Morphological analysis and lemmatization for Swiss German using weighted transducers

Baumgartner, Reto

Abstract: With written Swiss German becoming more popular in everyday use, it has become a target for text processing. The absence of a standard orthography and the variety of dialects, however, lead to a vast variation in different spellings which makes this task difficult. We built a system based on weighted transducers that recognizes over 90% of the tokens in certain texts. Weights ensure preferring the best analysis for most words while at the same time allowing for very broad range of spelling variations. Our morphological tagset that we defined for this purpose and lemmas in Standard German open the possibility for further processing. Besides our morphological analyzer and lemmatizer, a morphologically annotated corpus offers new resources for Swiss German and helps spreading our tagset.

Originally published at:
Baumgartner, Reto (2016). Morphological analysis and lemmatization for Swiss German using weighted transducers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS) Bochum, Germany September 19–21, 2016, Bochum, 19 September 2016 - 21 September 2016. Universitätsverlag Ruhr-Universität Bochum, 44-49.

Morphological analysis and lemmatization for Swiss German using weighted transducers

Reto Baumgartner
University of Zurich
retoflavio.baumgartner@uzh.ch

Abstract

With written Swiss German becoming more popular in everyday use, it has become a target for text processing. The absence of a standard orthography and the variety of dialects, however, lead to a vast variation in different spellings which makes this task difficult. We built a system based on weighted transducers that recognizes over 90% of the tokens in certain texts. Weights ensure preferring the best analysis for most words while at the same time allowing for very broad range of spelling variations. Our morphological tagset that we defined for this purpose and lemmas in Standard German open the possibility for further processing. Besides our morphological analyzer and lemmatizer, a morphologically annotated corpus offers new resources for Swiss German and helps spreading our tagset.

1 Introduction

With an increased use of written text in Swiss German (SwG), there is a growing interest in tools to process these texts. SwG dialects are spoken by more than 4 million people in Switzerland in everyday life around the centers Zurich, Basel and Bern whose dialects we covered in our system at this stage. In writing usually Standard German (StG) is preferred but for private communication many people use their SwG dialect.

SwG differs from StG in phonology, vocabulary and grammar. Its vowel system still resembles that of Middle High German (MHG) with *Ziit* "time" and *Huus* "house" (MHG *zît* and *hûs*; StG *Zeit* and *Haus*) while the differences in the consonant system and the loss of endings are more modern traits (Christen et al., 2012, p. 27). Over time SwG has lost its genitive case and the past preterite (Siebenhaar and Wyler, 1997, p. 37). Conversely it possesses infinitive particles that are not known to StG.

SwG consists of different local dialects that mainly differ in phonology and to a lesser extent in vocabulary. There is no standard orthography, but there are proposals for sound-character assignment like *Dieth-Schreibung* (Dieth, 1986) or *Bärndütschi Schrybwys* (Marti, 1985a) that are, however, not known to everyone. This results in a high variability of spellings influenced by both dialects and personal writing preferences. As an example for the StG word *Jahr* "year", we found in our corpus *Jahr* and *Jaar*, *Johr* and *Joor*, even *Joh* for different pronunciations and spelling preferences.

The lack of a standard orthography and the vastness of variants motivate the choices that have to be made to process these dialects. For lemmatization we use StG words. The variants can probably best be dealt with using finite-state technology that do not rely on huge corpora but on linguistic engineering. Weighted transducers can be used for a better trade-off between good coverage and overgeneration.

2 Related Work

The increase of SwG in writing led to a number of resources:

Corpora: By now two corpora consisting of everyday written language have been collected. The *Swiss SMS Corpus* (Stark et al., 2009 2015) counts 275 000 tokens in SwG from short messages. The corpus includes manually made glosses in StG. *NOAH's Corpus of Swiss German Dialects* (Hollenstein and Aepli, 2014) counts 115 000 tokens in SwG from different sources like blogs, wikipedia entries, literature, newspapers or a business report. The corpus has manually been annotated with parts of speech. With *Archimob – A corpus of Spoken Swiss German* (Samardžić et al., 2016), there is a corpus of transcribed spoken SwG, opposed to the others whose material was written first.

Taggers: Hollenstein and Aepli (2014) trained a Part-of-Speech tagger model on their collected data that reaches an accuracy of 90.62%.

