11-712: NLP Lab Report

David Klaper

April 25, 2014 **[due date –NAS]**

Abstract

[one paragraph here summarizing what the paper is about –NAS]

[brief introduction –NAS]

1 Basic Information about Swiss German

Swiss German is a group of Germanic dialects spoken in Switzerland. In 2000, about 4.6 million people in Switzerland spoke Swiss German (Lewis et al., 2013). By now, this number probably has increased since the overall population of Switzerland increased from 7.2 to 8 million people since 2000 (Swiss Federal Statistical Office, 2013). In general, Swiss German is quite similar to Standard German there are specific syntactic, lexical and other differences between Standard German and the Swiss German dialects. There are also considerable differences with regard to these features between different dialects. (Scherrer, 2011)

2 Past Work on the Syntax of Swiss German

The name Swiss German already indicates that it is closely related to (Standard) German. Many words and the basic syntactic structure are similar or equal to German. One often-cited characteristic of Swiss German is the existence of context-sensitive structures in some dialects as shown by Shieber (1985).

Due to the similarity to Standard German it makes sense to consider resources for Standard German syntax. There are many resources on Standard German syntax. One of them is the World Atlas of Language Structures (Dryer and Haspelmath, 2013). It summarizes the syntactic and morphologic properties of a language as a list of features. Furthermore, there exist dependency treebanks and dependency parsers for Standard German. As an example Sennrich et al. (2009) present a hybrid dependency parser, which combines hand-written rules with a statistical model. The parser relies on supervised training data, which is not available for Swiss German yet.

We have seen that for Standard German there exist many powerful tools and datasets. Unfortunately, there are still considerable differences between the two. First, Swiss German is mainly a spoken language and no unified writing system exists. Furthermore, even within Switzerland the dialects vary considerably regarding pronunciation and in consequence spelling.

Scherrer (2007) attempted to normalize Swiss German words to their Standard German counterparts, which would allow using the Standard German resources. Although he created a working system, the results are below 50% for both precision and recall. Scherrer states also that "for many dialect words, it yields no result at all" (Scherrer, 2007, p. 60). Scherrer and Rambow (2010) worked towards machine translation from Standard German to specific Swiss German dialects and outlined how to use Standard German resources for creating a Swiss German constituent parser. This work is

very interesting but it requires dialect identification and the performance is not good enough. Specifically, introducing a high error rate in a preprocessing step will make building a robust dependency parser even harder.

As stated before there are differences between Standard German and Swiss German as well as between the different Swiss German dialects. Scherrer (2011) proposes a system to normalize Swiss German dialects to Standard German syntax and explains some of the syntactic differences. Bucheli and Glaser (2002) started a project for mapping the differences between the Swiss German dialects in an atlas. However, as of the beginning of 2014 the atlas has not been published. As part of this project Glaser and Frey (2006) investigated reduplication phenomena in Swiss German dialects.

Beyond the resources describing Swiss German syntax and some tools to standardize Swiss German there are very few tools available. Also, there are no freely available corpora published yet, but Aepli and Hollenstein (2014) at the University of Zurich are working on a corpus of 50'000 tokens annotated with *Part-of-Speech* (PoS) tags. They are also building a PoS-Tagger for Swiss German. They trained a model which got them approximately 85% accuracy on Alemannic Wikipedia texts with 20'000 tokens training. (Aepli and Hollenstein, 2012). It does not discriminate between different dialects, which eliminates dialect identification as a source of error. Dialect identification is especially difficult because the borders are continuous and many people have lived in multiple dialect areas and have a mix of dialects (e.g. the writer of this document). This tagger could be very useful for the parser.

In conclusion, while Swiss German has many similarities with Standard German, for which many resources exist. The attempts to normalize Swiss German dialects to Standard German have yielded mediocre performance. There are some dialectometric studies detailing the differences between the distinct dialects but very few computational systems and no corpora. Aepli and Hollenstein (2014) are working on a corpus and PoS-Tagger for Swiss German. This tagger will not discriminate between different dialects which eliminates the dialect identification.

3 Available Resources

Since Swiss German is a mainly spoken language, written resources are sparse. In particular, organized corpora are basically non-existant or transcribed resources from interviews or otherwise artificial written text. The Zürcher Kompetenzzentrum für Linguistik (Zurich compentency center for linguistics) compiled a ist of corpora[1]. The problem with the list is that there are non-working links and most resources are only available on request and may not be shared. Finally, as mentioned above most of these resources are geared towards linguistic analysis use and are just transcriptions of spoken language.

For building their PoS tagger, Aepli and Hollenstein (2012) compiled data from the Alemannic Wikipedia[2] and annotated 20'000 tokens with PoS tags. Since the seminar thesis they have significantly increased the scope and size of their corpus. By now they have annotated over 50'000 tokens. First, there is a Swiss German edition of the 2012 corporate report of the Swiss firm Swatch[3]. The report contains more than 70'000 tokens of which 13'000 were annotated with PoS tags. Another source they used is the anniversary edition of a daily free newspaper in Switzerland [4], which was a special edition written in Swiss German. They annotated about 11'000 tokens. They also annotated

[1] `http://www.linguistik.uzh.ch/resources/korpora.html` last checked January 31^{st}, 2014.

[2] `http://als.wikipedia.org` last checked last checked January 31^{st}, 2014.

[3] Report available at `http://www.swatchgroup.com/de/investor_relations/jahres_und_halbjahresberichte` last checked January 31^{st}, 2014.

