Normalising orthographic and dialectal variants for the automatic processing of Swiss German

Tanja Samardžić[*], **Yves Scherrer**[†], **Elvira Glaser**[‡]

[*]CorpusLab, URPP Language and Space, University of Zurich, Freiestrasse 16, 8032 Zürich, Switzerland
[†]LATL-CUI, University of Geneva, Route de Drize 7, 1227 Carouge, Switzerland
[‡]German Department, University of Zurich, Schönberggasse 9, 8001 Zürich, Switzerland
tanja.samardzic@uzh.ch, yves.scherrer@unige.ch, eglaser@ds.uzh.ch

Abstract

Swiss dialects of German are, unlike most dialects of well standardised languages, widely used in everyday communication. Despite this fact, they lack tools and resources for natural language processing. The main reason for this is the fact that the dialects are mostly spoken and that written resources are small and highly inconsistent. This paper addresses the great variability in writing that poses a problem for automatic processing. We propose an automatic approach to normalising the variants to a single representation intended for processing tools' internal use (not shown to human users). We manually create a sample of transcribed and normalised texts, which we use to train and test three methods based on machine translation: word-by-word mappings, character-based machine translation, and language modelling. We show that an optimal combination of the three approaches gives better results than any of them separately.

1. Introduction

Swiss German refers to a range of German varieties spoken in the Northeastern two thirds of Switzerland. Swiss German dialects are widely used in speech, while standard German is used almost exclusively in written contexts. It is usually admitted that the sociolinguistic configuration of German-speaking Switzerland is a model case of diglossia, i.e. an environment in which two linguistic varieties are used complementarily in functionally different contexts.

Despite the preference for spoken dialect use, written Swiss German use has become popular in electronic media like blogs, SMS, e-mail and chatrooms; the Alemannic Wikipedia contains about 6000 articles written in Swiss German. This increased use of dialects in the written domain has not been followed by a development of processing tools, which is why Swiss German still belongs to less-resourced languages.

The rich resources and tools developed for standard German are of little use in treating Swiss German. Dialectological studies show considerable differences between standard and Swiss German not only in the lexicon and pronunciation, but also in morphology and syntax. The work of Hollenstein and Aepli (2014) on part-of-speech tagging shows that better performance is reached if the tools are trained on a small Swiss German corpus, than on a large standard German corpus. Processing Swiss German requires an independent approach that needs to address a range of issues characteristic for non-standard and spoken varieties in addition to the well-known regional variability.

This paper addresses the orthographic inconsistency and dialectological variation typical for Swiss German texts. Normalisation, i.e. mapping the variants of what can be identified as the same word to a single representation, is necessary for an efficient search of Swiss German sources such as web content or social media. While this step might not be crucial for tasks such as part-of-speech tagging (Hollenstein and Aepli, 2014), it is essential for any task that requires establishing lexical identities. Such tasks include building an efficient corpus query interface for linguistic research, semantic processing, and information retrieval.

We propose an approach to automatic normalisation that casts the task as simplified machine translation from inconsistently written texts to a unified representation. The resulting normalisation is treated as word-level annotation which is internally used for executing search queries, but is not intended to be presented to human users.

2. Related work

Swiss German has been the object of extensive dialectological research for more than 100 years. One of the major contributions is the *Sprachatlas der deutschen Schweiz* (SDS), a linguistic atlas that covers phonetic, morphological and lexical differences (Hotzenköcherle et al., 1962–1997). Other – still ongoing – projects include the *Idiotikon* (Staub et al., 1881–), a comprehensive dialect dictionary covering the Swiss German varieties, and a syntactic atlas (Bucheli and Glaser, 2002).

Recent work in computational dialectology relies on structured data from atlases and dictionaries. This work includes making the collected data electronically available (Bickel, 2006; Scherrer and Rambow, 2010), creating models of morphological and syntactic variation (Scherrer, 2011a; Scherrer, 2011b), studying dialect variation with dialectometrical methods (Goebl et al., 2013; Jeszenszky and Weibel, 2014), and comparing a sample of SDS data with present-day data collected through crowdsourcing (Kolly and Leemann, in press).