Morphological generation: A closely related task to ours is morphology generation. An approach from Scherrer (2011) uses replacement rules and information about the dialects' location to build SwG word forms. As this system follows specific spelling guidelines for consistency, it is not suited for analysis where it is important to recognize a broad range of different spellings.

3 Annotation scheme

3.1 Parts of speech

As both the *Swiss SMS Corpus* and *NOAH's Corpus* make use of the *Stuttgart–Tübingen–TagSet* (STTS) (Schiller et al., 1999), we chose the same tagset for our parts of speech. As it was developed for StG we had to make some changes for use with SwG:

Changed use: Some tags had to be opened to different words with the same function. The use of *wo* "where" as relative pronoun (*PRELS*) or as subordinating conjunction (*KOUS*) like StG *als* "when" demands the expanded use of these tags. Similarly *für* "for" and *zum* "to the" can now be conjunctions that govern an infinitive (*KOUI*).

In contrast to StG, prepositions can be combined with any article. In consequence *APPRART* is also applicable for plural forms as in *id* "into the" or indefinite articles as in *ime* "in a".

Lacking a corresponding form, the tag *PRELAT* for attributive relative pronouns will not be used.

Additions: For infinitive particles like *go* or *cho* we decided to use the tag *PTKINF* like in the *Swiss SMS Corpus* and in *NOAH's Corpus*.

For merged words like *hets* "there is" (literally "has it") we copied the treatment from Hollenstein and Aepli (2014) with the plus sign. *hets* is therefore tagged with *VAFIN+PPER*. Unlike in their Part-of-Speech tagging task, for our morphological analysis task all tags must be kept.

A completely new tag is *PTKAM* for the particle *am* (literally "at the") in the progressive verb form. In StG examples like *Ich bin am Schreiben* literally "I am at-the writing", *Schreiben* is commonly analyzed as a substantified verb forming a prepositional phrase together with *am*. In SwG this construction is expanded with verbal objects more often than in most areas outside Switzerland (Van Pottelberge, 2005). Such an example would be *Ich bi en Brief am schriibe* literally "I am a letter at-the writing". The fact that *am* here stands between the object and the infinitive makes an analysis as prepositional phrase impossible and speaks against the tag *APPRART* for *am*. The comparison with *en Brief z schribe* "to write a letter" with the particle *z* is a good argument for *am* to be analyzed as a particle too. Our tag would also make sense for other varieties of the German language where such constructions occur or where their interpretation as verbal forms are preferred over one as preposition–noun sequences.

3.2 Morphological features

Due to the absence of an established morphological tagset for SwG, we defined a character based tagset that extends the STTS to *STTS.gsw*. The characters that make up the tags are listed in table 1.

Category	Tags
Degree	**p** (positive), **c** (comparative), **s** (superlative)
Person	**1** (first), **2** (second), **3** (third)
Case	**n** (nom.), **a** (acc.), **d** (dat.), **r** (nom./acc.)
Number	**s** (singular), **p** (plural)
Gender	**m** (masc.), **f** (fem.), **n** (neutral)
Mode	**i** (indicative), **j** (subjunctive I), **k** (subjunctive II)
Inflection	**s** (strong), **w** (weak)
Definiteness	**i** (indefinite), **d** (definite)

Table 1: Morphological tags.

We decided against a tag for the mixed adjective inflection that is used by many descriptions of the StG language. The reasons behind this are that this distinction is solely syntactic and that different SwG dialects use the strong and weak inflection differently.

As there is no past preterite, the category *time* could be spared. In consequence the two subjunctive tenses are interpreted as different modes (as *subjunctive I* and *II* instead of *subjunctive present* and *preterite*).

We introduced a shared tag for nominative or accusative cases even though this would constitute a large intervention from a linguistic perspective. As only personal and some related pronouns make a distinction between these cases, different tags for these forms would lead to competing analyzes that could only be distinguished through semantics. Therefore we exclude the task of disam-

biguating these cases but mark this with the tag *r* (from *rectus*).

In our example *hets*, the tag *VAFIN* is extended with *3si* (3rd person, singular, indicative) and *PPER* is extended with *3snn* (3rd person, singular, neutral, nominative).

3.3 Lemmas

For the choice of lemmas we decided to follow the rules from the *Swiss SMS Corpus* to ensure compatibility between different resources for SwG. Their main principles are that closely related words must be used, no new StG words must be invented and that the meaning should not be changed (Ueberwasser, 2013).

