A Resource for Natural Language Processing of Swiss German Dialects

Nora Hollenstein
Institute of Computational Linguistics
University of Zurich
hollenstein@ifi.uzh.ch

Noëmi Aepli
Institute of Computational Linguistics
University of Zurich
naepli@ifi.uzh.ch

Abstract

Since there are only a few resources for Swiss German dialects, we compiled a corpus of 115,000 tokens, manually annotated with PoS-tags. The goal is to provide a basic data set for developing NLP applications for Swiss German. We extended the original corpus and improved its annotation consistency. Furthermore, we trained dialect-specific PoS-tagging models and implemented a baseline system for dialect identification.

1 Introduction

Swiss German is a dialect continuum which includes dialects derived from Standard German. However, NLP tools for Standard German do not perform well on Swiss German as it differs from Standard German in terms of phonetics, vocabulary, morphology and syntax. Moreover, there is no official orthography standard. Nevertheless, Swiss German has been used more frequently in recent years not only in informal contexts but also in the media and by literary authors. Therefore, the need for dialect-specific resources as a foundation for Swiss German NLP becomes evident. In previous work, Scherrer and Rambow (2010a) reviewed the existing resources and applications that have been developed for Swiss German language processing.

The present work builds on an existing corpus for Swiss German dialects (Hollenstein and Aepli, 2014). *NOAH's Corpus* includes written texts from different genres: Wikipedia articles, news articles, the SWATCH annual report (2012), chapters from novels and web blogs. We extended the corpus through further manual annotation in order to provide a larger lexical resource, which is freely available for research purposes[1]. Moreover, we explored two applications of this corpus: dialect-specific PoS-tagging and dialect identification.

2 Annotation

We extended the corpus by adding additional texts of the genres mentioned above to reach 115,000 tokens. The manual annotation of PoS-tags was conducted by the authors (Swiss German native speakers) and followed the same guidelines from Hollenstein and Aepli (2014). The STTS (Schiller et al., 1999) was chosen as the basic tag set with the addition of a few new tags to cover phenomena not present in Standard German. Swiss German requires a tag *PTKINF* (infinitive particle) for sentences such as *"Am erschtä Tag simmer go (PTKINF) poschtä."*[2], which has no direct translation to Standard German. Furthermore, a "+"-sign is added to the tag of merged words, which appear due to the lack of an orthography standard.

2.1 Annotation Consistency

In order to provide a reliable resource for language technology, we place great importance on the manual annotation process. To ensure that the annotation is consistent throughout the whole corpus and to further specify the annotation guidelines, we applied a simplified version of the *variation n-gram method* (Dickinson and Meurers, 2003). This technique detects sequences of *n* tokens which occur multiple times in the corpus with varying annotation. The detected n-gram sequences were manually analyzed.

3 Dialect-specific PoS-Tagging

We trained the BTagger (Gesmundo and Samardžić, 2012) on the annotated data. Over the whole corpus a tagging accuracy of over 90% was reached (Hollenstein and Aepli, 2014). This on-going work places emphasis on the variety of dialects present in Swiss German. Taking advantage of the fact that dialect information is

[1] Download link: http://kitt.cl.uzh.ch/kitt/noah/corpus

[2] Translation: *"The first day we went shopping."*

Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology,
pages 108–109, University of Duisburg-Essen, Germany, Sep 30–Oct 2 2015.

available as metadata in our corpus, the PoS-tagger was trained for each dialect separately. We focused on the five dialects for which the largest amount of training data is available and evaluated these through a 10-fold cross-validation. Each model was trained on 4,000 tokens. The results can be observed in Table 1. More training data will be needed to improve the performance of these models.

Dialect	Accuracy
Aarau	85.73%
Basel	85.28%
Bern	87.85%
Ostschweiz	85.77%
Zürich	87.47%

Table 1: PoS-Tagging accuracy for each of the five dialects.

4 Dialect Identification

As a second application to this corpus, we are in the process of building a dialect identification system. Scherrer and Rambow (2010b) showed that dialect identification via a character n-gram approach could indeed perform well even for very similar dialects, given that sufficient training data from different sources is available. With this corpus, which provides a variety of text genres and dialect information as metadata for most of the included articles, we believe that we have a solid data set to develop a dialect identification model.

In order to implement a baseline system for five major Swiss dialects (Aarau, Basel, Bern, Ostschweiz and Zürich) we compiled a development set of 1,470 sentences and a test set of 250 sentences (50 per dialect). The trained dialect ID model uses a character-based trigram approach. We trained a trigram language model for each dialect and scored each test sentence against every model. The predicted dialect was chosen based on the lowest perplexity.

Table 2 shows the results of this baseline system. Overall, this model reached an F-score of 0.66. To improve this model in the future, the limited amount of training data and the similarity between dialects will have to be taken into account.

5 Conclusion and Future Work

We have extended the corpus for Swiss German dialects to 115,000 manually PoS-tagged tokens.

Dialect	P	R	F
Aarau	0.30	0.36	0.33
Basel	0.54	1.0	0.70
Bern	0.52	0.76	0.62
Ostschweiz	0.68	1.0	0.81
Zürich	0.74	1.0	0.85
Average	0.56	0.82	**0.66**

Table 2: Performances of the trigram model on the test sentences. The columns refer to precision, recall and F-score respectively.

Furthermore, we improved the quality of the data by conducting consistency tests. Employing this improved corpus, we experimented with two possible NLP applications. First, we trained dialect-specific PoS-tagging models which reached between 85% and 88% accuracy. Second, we implemented a baseline system for dialect identification for future research. This system based on character-based language models achieved an overall F-score of 0.66. The dialect ID model for the Swiss German dialect continuum is subject to future work.

References

M. Dickinson and D. Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 107–114. Association for Computational Linguistics.

A. Gesmundo and T. Samardžić. 2012. Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 368–372. ACL.

N. Hollenstein and N. Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to pos tagging. *COLING 2014*, page 85.

Y. Scherrer and O. Rambow. 2010a. Natural language processing for the Swiss German dialect area. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 93–102, Saarbrücken, Germany.

Y. Scherrer and O. Rambow. 2010b. Word-based dialect identification with georeferenced rules. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1161. Association for Computational Linguistics.

A. Schiller, S. Teufel, C. Stöckert, and C. Thielen. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS, August.