Unstructured textual data have started being processed only recently. Scherrer (2012) uses a small corpus for exploratory experiments in corpus-based dialectometry. Hollenstein and Aepli (2014) collect a corpus of written Swiss German texts and use it to train and test part-of-speech models. Stark et al. (2009–2015) compile, normalise, and part-of-speech tag a corpus of SMS messages. An archive of short recordings is being digitised by Studer-Joho (2011).

Regarding normalisation, dialect texts face problems that are similar to other types of non-standard data such as historical, spoken or computer-mediated communication (chat, Twitter, SMS, . . .) data. Automatic word normalisation has been a popular topic in historical NLP over the last few years. For instance, the VARD 2 tool (Baron and Rayson, 2008) approaches word normalisation as a particular case of spellchecking and provides a complete environment with an editor, a rule induction mechanism operating on already corrected texts, and a combination of several mechanisms to guess the modern word from a historical one. Bollmann (2012) presents a method to automatically extract rewrite rules from a parallel corpus. These rules are then used to normalise historical word forms, either on their own or in conjunction with a filtering step based on a reference lexicon of the modern language. Pettersson et al. (2013) do not induce explicit transcription rules, but rather find for each historical word its most similar modern counterpart in a lexicon; similarity is measured with Levenshtein distance. These methods are primarily useful for treating small edits in words that are mostly similar.

More recently, character-level statistical machine translation (CSMT) has been successfully applied to normalisation of computer-mediated communication (De Clercq et al., 2013; Ljubešić et al., 2014) and historical texts (Pettersson et al., 2013b; Pettersson et al., 2014; Scherrer and Erjavec, 2015). This method has originally been proposed for translation between closely related languages (Vilar et al., 2007; Tiedemann, 2009). It requires less training data than word-level SMT but is limited to applications where regular changes occur at the character level.

For Swiss German dialects, word normalisation has already been manually performed by Stark et al. (2009–2015). The normalisation strategy proposed in our study is slightly different. The goal of Stark et al. is primarily to represent the content of the SMS messages, while we aim at representing morphosyntactic differences between variants, which is necessary for studying formal linguistic variation.

3. Neutralising the variation

Variation in written Swiss German is observed at two levels. First, a lexical unit that can be identified as "the same word" is pronounced, and therefore also written, in a different way in different regions of Switzerland. Second, a lexical unit that can be considered phonetically invariant (within a region) is written in a different way on different occasions. This is due to the lack of a widely accepted orthographic standard, and to occasional intra-speaker variation. Writing is thus not standardised and highly inconsistent. A set of transcription recommendations, proposed by Dieth (1986), is often used in expert transcriptions. However, these recommendations tend to be interpreted and implemented in different ways, resulting in inconsistencies even within a single text transcribed by the same expert.

The two types of variation combined result in a great number of potential variants that need to be reduced to a single form in order to establish identity between words that are felt to be the same across variants. We illustrate in Table 1 the range of potential variation by an arbitrarily chosen segment from our corpus (details below). The table shows all the strings that are assigned the same normalisation. We can see that different variants of common words are found even in the same document (Var1), that is within a sample of the size of around 10 000 tokens transcribed by the same trained expert. In addition to these, many more variants are found in other documents containing samples from other varieties (Var2). The shown variants include cases of morphosyntactic syncretism (e.g. *mis, miis*, which are neuter forms of masculine *min, miin* and are normalised by the same word *mein*), variants caused by code-switching (e.g. *määin, mäin, main, man, hat*), and transcription errors (e.g. *hüt, händ, haa, iiimer*).