For example *hets* is annotated as *haben/VAFIN.3si+es/PPER.3snn* after including the morphological tags and lemmas.

4 Material

4.1 Corpus

For the calculation of the weights and for testing we annotated two sets of around 14 500 tokens taken from *NOAH's Corpus* using the morphology analysis tool in its development stage and selecting or adding the correct analysis.

4.2 Standard German resources

To avoid having to collect word stems and classifying them by inflection class, we took the allomorph list from *Morphisto* (Zielinski et al., 2009). Our material taken from this source counts 7833 nouns, 4300 verbs, 3178 adjectives, 1052 proper nouns and 781 adverbs. We used the lemma stem for our lemmas and the allomorph stems for later converting to SwG sounds. The inflection classes enable us to select the right endings in SwG and the word frequency classes are used as base for the weights of our tool.

With this connection to StG, the selection of stems can easily be changed without the need to collect more SwG stems and the lemmas are consistent with the resources used in this task.

For words from other parts of speech we had to take the frequency class from the *DeReWo* list from IDS (2012).

4.3 Swiss German resources

The forms of the closed word classes like pronouns, particles and similar were added with consulting dialect grammars from Weber (1948), Marti (1985b) and Suter (1992). In addition we added 11 for adjectives plus ordinal numbers (as *ADJA*), 127 adverb stems, about 50 noun stems and around 90 full verb stems (21 roots plus different prefixes) that cannot easily be derifed from StG forms.

5 Implementation

Our system is intended to be run with the *Helsinki Finite-State Transducer Technology* (HFST) (Lindén et al., 2009). HFST allows building and applying weighted finite-state transducers with tropical semi-rings. That means paths can be punished with weights that are added on the way and the paths with the lowest weights are preferred.

5.1 Forms

The implementation of the SwG word forms happens in two stages. The first stage is producing a hidden layer that represents the phonemes of different dialects. For open word classes like nouns or verbs we use replacement rules that we apply on the stems from *Morphisto*. For example *Zeit* "time" has to be converted to `zīt` while *heiß* "hot" will become `haĭss`. Those different replacements (see figure 1) for ⟨ei⟩ will be weighted by their probability, including phonological context as far as possible.

```
# ei before er
define EI1 [ {eier} (->) {īr}::0.1 ];
define EI2 [ {eier} (->) {īĕr}::4.7 ];
define EI3 [ {eier} -> {aĭĕr}::5.4 ];
# ei else
define EI4 [ {ei} (->) {ī}::0.9 ];
define EI5 [ {ei} (->) {ĩ}::5.4 ];
define EI6 [ {ei} -> {aĭ}::1.1 ];
# combined rule for ei
define EI [ EI1 .o.  EI2 .o.  EI3 .o.
EI4 .o.  EI5 .o.  EI6 ] ;
```

Figure 1: Replacement rules for ⟨ei⟩. First EI1–EI3 deal with ⟨ei⟩ before ⟨er⟩, then EI4–EI6 replace ⟨ei⟩ in all other cases. Higher weights indicate less frequent options.

For the closed word classes and words that do not exist in StG like *gheie* "to drop" we wrote the forms directly in phonemes.

In the second stage these phonemes are replaced by dialect specific spellings using a different set of rules for every dialect. Here we limited using weights to specific sound changes that are not represented by the chosen phonemes, as in most cases

list	phonemes	dialects	translation
Zeit	zīt	Ziit, zyt	“time”
Zeit-e	zītĕ	ziite, Zytä	“times”
ghĩĕ		gheie, ghie	“[I] drop”
ghĩt		gheit, ghiit	“[it] drops”
mīn		min, miin	“my”
mīnĕ		myne, minä	“mine”

Table 2: Form generation for open classes over a step in-between and for exclusively SwG words and closed classes directly in phonemes.

the different results are just different spellings for the same sounds. So far we made dialect modules for Basel, Bern and Zurich. Table 2 gives some examples how the different forms are generated.

Clitics are added between these steps and flag diacritics – a feature offered by HFST – are used to disable ungrammatical combinations.

5.2 Weights

The way the weights are calculated is motivated by the tropical semi-ring and the word frequency classes. For every competing alternative at a decision (e. g. in replacement rules), the the absolute value of the binary logarithm of the probability is added to the weight. Using this formulae all the weights are in the same currency and the different reasons for weights can be treated the same.