[4] Edition May 28^th, 2013 in the archive at `http://www.blick.ch/blickamabend/epaper/` last checked January 31^{st}, 2014.

Source	Tokens with PoS	Tokens overall
Wikipedia	20'000	60'000
Newspaper	11'000	17'000
Swatch Report	13'000	70'000
Blog Text	0	277'000
Total	**44'000**	**424'000**

Table 1: Rough estimated token counts

extracts from murder stories by Viktor Schobinger[5] but I decided not to use this part, because the domain is quite different and probably even more difficult than the rest. (Aepli and Hollenstein, 2014)

In order to complement these resources, I decided to compile a new collection of Swiss German texts from the web, as an additional ressource. I was considering Twitter text but this will probably be too noisy. Also, it might be hard to detect whether a short tweet is Swiss German or not. Therefore, I decided to search for Swiss German blogs. I restricted myself to blogspot blogs, in order to facilitate subsequent extraction. Using google queries involving typical Swiss German words and 'blog' I searched for mainly Swiss German blogs. Using this technique, I selected 17 blogs from people posting entries written in Swiss German. Interestingly, the queries I used returned almost exclusively returned blogs from Swiss people doing an exchange program or living abroad. I extracted 885 blog entries (but some contain only pictures no text and a few are in languages other than Swiss German).

I then proceeded to filter out useless posts. I extracted the xml, removed html tags and wrote each blog entry into its own file with the pattern 'blog_*blognumber*_*postnumber*.txt'. The blog number relates to one of the original 17 blogs. The postnumber indicates how old an entry is. Hence, the postnumber is in inverse chronological order. After extracting the text from the xml, I removed all files that were empty or less than 100 Bytes, since many posts only included pictures. The, I manually went through all blog posts, deleting the entries with more than half written in another language. In a few special cases, e.g. song lyrics, I just removed these and kept the blog post if they were not vital to the content. At the end 615 blog posts remained.

After cleaning the data I wrote a script to count the tokens. As a token I defined a sequence of alphabetical or numerical characters of length 2 or more. Note that this definition ignores punctuation. I will later use a tokenization script provided by Aepli and Hollenstein (2014), where the count will be higher, but for now this is precise enough. The token count was 277'000 with 32'402 types. This is a type-to-token ratio of 11.7%, which is fairly high. Moreover, over half of the types (19'397) occur only once, i.e., they are hapax legonoma.

Also, I will forward the data to Aepli and Hollenstein (2014) and they will annotate part of it to make it part of their tagger training material. A side resource I might use (depending on the parsing framework I use) is a list of Swiss German words without German roots[6]. If this word list helps, I might also consider German word lists.

For the reference corpora A and B, I plan to select data from each data set. In particular, corpus A will consist of a text-section from the Swatch report, Wikipedia, or the newspaper, which is at least 1000 tokens in one piece. Corpus B will consist of one or multiple blog posts from the same author, since the posts are usually reasonably contiguous. In short, I plan to use over 150'000 tokens Swiss German text from Aepli and Hollenstein (2014) of which about 50'000 tokens are annotated

[5] `http://www.zuerituetsch.ch/zueri-krimi.html` last checked January 31^{st}, 2014.

[6] `http://www.dialektwoerter.ch` last checked January 31^{st}, 2014.

with PoS-texts. In addition, I will use Swiss German blog text from mainly exchange students. This collection will have over 250'000 tokens Swiss German text, of which I can use most for training. This is most likely the largest collection of Swiss German written text in existence with over 400'000 tokens. I hope that such data will reduce the effect of the many different ways to write the same word in Swiss German.

4 Survey of Phenomena in [Your Language/Genre –NAS]

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

11 Acknowledgements

I'd like to thank Noëmi Aepli and Nora Hollenstein for supporting me with information about Research in Swiss German and especiallz for giving me early access to the Swiss German resources that they are developing.

References

Noëmi Aepli and Nora Hollenstein. Part-of-Speech tagging für Schweizerdeutsch. Seminar Thesis, Institute of Computational Linguistics, University of Zurich, 2012.

Noëmi Aepli and Nora Hollenstein. Personal Communication (January), 2014.

Claudia Bucheli and Elvira Glaser. The syntactic atlas of Swiss German dialects: Empirical and methodological problems. In *Syntactic Microvariations*. Meertens Institute, 2002.

Matthew S. Dryer and Martin Haspelmath, editors. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, 2013. URL `http://wals.info/languoid/lect/wals_code_ger`.

Elvira Glaser and Natascha Frey. Doubling phenomena in Swiss German dialects. In *Workshop on Syntactic Doubling in European Dialects*, 2006.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world. 17th Edition. SIL International. `http://www.ethnologue.com/language/GSW`, 2013.

Yves Scherrer. Adaptive string distance measures for bilingual dialect lexicon induction. In *Proceedings ACL*, pages 55–60, 2007.

Yves Scherrer. Syntactic transformations for Swiss German dialects. In *Proceedings of EMNLP*, 2011.

Yves Scherrer and Owen Rambow. Natural language processing for the Swiss German dialect area. In *Proceedings KONVENS*, 2010.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. A new hybrid dependency parser for German. In *Proceedings GSCL Conference*, 2009.

Stuart M. Shieber. Evidence against the context-freeness of natural languages. *Linguistics and Philosophy*, 8:333–343, 1985.

Swiss Federal Statistical Office. Ständige wohnbevölkerung nach geschlecht und staatsangehörigkeit-skategorie, am ende des jahres. `http://www.bfs.admin.ch/bfs/portal/de/index/news/04/01.Document.141977.xls`, 2013.