Var2	mine mini määin mäin main	man	hed hèd hèt heet haa händ hüt	ime imer emmer iiimer	gsäit gsääit
Var1	mi mii miin mis miis	ma	hat hät		gsait
Orig.	*min*	*maa*	*het*	*immer*	*gsaait*
Norm.	**mein**	**mann**	**hat**	**immer**	**gesagt**
Trans.	My husband has always said...				

Table 1: A segment of a transcribed (Orig.) and normalised (Norm.) text with corresponding variants found in the same document (Var1) and in other documents (Var2).

As it can be seen in Table 1, the normalised forms resemble standard German. However, they diverge from standard German in two aspects: lexical mismatches and word boundaries.

Lexical mismatches Many Swiss German lexical items do not have any etymologically related standard counterparts. For example, Swiss *öpper* corresponds to standard *jemand* 'someone', *töff* to *Motorrad* 'motorbike', and *gheie* to *fallen* 'to fall'. Two normalisation approaches are possible here. One can replace the Swiss items with the semantically equivalent (but formally dissimilar) standard items. The choice of these items is a matter of interpretation and a potential source of inconsistency, which should be avoided in an annotation designed to eliminate the original inconsistencies. Alternatively, one can propose specific dialect-independent items that are usually based on a historically common form, but that are not used in contemporary standard German. A disadvantage of this approach is that it renders the normalised text artificial and hard to understand. It also requires expertise in history of German.

With a goal to represent the local varieties as accurately as possible, we do not translate specific Swiss lexical items, but rather normalise them using a convenient, etymologically motivated common construction. Thus, *öpper* is normalised as *etwer*, *töff* as *töff*, and *gheie* as *geheien* (these normalised forms do not exist in standard German); Swiss German *vorig* 'remaining' is normalised as *vorig*, although this word means 'previous' in standard German.

Word boundaries Standard conventions regarding word boundaries are often not applicable to Swiss German, where articles and pronouns tend to be cliticized. For example, *hettemers* corresponds to the standard German sequence *hätten wir es*, and *bimene* corresponds to *bei einem*. Again, one can either impose standard boundaries, which requires considerable reconstruction and departure from the actual variety, or one can introduce word boundaries that correspond better to the intuition of native speakers, also observable in spontaneous writing. The latter choice leads to potential inconsistency because the decisions on what is written as one word need to be taken case by case. We decide to keep the standard word boundaries whenever this is possible. Thus, *hettemers* is normalised as *hätten wir es*, *bimene* as *bei einem*.

An important feature of our approach is that we regard normalisation as a hidden annotation layer used only for automatic processing. The users are expected to formulate queries and the results are presented in a form of original writing (keeping the original inconsistency). This allows us to choose arbitrary representations, which users would find artificial and hard to adopt.

4. Manual normalisation

To construct the data set for our initial experiments in automatic normalisation, we selected a sample of documents from a corpus of transcribed Swiss German under construction. The corpus contains transcriptions of video recordings collected by the ArchiMob association[1] in the period 1999–2001. Informants come from all linguistic regions of Switzerland and represent both genders, different social backgrounds, and different political views. Each recording is produced with one informant using a semi-directive technique and is between 1h and 2h long.

We selected transcriptions of 6 documents and annotated them manually with the normalised forms, as described in the previous section. These 6 documents were transcribed by a single expert and normalised by 3 other expert annotators. To ensure the consistency of annotation, we produced guidelines which listed case-based decisions. We also used annotation tools that allowed annotators to quickly look up previous normalisations for each word which had already been normalised. We initially used VARD 2 (Baron and Rayson, 2008), but we later switched to the better adapted IGT tool (Ruef and Ueberwasser, 2013).

5. Optimising machine translation for normalisation

We approach the task of automatic normalisation by adapting statistical machine translation techniques to the specific properties of the process. The gold standard normalisation is assigned at the word-level, which results in the same word order in the originally transcribed and in the normalised text. This makes the task easier than standard machine translation, because word alignment does not need to be computed.

