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 future 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

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

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

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 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 Swiss German

The basic word order is a bit tricky in German since the main clause has a subject-verb-object (SVO) order but the subordinate clauses are verb-final, i.e., subject-object-verb (SOV). German has a richer morphology than English but in Swiss German some of the distinct cases collapse to the same wordform. Furthermore, Swiss German lacks a preterite tense, all sentences in the past are expressed using perfect. Finally, Swiss German does not have the genitive case. It is replaced by dative constructions. (Scherrer, 2011)

These are some general observations but the focus here is on syntactic phenomena that influence the dependency structure in crucial ways. Scherrer (2011) lists more interesting differences. One point are the non-projective dependencies in Swiss German. One reason for them is the verb raising with the auiliary verb "lo" (*let*), where the dependency between the auxiliary and its argument and the main verb and its object will cross. There exist other forms of verb raising, where verb order differs from the Standard German word order. Also, this may depend on the dialect. In short, the word order of Swiss German is more free than Standard German.

Some more phenomena that illustrate differences between the well-studied Standard German and Swiss German, are a preposition before a dative in the case it is not already part of a prepositional phrase. This would have more impact on constituent parsing though. Scherrer (2011) lists more phenomena but some of them seem rather uncommon in central Switzerland. In particular, the very southern dialects have even more deviations, but most of the population lives in central and northern Switzerland, therefore these effects might not occur very much in the corpora. A last part that might be particularly tricking for dependency annotation are doubling phenomena in Swiss German as explained by Glaser and Frey (2006). For many of these phenomena it is not clear, which word is the head and which the dependent even if they are clearly related.

By inspection of the data, I looked for other interesting phenomena. One phenomena is the melting of verbs and personal pronouns (and similar combinations) that is also reflected in the tagset by Aepli and Hollenstein (2012). This is a major challenge for the annotation of dependencies, because the two words have independent syntactic roles but both its dependents will be added to this single token. There are many spelling variations, even within a document a term may not be written the same way or spaced differently in some cases. Hence, for data-oriented methods this may pose a problem.

For the blog corpus, there are some specific observations. First, often vowels are duplicated multiple times to place emphasis on a word, such as "gaaanz elei" (*totally alone*), which increases the number of types for the same lemma. Also, because the blogs are by Swiss people abroad there are many insertions of foreign language material, e.g. "sWetter isch totally crazy gsi" (*the weather was totally crazy*).

In conclusion, Swiss German is generally more free regarding word order than Standard German, especially with regard to auxiliary and main verb. This may yield non-projective dependency relations. Furthermore, verb and personal pronouns can sometimes be contracted to one word, as other combinations can. The spelling is not uniform and duplication of vowels is common. For

dependency annotation especially the frequent use of auxiliaries and the duplication of words can be difficult, since the head-dependent relation is not obvious in these cases.

5 Initial Design

This section describes the annotation design and briefly outlines the next steps, i.e., building the actual parser. As stated before my two test corpora have a different focus. Corpus A consists of a small part of the data by Aepli and Hollenstein (2012). Corpus B consists entirely of blog posts. The data was annotated manually. To ensure annotation consistency the following documents the annotation decisions made.

Table 2 gives an overview of the annotation decisions. Before going into the detailed decisions an overall observation. The sentence splitting of my tokenizer is far from perfect, since capitalization can not be used as a feature of sentence splitting. First, because some people write all lower case and second, because in German common nouns are capitalized. Therefore, it is possible to have multiple words attached to the root, which mean that the clauses are independent. Most of the decisions are semantically oriented. Also, I sometimes give English-only examples, when the matter is the same in both languages.