6 Results

6.1 Coverage

For 79% of the tokens in the test corpus, our system could produce the correct analysis according to their positions. With exclusion of foreign language material (*FM*), named entities (*NE*) and non-words (*XY*) this quota reached 86%.

In the blog data we could even observe that 90.8% of the tokens (without *FM*, *NE* and *XY*) could be reached. On the other side, the business report and wikipedia entries proved to be more difficult with 81.7% resp. 81.9%.

Table 3 shows the coverage for all tokens and some selected parts of speech. The closed word classes like negation particles (*PTKNEG*), indefinite pronouns (*PIDAT*) or interrogative pronouns (*PWS*) are fully covered. The open word classes like named entities (*NE*), nouns (*NN*) and adjectives (*ADJA*) are more difficult. While it was not the goal to include a lot of named entities, the nouns are

Parts of speech	correct
all	0.790
w/o FM, NE, XY	0.860
NN	0.583
ART	**0.970**
NE	*0.129*
APPR	**0.959**
VAFIN	**0.980**
ADJA	0.662
APPRART	**0.970**
KON	**0.992**
VVPP	0.851
VVFIN	0.881
PTKNEG	**1.000**
PIDAT	**1.000**
PWS	**1.000**

Table 3: Coverage of all tokens in the test corpus, the most frequent PoS and some selected PoS.

an open problem. Like most Germanic languages, SwG allows building a theoretically unlimited number of compounds which were hard to grasp and which shows in the low coverage in our system. Similarly also adjectives can be derived from other word classes. The most frequent case of this type proofed to be participles that had been turned into adjectives and declined accordingly.

6.2 Weights

Evaluating weights in a group of non-stanardized dialects is difficult because different speakers might not agree on what analyses are acceptable or not. Hence we chose a purely data driven approach which compares the ranking be our system with a random order of analyses using the *mean reciprocal rank* (MRR) (Büttcher et al., 2010, p. 409) and (Neumann, 2010, p. 587). The MRR averages the multiplicative inverse of the rank of the first correct solution for all evaluated tokens. To reduce the impact from uncovered words, we only looked at them where the correct solution is provided by the system.

The overall MRR of 0.843 by the system compared to the 0.531 for random orders shows that the weights successfully order the analyses.

Besides the overall MRR, table 4 shows that both open word classes like verbal participles (*VVPP*) and closed word classes like cardinal numbers *CARD* profit from the weights. Parts of speech like infinitives with *zu* “to” (*VVIZU*) could also do

Parts of speech	system	random
all	**0.843**	0.531
w/o FM, NE, XY	**0.842**	0.530
NN	**0.911**	0.595
ART	**0.721**	0.386
NE	**0.957**	0.783
APPR	**0.759**	0.375
VAFIN	**0.931**	0.450
ADJA	**0.439**	0.324
APPRART	**0.792**	0.423
KON	**0.908**	0.474
VVPP	**0.983**	0.645
VVFIN	**0.708**	0.349
CARD	**0.993**	0.886
VVIZU	1.000	1.000
PTKZU	*0.348*	0.587
PTKA	*0.417*	0.459

Table 4: Mean reciprocal rank on the correctly analyzed tokens, for all PoS, the most frequent and for some selected PoS. The random numbers set the baseline.

without weights.

On the other hand particles before infinitives (*PTKZU*) or before adjectives (*PTKA*) even suffer from weights. The cause there is that they are beaten by more frequent prepositions of the same form and are thus always on a deeper rank.

A small profit can be seen with adjectives (*ADJA*). There a large number of analyses for certain types pull down the MRR. For example *schöööni* "beautiful" there can be found up to 5 valid analyses (out of 24).

For these problems with particles and adjectives, a word based procedure cannot solve the problem. However, with a language model the problem of competing analyses should be solved easily.

7 Conclusion

With a token coverage of the treated parts of speech of 86% up to 90% on selected texts our system clearly can help with the production of annotated resources for the SwG dialects.

An open problem is still the low coverage on nouns due to large potential to build new words. Enabling composition and derivation is a possible answer to this problem. For words unknown due to the lack of corresponding StG words, adding more stems seems the best way.

For the future we see much potential in language models to be able to distinguish between competing analyses. For this task our corpus can be used as training data. Experiments will have to decide if the weights can be used as emission models.

