Universität Zürich[UZH]

Institute of Computational Linguistics

NOAH 3.0: Recent Improvements in a Part-of-Speech Tagged Corpus of Swiss German Dialects

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

Introduction

- NOAH is a part-of-speech tagged (POS) text corpus of different Swiss German dialects
- for **version 3.0** we applied machine learning to spot annotation inconsistencies

Swiss German

- a **low-resourced** language belonging to the Alemannic group of dialects
- a dialect continuum where differences between dialects can be found in every aspect and whose dialects are very different from Standard German
- **differences** in all linguistic aspects; phonetics, lexicon, morphology, syntax
- used in **spoken language** & **informal written** texts (emails, blogs, text messages, etc.)

Examples of differences to Standard German

- **vocabulary**: different **gender** 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 liess ihn gehen.*
  - Swiss German: *sii hät ihn **ga la** / **la ga**.*
- unused **cases** in Swiss German
  - Standard German: Die Augen **des** Froschs.
  - Swiss German: **am** frosch **sini** auge
- **merged words** in Swiss German
  - Standard German: *gehen wir*
  - Swiss German: ***gömmer***
- …

Tagset

- **basic tagset:** *Stuttgart-Tübingen-TagSet* (STTS), the fine grained standard tag set for German (54 tags)
- **two tagset extensions**

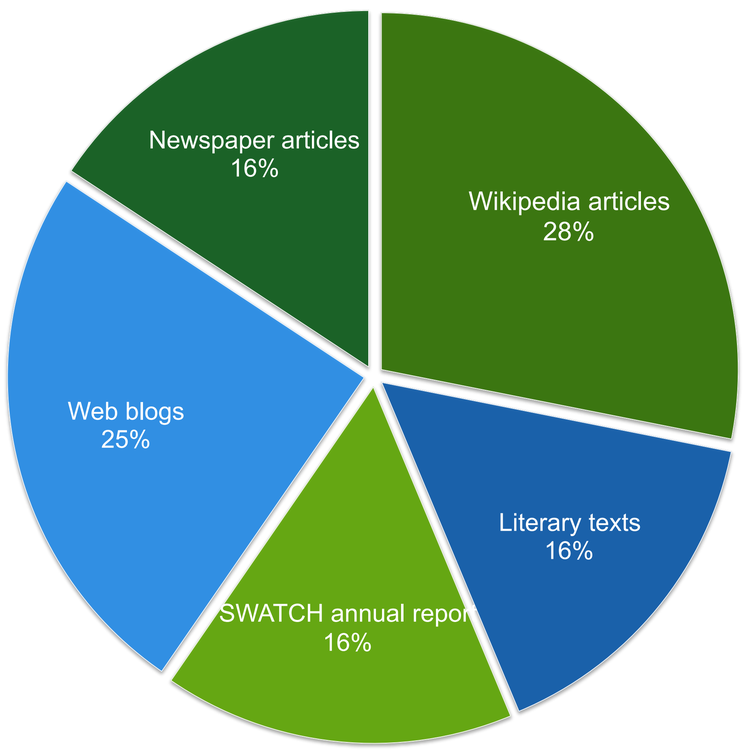
  (1) PTKINF for infinitive particles
  - Standard German: *Ich gehe einkaufen.*
  - Swiss German: *Ich go **go**/PTKINF poschte.*

  (2) "**+**"-marker for indicating merged words in total: 1,767 POS tags with "**+**", i.e. 1.6% some examples:

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

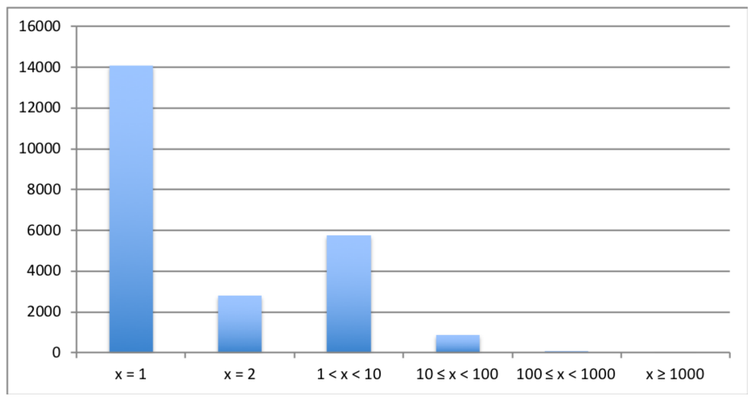
NOAH's Corpus of Swiss German Dialects - V3.0

- various text **genres** (oral and written style)
- **dialects** of many German-speaking regions of Switzerland
- 113,565 tokens
- **manually annotated with** part-of-speech tags
- http://pub.cl.uzh.ch/purl/NOAH



tagger	accuracy
TNT Tagger	92.4%
Wapiti (CRF)	92.4%
BTagger	93.5%

Swiss German Variability



frequencies of type frequencies (x) in a Swiss German text

- 6,155 sentences | 105,692 tokens | 20,882 types (NOAH + parts of 2 additional novels)
- 14,099 hapax legomena | 2,804 hapax dislegomena
- 19,874 < 10 times | 29,767 < 100 times

→ **high degree of data sparseness!**

Spotting Inconsistencies

- we applied **machine learning** to efficiently spot annotation inconsistencies
- **manual verification** and correction of the differences between a PoS tagger's output and the hitherto gold standard
- **corrections**/modifications of 2,205 **POS tags**, i.e. 1.94% of total 113,565 tags of V3.0
- **corrections** of 81 **segmentation** errors, i.e. 1.1% of total 7,303 sentences/segments of V3.0
- **guideline changes** for problematic cases
- anticipating future dependency annotations, we systematically chose the **syntactically more consistent option**, e.g.

Modal verb's past participles

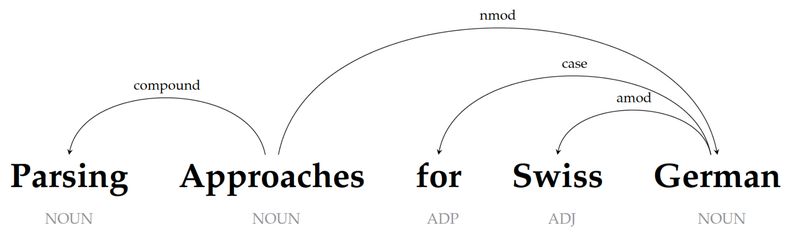
We deviate from the TIGER scheme and categorise them as such (**VMPP**) even though they morphologically look like infinitives.

ich/PPER
ha/VAFIN
das/PDS
*müesse/**VMPP***
läse/VVINF

Standard German: *ich hab das lesen müssen*

Parsing Approaches

- next step – from POS tagging to syntactical parsing: **universal dependency parsing** for Swiss German
- application of **different cross-lingual parsing strategies** exploiting Standard German resources
- **3 approaches**
  - lexicalised **annotation projection**
  - delexicalised **model transfer**
  - direct cross-lingual transfer
- results show **~60%** Labelled Attachment Score (LAS) for all approaches
- provide a first step towards Swiss German dependency parsing



- https://github.com/noe-eva/SwissGermanUD

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