On the other hand, learning the mapping between original and normalised words is complicated by other factors. First, normalisation depends on the context. For example, *es* in *es git chrieg* 'there will be war' is normalised as *es*, while its normalisation in *es ross* 'a horse' is *ein*. Second, Swiss German transcriptions tend to contain words whose normalisations consist of several words (see above). Third, as illustrated in Table 1, the number of possible variants for many words is very big, which means that the mappings are hard to learn in a small sample such as ours.

5.1. Training and test data

For our experiments, we use the 6 documents with manual normalisations for training and testing. In particular, we perform cross-validation on 5 folds, each of which contains 80% of each of the 6 documents as training data and the remaining 20% of each document as test data. Hence, each fold contains data from all six dialects, in the training set as well as in the test set. The training sets consist of 848 turns each (between 75 070 and 85 031 words) and the test sets of 212 turns each (between 14 668 and 24 629 words). This setup simulates a realistic scenario where the dialect of the data to be normalised is unknown but assumed to be partially observed at training.

We partition the words occurring in the test sets in four classes, depending on the applicable normalisation methods: *Unique* words are associated with exactly one normalisation in the corresponding training set; *Ambiguous 1* words are associated with more than one normalisation candidate, but a unique most frequent normalisation can be determined; for *Ambiguous 2* words, no single most frequent normalisation can be selected because of tied frequency counts; *New* words have not been observed in the training set and therefore no normalisation candidates are available. The proportions of these four classes, averaged over the 5 folds, are given in the first column of Table 2.[2]

5.2. Experiments and results

No normalisation As a baseline, we do not normalise the dialect words at all. This yields an average accuracy of 20.28%, i.e. for one in five words, the original form is identical to the normalised form (Table 2, second column).

Word-by-word translation We first test a simple approach to normalisation, where we define mappings between originally transcribed and normalised forms as a word-by-word translation model with fixed alignment. In this setting, we select the normalisation that is most frequently associated with the target original form in the training set for the *Unique* and *Ambiguous 1* classes. We randomly choose one of the most frequent normalisations in

[1] More information available at http://www.archimob.ch

[2] An alternative setting, potentially leading to a bigger proportion of unknown words, would be a document-based split. The proportion of new words in this alternative setting varies between 8.78% and 15.48% with an average of 13.27% that is close to 11.27% in our setting.

	Proportion (%)	Accuracy (%)					
		No norm.	Word-by-word	CSMT	Archimob LM	Extended LM	Combi.
Unique	44.53	23.06	**87.83**	87.83	87.83	87.83	87.83
Ambiguous 1	43.68	21.92	**80.63**	80.63	70.11	67.12	80.63
Ambiguous 2	0.52	4.17	**38.43**	**39.81**	**46.06**	**49.54**	49.54
New	11.27	3.64	3.64	**23.88**	23.88	23.88	23.88
All	100	20.28	74.94	77.23	72.67	71.38	77.28

Table 2: Accuracies of the automatic normalisation methods for the different word classes, averaged over the 5 folds.

the *Ambiguous 2* class, whereas we use the original, non-normalised form for *New* words.

Results are shown in the third column of Table 2.[3] This method yields accuracy values of more than 80% for the first two classes, which represent 88.21% of data. However, it performs less well where frequency cues are missing (*Ambiguous 2*) and obviously, where no normalisations at all have been observed at training time (*New*).

Character-based translation To address the unknown words (*New*), we train a CSMT model. Character-level SMT is different from standard (i.e. word-level) SMT in that, instead of aligning words occurring in sentence pairs, one aligns characters occurring in word pairs. The resulting translation models contain phrases which consist of character sequences instead of word sequences, and language models are trained on character n-grams instead of word n-grams. This system will thus learn the most frequent character correspondences used to normalise dialect words, and will be able to generate normalisations also for the words that were not seen in the training set. We also apply the CSMT system to disambiguate the possible normalisations in the *Ambiguous 2* class.[4]

As shown in the CSMT column of Table 2, this model improves normalisation of unseen words by 20% absolute, and is also able to slightly improve the disambiguation of *Ambiguous 2* words compared to random selection.

