PTKINF) poschte.*
  - Adding of a "+"-sign to any PoS tag of a merged word:

PoS tag	Swiss German	Standard German	English
VAFIN+	*isches*	ist es	is it
KOUS+	*dasme*	dass man	that one
VMFIN+	*chame*	kann man	can one
PTKZU+	*zflügä*	zu fliegen	to fly
ADV+	*deetobe*	dort oben	up there

NOAH's Corpus of Swiss German Dialects

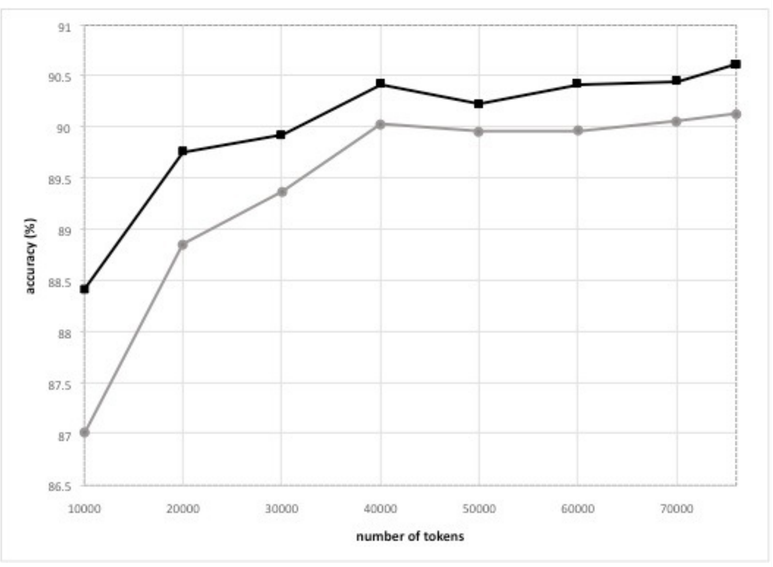
- **Download**: http://www.cl.uzh.ch/research/downloads.html
- Compilation of a corpus consisting of various text genres.
- Including dialects of most German-speaking regions of Switzerland.
- Manually annotated with Part-of-Speech tags.
- Training and evaluation of a statistical Part-of-Speech tagger, achieving an accuracy of 90.62%.



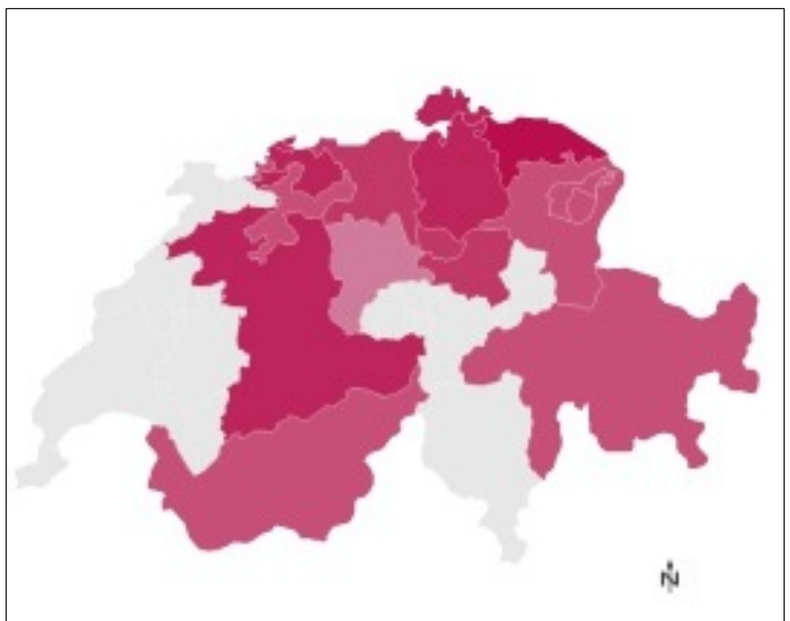
Corpus composition	Number of tokens	Tagging accuracy
Wikipedia articles	*20135*	*90.92%*
Literary texts (novels)	*11165*	*89.37%*
SWATCH annual report	*11386*	*88.82%*
Web blogs	*17671*	*88.10%*
Newspaper articles	*11259*	*87.17%*
Total	***73616***	***90.62%***

Part-of-Speech Tagging

- Training of 6 statistical PoS-Taggers
- Best results achieved with *BTagger*
- *BTagger* makes use of context information and emphasises the transition probability by learning sequences of tags.
- 10-fold cross validation over the complete corpus
- Most frequent errors:
  - Confusion of nouns (NN) and proper names (NE)
  - Confusion of articles (ART) and personal pronouns (PPER)
- Accuracy: 90.62%



Relation between PoS tagging accuracy and corpus size for the *TnT* tagger (grey line) and the slightly better results from the *BTagger* (black line).



The map shows in red the different dialect regions represented in the corpus. It covers almost all German-speaking areas of Switzerland.

Conclusion

- There is a need for more language processing tools for Swiss German.
- *NOAH's Corpus* is a basis for continuative research in Swiss German language processing.