Another task for the future is the expansion of the program to process more dialects. Especially the alpine dialects differ from those covered here and could profit from this.

Acknowledgments

I would like to thank Simon Clematide from University of Zurich for his help and valuable input during this project. Also, I am thankful to Noëmi Aepli for explanations about the tagging in NOAH's corpus.

References

Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. 2010. *Information Retrieval - Implementing and Evaluating Search Engines*. MIT Press, Cambridge MA, USA.

Helen Christen, Elvira Glaser, and Matthias Friedli, editors. 2012. *Kleiner Sprachatlas der deutschen Schweiz*. Huber, Frauenfeld, Switzerland, 4th edition.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*. Lebendige Mundart. Sauerländer, Aarau etc., Switzerland, 2nd edition.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann, editors, *COLING 2014, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 85–94.

IDS Institut für Deutsche Sprache, Programmbereich Korpuslinguistik. 2012. Korpusbasierte Wortgrundformenliste DeReWo, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation. `http://www.ids-mannheim.de/derewo`.

Krister Lindén, Miikka Silfverberg, and Tommi A. Pirinen. 2009. HFST tools for morphology - an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology. Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009, Zurich, Switzerland, September 2009. Proceedings*, pages 28–47.

Werner Marti. 1985a. *Bärndütschi Schrybwys: ein Wegweiser zum Aufschreiben in berndeutscher Sprache: mit einer Einführung über allgemeine Probleme des Aufschreibens und einem Wörterverzeichnis nebst Beispielen*. A. Francke, Bern, Switzerland, 2nd edition.

Werner Marti. 1985b. *Berndeutsch-Grammatik für die heutige Mundart zwischen Thun und Jura*. A. Francke, Bern, Switzerland.

Günter Neumann. 2010. Text-basiertes Informationsmanagement. In Kai-Uwe Carstersen, Christian Ebert, Cornelia Ebert, Susanne J. Jekat, Ralf Klabunde, and Hagen Langer, editors, *Computerlinguistik und Sprachtechnologie. Eine Einführung*, pages 576–615. Spektrum Akademischer Verlag, Heidelberg, Germany, 3rd edition.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. Archimob - a corpus of spoken Swiss German. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Yves Scherrer. 2011. Morphology generation for Swiss German dialects. In *Systems and Frameworks for Computational Morphology - Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings*, pages 130–140.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`.

Beat Siebenhaar and Alfred Wyler. 1997. *Dialekt und Hochsprache in der deutschsprachigen Schweiz*. Edition „Pro Helvetia“, Zurich, Switzerland, 5th edition.

Elisabeth Stark, Simone Ueberwasser, and Beni Ruef. 2009–2015. Swiss SMS Corpus. `https://sms.linguistik.uzh.ch`.

Rudolf Suter. 1992. *Baseldeutsch-Grammatik*. Grammatiken und Wörterbücher des Schweizerdeutschen in allgemeinverständlicher Darstellung. Christoph-Merian-Verlag, Basel, Switzerland, 3rd edition.

Simone Ueberwasser. 2013. Non-standard data in Swiss text messages with a special focus on dialectal forms. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research. (=TSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln 5. Hrsg: Christiane M. Bongartz und Claudia M. Riehl)*, pages 7–24, Aachen. Shaker Verlag.

Jeroen Van Pottelberge. 2005. Ist jedes grammatische Verfahren Ergebnis eines Grammatikalisierungsprozesses? Fragen zur Entwicklung des *am*-Progressivs. In Thorsten von Leuschner, Tanja Mortelmans, and Sarah Groodt, editors, *Grammatikalisierung im Deutschen*, pages 169–192. De Gruyter, Berlin, Germany.

Albert Weber and Bund Schwyzertütsch. 1948. *Zürichdeutsche Grammatik: ein Wegweiser zur guten Mundart*. Grammatiken und Wörterbücher des Schweizerdeutschen in allgemeinverständlicher Darstellung. Schweizer Spiegel-Verlag, Zurich, Switzerland.

Andrea Zielinski, Christian Simon, and Tilman Wittl. 2009. Morphisto: Service-oriented open source morphology for German. In *State of the Art in Computational Morphology - Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009, Zurich, Switzerland, September 2009. Proceedings*, pages 64–75.