Language models To account for the fact that normalisation depends on the context of the target word, we add a language model to disambiguate the normalisations of the *Ambiguous 1* and *2* classes. We first learn the language model on the normalised part of the Swiss German training set (*Archimob LM* in Table 2). Such a model represents well our target text, but the training corpus is small. As an alternative, we learn a larger language model by extending the training set with the TüBa-D/S corpus (Hinrichs et al., 2000), which contains transcribed standard German speech data, i.e. a similar genre as the one of our dialect corpus (*Extended LM* in Table 2). This allows us to build a language model that is more reliable but also noisier with respect to our target data. In both cases, we create trigram language models using IRSTLM (Federico et al., 2008).

The Archimob language model improves the disambiguation of *Ambiguous 2* normalisations, but is not able to outperform the simple maximum likelihood selection used in the word-by-word model for the *Ambiguous 1* class. Due to the higher proportion of *Ambiguous 1* words, the overall accuracy decreases. This effect is intensified with the extended language model.

Combination The best overall accuracy is obtained with word-by-word normalisation for the *Unique* and *Ambiguous 1* classes, CSMT for *New* words, and with the extended language model for *Ambiguous 2* words (Table 2, last column). The resulting accuracy lies 2.34% absolute above the word-by-word model and 57% above the baseline.

6. Conclusion and future work

This paper formulates the main challenges of normalising Swiss German and tests several methods to automatically normalise unseen texts. We have shown that a relatively good automatic normalisation of a wide range of variants in Swiss German can be obtained using a small training set. The key to the success is an appropriate combination of different methods that takes advantage of the specific properties of the task.

In future work, we will increase the training set and investigate optimisation techniques to improve combinations of methods, focusing more on highly ambiguous and unknown words, which still pose a problem for all the methods presented in this paper.

7. References

Baron, Alistair and Paul Rayson, 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Aston University.

Bickel, Hans, 2006. Idiotikon digital: Überlegungen zu einer elektronischen Ausgabe des Schweizerdeutschen Wörterbuchs. In *Schweizerdeutsches Wörterbuch – Bericht über das Jahr 2006*. Zürich.

Bollmann, Marcel, 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*. Lisbon.

[3] Figures in bold indicate the word classes for which the method provides new results.

[4] In this experiment, each unique original–normalised word pair forms a training instance. Spaces are inserted between the characters of each word so that each character is interpreted as an atomic translation entity. We used the Moses toolkit (Koehn et al., 2007) with the following settings: the language model is trained on the same data, using character 6-grams (following earlier experiments); Good-Turing discounting is applied to smooth translation probabilities; distortion is disallowed; we use default weights instead of MERT tuning for the different parameters of the translation model due to the small sizes of our training sets.

Bucheli, Claudia and Elvira Glaser, 2002. The syntactic atlas of Swiss German dialects: Empirical and methodological problems. In Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij (eds.), *Syntactic Microvariation*, volume 2. Amsterdam: Meertens Institute Electronic Publications in Linguistics.

De Clercq, Orphée, Bart Desmet, Sarah Schulz, Els Lefever, and Véronique Hoste, 2013. Normalization of Dutch user-generated content. In *Proceedings of RANLP 2013*. Hissar.

Dieth, Eugen, 1986. *Schwyzertütschi Dialäktschrift*. Aarau: Sauerländer, 2nd edition.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo, 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech 2008*. Brisbane.

Goebl, Hans, Yves Scherrer, and Pavel Smečka, 2013. Kurzbericht über die Dialektometrisierung des Gesamtnetzes des Sprachatlasses der deutschen Schweiz (SDS). In Karina Schneider-Wiejowski, Birte Kellermeier-Rehbein, and Jakob Haselhuber (eds.), *Vielfalt, Variation und Stellung der deutschen Sprache*. De Gruyter, pages 153–176.