Since Swiss German does not use a past form, many sentences have auxiliaries. **Decisions 1 and 2** say that in sentences with multipe verbal parts, the inflected auxiliary is always the head. This is illustrated in the example below. The subscript numbers indicate the position of the parent in the senence. Zero means root. The inflected verb *het* dominates *welle*, which dominates the main verb *lösche*. **Decisions 3 and 4** concern the attachment of the arguments. The subject always attaches to the inflected verb as a child. Thus, *hans* is a child of *het*, all other arguments and adjuncts attach to the main verb, e.g. *lösche*. There are certain special cases. Especially, modal verbs, such as *cha* ('can') can be modified by adverbial constructs. These attach to the verb they modify. Sometimes this is hard to decide, then they are attached to the main verb. Note that the verb *sött* ('should') can be used as a main verb. The reason for this is that the importance of the first auxiliary is reflected and the connection between subject and form of the inflected verb is explicit. Furthermore, the main verb defines what kind of objects are allowed, therefore it dominates the arguments.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
De	hans_3	het_0	welle_3	d'	date_{10}	mit	em	magnet_{10}	lösche_4
The	Hans	has	want	the	data	with	the	magnet	delete

'Hans wanted to delete the data with the magnet'

Decision 5 states that subordinate clauses are headed by the subjunction that introduces them. Hence, the head verb in the embedded clause is dependent on the relative pronoun (or other subjunctive word). The reasoning is that the subjunction defines how the clause relates to the main sentence and hence, acts as the root of the embedded clause. **Decision 6** states that in a coordination the conjunction, typically *ond* or *und* ('and'), is the head of the conjuncts. Furthermore, should there be multiple parts to a conjunction of which all but the last are connected by commas, all parts of the conjunction will be direct dependents of and. (The comma is not reflected, since it is attached to the root, see below). If there are multiple 'and' (that are not semantically in a hierarchy), the 'and' dominate from left to right, i.e., the first and will be the highest. **Decision 7** determines that in numerical expressions like *öppe 3 stund* ('about 3 hours'), the unit is the head of the number and the number is the head of the modifier. The reason for this is that the number quantifies the unit and themodifier modifies the quantity. **Decision 8** is related to the fact that certain verbs are built by prepending a particle in front of the verb. In infinitive constructions this particle can be split from the verb. It is clear that this particle is a direct dependent of the verb it belongs to.

The important **Decision 9** says that prepositional phrases are headed by the noun. The reason is that there are large variations in the use of prepositions among the dialects, while the noun is always there. Furthermore, some dialects use sequences of prepositions and articles, while others contract both to one word. **Decision10** states that discourse markers, such as interjections and smileys are attached to the corresponding head of the phrase they belong to. Note that smileys might be attached to the root if they are outside of a sentence or clause context. Also, Swiss people use 'and' and other conjunctions sometimes as discourse fillers or simple linking elements, in that case this rule applies and not the coordination rule.

A special phenomena is the contraction of different words in Swiss German. For instance, one can contract a relative pronoun *wo* ('where, when') and the following personal pronoun *ich* ('I') to *wonich*. This was also observed by Aepli and Hollenstein (2012). Therefore, they introduced multiple PoS tags to deal with such phenomena. The problem is where in the dependency parse these constructions with multiple roles should end up. **Decision11** is the word should always adopt its strongest role, i.e., the word part that is higher in the dependency parse will determine the position of the contracted word. **Decision 12** concerns comparison phrases of the form below.

1	*2*	*3*	*4*	*5*	*6*
So	schön$_0$	wie$_2$	de	himmel$_6$	esch$_3$
So	beautiful	how	the	sky	is

'As beautiful as the sky (is)'

The point is that the main content of the phrase is the beauty. The verbal phrase is what is being compared to. It is a semantic construction to introduce a contrast. It cannot be inverted, thus, I decided that the verb is not the head in this case. **Decision 13** is similar to number 8. In German, 'no more' can be split and the object that is not available anymore will be between the two parts. Here, the more is clearly a dependent of no. **Decision 14** concerns sequences of multiple nouns. While general compounds are written together, some compounds can be split. There the head is the main content. However there is another special case. It is 'the word -bus-'. Here word will be the head and bus will be dependent. This is also a semantic decision, because the phrase is about a word not a bus. **Decision 15** is also such a 'meta'-concern. In the phrase "for example [X]", X will be the head and example will be dependent, because the semantic content is X.

Decision 16 states that in a sentence that has no verb, the next most important word becomes the head. There are even sentences where a noun is the head and then a finite relative clause is embedded. This inheritance principle makes sense, although in the case of ellipsis it might be hard to determine what exactly should be the head. The decision should also be semantically motivated. **Decision 17** is simply all punctuation is strictly attached to the root. This is because punctuation is very unreliable in Swiss German, because the use by different speakers is inconsistent.

Finally. a couple of decisons not listed in the table, because they are minor. Foreign material, e.g. English, is annotated as much as possible according to English standards, where they exist or otherwise according to this manual. A sensitivity regarding passive formulations. In the passive voice, the main verb is a participle that may be hard to distinguish from an adjective, but it is as treated as a verb as long as it is clearly related to a verb. This leads to a distinction between 'has been possible', where 'been' will be the main verb that gets the arguments as its dependents and 'has been registered', where 'registered' is the main verb. There are subjunctions, such as *aber* ('but'), which can also be used instead to just modify a verb phrase, then they are dependent of the verb. Last, in 'dangerous to deadly', 'dangerous' is the head of 'to', which in turn is head of 'deadly'.

While in general the annotation worked quite well, it turns out that using and as the head of its conjuncts leads to unintuitive phenomena. In particular, if there is a coordinated verb phrase that applies to the same subject, then 'and' dominates both the conjuncts and the subject, which clearly

ID	Pattern	Head
1	Mainverb Aux	Aux
2	Aux [...] Mainverb	Aux
3	Subj	Inflected Verb
4	Object, other Args, Adjuncts	Mainverb
5	RelPron Verb	RelPron
6	[XP] *and* [XP]	*and*
7	Mod Num Unit	Unit
8	Mainverb VerbPart	Mainverb
9	Prep N	N
10	DiscourseMarker	Corresponding Clause Head
11	Contractions	Placed At Highest
12	Adj Comparison MainVerb Aux	Comp
13	2 Part Negation (*no more*)	*no*
14	N_descriptor N_term	N_descriptor
15	Example	Main content
16	No Verb Sentence	Next Best, e.g. Noun
17	Punctuation	ROOT

Table 2: Annotation Decisions

do not have the same function. However, I decided to stick to the annotation schema because I did want to avoid preferring one of the two conjunct phrases over the other. Arguably, 'and' is also used as a linking word instead of a coordination, but this is a different case and should in my opinion not be mixed with the real coordination of phrases.

Finally, a couple of comments about the data. First, the data does not use Gold standard PoS but the ones that are assigned by the tagger model trained on 57'000 tokens Swiss German text. Second, the tokenization and in particular the sentence splitting are far from perfect. Part of this is due to the sloppy use of punctuation and part of it is due to the simplicity of the sentence splitter. Third, Wikipedia in paricular tends to have special formatting characters which impede proper tokenization.

The parser will operate in a semi-supervised fashion, making use of annotated training material as much as possible. It is important to use unsupervised features as well, because a large amount of the available texts will not be annotated. It would be bad not to use this data. The semi-supervised method by Koo et al. (2008) based on Brown Clustering serves as the model for my unsupervised features. I consider using the TurboParser (Martins et al., 2010) as a baseline. Then I would try adding unsupervised features to improve over the supervised baseline. After annotation of the test data, I annotated 10'000 tokens (650 sentences) for training material in order to have the necessary supervised training. For annotation I used the DG Annotator by Giuseppe Attardi (`http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/` last visited February 22^{th}, 2014.)

In conclusion, I tried to adopt a rather semantic notion of head and used this throughout the annotation. I will build a semi-supervised parser.

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 especially 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.

Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In *Proceedings ACL*, 2008.

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.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings EMNLP*, pages 34–44, 2010.

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 staatsangehörigkeitskategorie, am ende des jahres. `http://www.bfs.admin.ch/bfs/portal/de/index/news/04/01.Document.141977.xls`, 2013.