Hinrichs, Erhard W., Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann, 2000. The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.

Hollenstein, Nora and Noëmi Aepli, 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*. COLING 2014, Dublin.

Hotzenköcherle, Rudolf, Robert Schläpfer, Rudolf Trüb, and Paul Zinsli (eds.), 1962–1997. *Sprachatlas der deutschen Schweiz*. Bern: Francke.

Jeszenszky, Péter and Robert Weibel, 2014. Correlating morphosyntactic dialect variation with geographic distance: Local beats global. In *GIScience 2014: Eighth International Conference on Geographic Information Science*, number 40 in GeoInfo Series. Vienna. Extended abstract.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 demonstration session*. Prague.

Kolly, Marie-José and Adrian Leemann, in press. Dialäkt Äpp: communicating dialectology to the public crowdsourcing dialects from the public. In Adrian Leemann, Marie-José Kolly, Volker Dellwo, and Stephan Schmid (eds.), *Trends in Phonetics and Phonology. Studies from German-speaking Europe*.

Ljubešić, Nikola, Tomaž Erjavec, and Darja Fišer, 2014. Standardizing tweets with character-level machine translation. In *Proceedings of CICLing 2014*, Lecture notes in computer science. Kathmandu: Springer.

Pettersson, Eva, Beáta B. Megyesi, and Joakim Nivre, 2013a. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (Nodalida 2013)*. Oslo.

Pettersson, Eva, Beáta B. Megyesi, and Joakim Nivre, 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Gothenburg.

Pettersson, Eva, Beáta B. Megyesi, and Jörg Tiedemann, 2013b. An SMT approach to automatic annotation of historical text. In *Proceedings of the Nodalida Workshop on Computational Historical Linguistics*. Oslo.

Ruef, Beni and Simone Ueberwasser, 2013. The taming of a dialect: Interlinear glossing of Swiss German text messages. In Marcos Zampieri and Sascha Diwersy (eds.), *Non-standard Data Sources in Corpus-based Research*. Aachen.

Scherrer, Yves, 2011a. Morphology generation for Swiss German dialects. In Cerstin Mahlow and Michael Piotrowski (eds.), *Systems and Frameworks for Computational Morphology, Second International Workshop (SFCM 2011)*. Berlin, Heidelberg: Springer.

Scherrer, Yves, 2011b. Syntactic transformations for Swiss German dialects. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Edinburgh.

Scherrer, Yves, 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Workshop on Visualization of Language Patterns and Uncovering Language History from Multilingual Resources*. Avignon.

Scherrer, Yves and Tomaž Erjavec, 2015. Modernising historical Slovene words. *Natural Language Engineering*:1–25. Available on CJO.

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

Stark, Elisabeth, Simone Ueberwasser, and Beni Ruef, 2009–2015. Swiss SMS corpus, University of Zurich, https://sms.linguistik.uzh.ch.

Staub, Friedrich, Ludwig Tobler, Albert Bachmann, Otto Gröger, Hans Wanner, and Hans-Peter Schifferle (eds.), 1881–. *Schweizerisches Idiotikon : Wörterbuch der schweizerdeutschen Sprache*. Frauenfeld: Huber.

Studer-Joho, Dieter, 2011. Digitising vernacular recordings: Preservation efforts at the phonogram archives of the University of Zurich. In *Preserving Endangered Audio Media — Rethinking Archival Strategies for Conservation of Analogue Audio Carriers*. Berlin: Staatliche Museen zu Berlin.

Tiedemann, Jörg, 2009. Character-based PSMT for closely related languages. In *Proceedings of EAMT 2009*. Barcelona.

Vilar, David, Jan-Thorsten Peter, and Hermann Ney, 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague.