Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation

Scherrer, Yves; Ljubešić, Nikola

Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation

Yves Scherrer
CUI / Department of Linguistics
University of Geneva
Geneva, Switzerland
yves.scherrer@unige.ch

Nikola Ljubešić
Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Abstract

The Swiss German dialect corpus ArchiMob poses great challenges for NLP and corpus linguistic research due to the massive amount of variation found in the transcriptions: dialectal variation is combined with intra-speaker variation and with transcriber inconsistencies. This variation is reduced through the addition of a normalisation layer. In this paper, we propose to use character-level machine translation to learn the normalisation process. We show that a character-level machine translation system trained on pairs of segments (not pairs of words) and including multiple language models is able to achieve up to 90.46% of word normalisation accuracy, an error reduction of 45% over a strong baseline and of 34% over a heterogeneous system proposed by Samardžić et al. (2015).

1 Introduction

The term Swiss German covers a range of German varieties spoken in the Northeastern two thirds of Switzerland. Despite the widespread (almost exclusive) use of dialects in speech and in electronic media, only few resources adapted for research in NLP are currently available. One recent resource is the ArchiMob corpus of transcribed speech (Samardžić et al., 2016), which is used in the experiments presented here.

This paper addresses the orthographic inconsistency and dialectological variation typical for Swiss German texts through the addition of a normalisation layer. Normalisation, i.e. mapping the variants of what can be identified as the same word to a single representation, is necessary 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 to view the normalisation task as a translation task from inconsistently written texts to a unified representation. We show that a single, optimised character-level machine translation system fares better than the heterogeneous system proposed by Samardžić et al. (2015).

2 Related work

Swiss German has been the object of extensive dialectological research for more than 100 years, which has led to major contributions such as dialect atlases (Hotzenköcherle et al., 1962–1997; Bucheli and Glaser, 2002) and comprehensive dialect dictionaries (Staub et al., 1881–). However, dialect corpora have started being collected only recently. Siebenhaar (2005) creates a corpus of interactions in Swiss German internet relay chat rooms. Hollenstein and Aepli (2014) compile a corpus of written Swiss German texts and use it to train and test part-of-speech tagging models. Stark et al. (2009–2015) collect, normalise, and part-of-speech tag a corpus of SMS messages. The ArchiMob corpus used in this work has been presented together with first experiments on automatic normalisation and part-of-speech tagging in Samardžić et al. (2015) and Samardžić et al. (2016).

Due to the lack of standardised spelling, dialect texts face problems that are similar to other types of non-standard data such as historical text, spoken language or computer-mediated communication. Normalisation (also called modernisation in the context of historical language) has been proposed to deal with this heterogeneity. Automatic word normalisation has been addressed through several approaches in the community of historical NLP, such as automatic induction of rules (Reffle, 2011; Bollmann, 2012), similarity-based form matching inspired by spellchecking (Baron and Rayson, 2008; Pettersson et al., 2013), and character-level machine translation (Pettersson et al., 2014; Scherrer and Erjavec, 2015). Character-level machine

translation (CSMT) has originally been proposed for translation between closely related languages (Vilar et al., 2007; Tiedemann, 2009), but has proven successful in many other settings where regular changes occur at the character level, including the normalisation of computer-mediated communication (De Clercq et al., 2013; Ljubešić et al., 2014).

3 Data

The ArchiMob corpus contains transcriptions of video recordings collected in the context of an oral history project (see `http://www.archimob.ch`) between 1999 and 2001. Currently, the corpus consists of 34 transcriptions of interviews conducted in various Swiss German dialects (Samardžić et al., 2016).

The recordings were transcribed manually by native speakers of Swiss German, using the Dieth guidelines (Dieth, 1986). These are general guidelines that can be interpreted and implemented in several ways, leading to some inconsistencies in the transcriptions. Furthermore, there is a considerable amount of pronunciation variation in the texts, on the intra-speaker level as well as on the dialect level. For example, the first person possessive pronoun in its masculine singular form (Standard German *mein*) has been transcribed in the four variants *min*, *miin*, *mi*, *mii* in a single text, reflecting different pronunciations by the same speaker. When other texts are considered, a fifth variant, *mine*, can be added. While the transcription inconsistencies could be eliminated with more precise guidelines, there is no obvious way to reduce the intra-speaker variation and the dialectal variation at the transcription level. Therefore, it was decided to annotate each original word form with a normalised form. The goal of normalisation is to reduce all variants that can be identified as "the same word" to a single form. At the moment, a subset of 6 recordings have been manually normalised, and the plan is to normalise the remaining documents in a semi-automatic way.

Normalisation is performed word by word. In most cases, the normalised forms resemble Standard German (see Figure 1 for an example), with two major divergences from this principle. First, Swiss German lexical items that do not have an etymologically related Standard German counterpart are not translated, but rather normalised using a convenient, etymologically motivated common construction. For instance, *gluegt* 'looked' is not translated to the semantic equivalent *geschaut*, but normalised to the reconstructed form *gelugt*.[1] Second, word boundaries in Swiss German may differ from the Standard German ones due to cliticisation effects, in which case one Swiss German word corresponds to more than one word in the normalisation layer, as illustrated by *dasch* 'this is' and its normalisation *das ist*.

Transcription	Normalisation	
jaa	ja	'yes'
de	dann	'then'
het	hat	'has'
me	man	'one'
no	noch	'still'
gluegt	gelugt	'looked'
tänkt	gedacht	'thought'
dasch	das ist	'this is'
ez	jetzt	'now'
de	der	'the'
genneraal	general	'general'

Figure 1: A transcribed and normalised utterance extracted from the corpus.

4 Automatic normalisation

First experiments aiming at learning the normalisation process were reported in Samardžić et al. (2015) and Samardžić et al. (2016). To this end, the words in the test set were partitioned in four classes, and different normalisation methods were chosen according to the word class:

- *Unique* words are associated with exactly one normalisation in the training set. At test time, these words were normalised using the normalisation seen during training.
- *Ambiguous 1* words are associated with more than one normalisation candidate, but a unique most frequent normalisation can be determined. In this case, the most frequent normalisation was proposed at test time.
- For *Ambiguous 2* words, no single most frequent normalisation can be selected because of tied frequency counts. For this class, it was proposed to select the best normalisation

[1] These reconstructions were generally inspired by the lemmas of the *Idiotikon* dialect dictionary (Staub et al., 1881–).

		Baselines and ceilings				Isolated words		Segments		Constr.
	Prop.	Ident.	Baseline	Combi	Ceiling	1 LM	2 LM	1 LM	2 LM	2 LM
Unique	46.63	22.84	98.79	98.79	98.98	98.30	98.22	97.25	97.64	98.69
Ambig.	42.12	23.52	84.06	84.20	84.64	83.45	82.52	86.27	87.54	87.92
New	11.25	9.46	9.46	35.33	99.57	53.15	53.91	52.50	63.59	65.87
All		21.62	82.54	85.51	93.00	86.96	86.62	87.59	89.56	90.46

Table 1: Percentages of correctly normalised words using the different comparison methods (left), the CSMT system applied to isolated words (center), the CSMT system applied to segments (right), and the segment-level system with constraints (rightmost). The first column shows the proportion of the three word classes in the test corpus. The *All* row refers to the micro-averages over the three word categories.

candidate either by character-level machine translation or by a word-level language model. The latter approach yielded the best results, but the overall impact was limited as this class only accounts for about 0.5% of words.

- *New* words have not been observed in the training set and therefore no normalisation candidates are available. These words were normalised using a character-level machine translation system trained on the training data.

In this contribution, we take advantage of additional annotations that have been made available in the meantime and show that a single normalisation method based on CSMT can outperform the heterogeneous method summarised above.

The current version of the ArchiMob corpus differs from previous versions in three respects. First, the manually annotated normalisations were double-checked to ensure best consistency (Samardžić et al., 2016). Second, the utterances were split into syntactically and prosodically motivated segments of 4-8 seconds. Third, hesitations and false starts were annotated as such and could be excluded from the normalisation task, since their normalisation is not meaningful.

We argue that the CSMT approach can be extended to all four categories of words, without loss of accuracy. To this end, four types of improvements are proposed:

- All CSMT models are tuned using MERT; this has not been done in previous work.

- We propose a setting in which each word is normalised in isolation, and a setting in which entire segments are translated. While normalisation is mainly performed at the token level in related work, we obtain significant improvements by normalising entire segments, thanks to the possibility of capturing parts of the context during the normalisation process.

- We add, beside the training data language model, an additional language model of spoken Standard German to the CSMT systems.

- We make use of the possibility to include translation constraints in the CSMT system. These constraints improve the normalisation of *Unique* words while maintaining the advantage of a single decision process.

Finally, instead of using cross-validation as in Samardžić et al. (2015), given that in these experiments we need development data for tuning, we create a single data split with 80% of utterances used for training, 10% for tuning and 10% for testing. As in earlier work, utterances from all six documents are represented proportionally in each file. The training set contains 8443 segments with 65 671 words, the development set contains 1054 segments with 9032 words, and the test set contains 1055 segments with 8212 words.

5 Experiments

5.1 Baselines and ceilings

The left half of Table 1 shows several measures that indicate the difficulty of the normalisation task. The *Proportion* column shows the distribution of the three word classes established above in the test set (we merge the *Ambiguous 1* and *Ambiguous 2* classes as their distinction is not relevant for the experiments presented here). The *Identical* column indicates for how many words the normalised form is identical to their original form; overall, only

Transcription	Normalisations	Occurrences	
de	der / dann	168	‘the / then’
das	das / dass	168	‘the / that’
es	ein / es	69	‘a / it’
i	ich / in	59	‘I / in’
mer	wir / man / mir	58	‘we / one / me’
s	das / es	44	‘the / it’
bi	bei / bin	22	‘at / am’
sii	sie / sein	22	‘she / be’
e	ein / eine	79	Neut / Fem
en	ein / einen	77	Masc / Neut
cho	kommen / gekommen	3	Pres / PP
chliine	kleiner / kleinen	3	Nom / Acc

Table 2: Normalisation ambiguities observed in the ArchiMob corpus – part-of-speech ambiguities in the upper part, inflectional ambiguities in the lower part. The third column shows the number of occurrences of the transcribed form in the test set. Example sentences and phrases can be found in Table 3 at the end of the paper.

about one fifth of all words are identical to their normalisations.

The *Baseline* normalises every word by assigning it the most frequent normalisation seen in the training set. In case of ties, a normalisation is chosen randomly; for new words, no normalisation process is applied at all. This exactly corresponds to the *Word-by-word* setting in Samardžić et al. (2015). The *Combi* column shows the figures obtained by applying the *Combi* method of Samardžić et al. (2015) and Samardžić et al. (2016) to the revised data set.

In order to estimate the ceiling of the approach normalising each word in isolation, we measure the level of word ambiguity by applying the *Baseline* method trained on the whole dataset (development and test sets included) and tested on the test set. We present the results in the *Ceiling* column.[2] The presented figures show that, as expected, the highest word ambiguity is present among the *Ambiguous* words. For the *New* words, which are generally low-frequency words, ambiguities are rarely observed due to the small size of the sample. The results in Table 1 show that overall 7% of words cannot be normalised regardless of the amount of training data available if contextual information is not taken into account.

Table 2 shows the most frequent ambiguities observed in the corpus. While most ambiguities concern short words of different parts-of-speech, there are also some ambiguities that arise due to the inflectional systems of Swiss German being less rich than the Standard German ones.

5.2 Applying CSMT to isolated words

The goal of this first experiment is to show that a single CSMT model, applied indiscriminately to all three word classes, performs equally well as the *Baseline* or the *Combi* model. To this end, we train a CSMT system on the training set, tune it using MERT on the development set and apply it to the test set.[3] Each word is considered in isolation for training and testing.[4] We have found 7-gram language models to work best, and we have disabled distortion throughout all steps of the process as there is no evidence of such phenomena in our data. Table 1 (*Isolated words*) shows results with

[2]To facilitate the comparison with the other methods, the attribution of the words to the categories *Unique, Ambiguous, New* is still based on the training data only. This means e.g. that 98.98% of words that were observed with a unique normalisation in the training set are effectively unique, whereas for the remaining 1.02% of words, a second normalisation was seen in the test data, making these words ambiguous.

[3]We use the Moses toolkit (Koehn et al., 2007) together with the KenLM language model toolkit (Heafield, 2011) for all experiments. We use the standard settings except for distortion, which is completely disabled, and for the MERT optimisation objective, where we choose WER (word error rate, which *de facto* becomes character error rate in a CSMT setting) instead of BLEU.

[4]For instance, the word *tänkt* is transformed to _ *t ä n k t* _ before feeding it to the translation system. The leading and trailing underscores have proved useful for explicitly modelling word boundaries.

two settings: in the first setting (*1 LM*), we use a single language model estimated on the target side of the training set, whereas in the second setting (*2 LM*) we add a second language model estimated on the Standard German OpenSubtitles2016 corpus (Lison and Tiedemann, 2016), 108 million tokens in size.[5]

The results show that the *1 LM* system achieves only slightly lower performance than the baseline for *Unique* and *Ambiguous* words, but generalises much better than the *Combi* system for *New* words, leading to a higher overall accuracy. A comparison of the two new systems shows that the second language model does not yield any improvements. Our assumption is that there are two reasons for that: (1) the data used for estimating the second language model (Standard German) is quite different to the target data (normalised Swiss German) and (2) word-level systems do not need as much target language data as segment-level systems because there is much more variation between words than inside words.

5.3 Applying CSMT to segments

Normalising each word in isolation means that contextual clues such as the preceding and following word cannot be used for disambiguation. By evaluating our *Ceiling* system we have shown that in this dataset we cannot correctly normalise 7% of words if we translate words in isolation, regardless of the amount of training data available. Therefore, in this second experiment we propose to translate complete segments.[6] By selecting phrases that span word boundaries, the system will be able to perform (at least local) context-dependent disambiguation. The evaluation is still performed word-by-word, as before.[7]

The training, tuning and testing steps in this experiment are the same as in the first one, except that in these experiments 10-gram language models have shown to perform best, not 7-gram language models. This is expected as this system requires as much word context information as possible. Using language models of greater order than 10 did not yield any significant improvements.

The results are shown on the right side of Table 1 (*Segments*). In the *1 LM* setting, the accuracy of *Ambiguous* words improves by 2.82%, as expected. However, contextual influence has a slight negative effect on the *Unique* (-1.05%) and *New* (-0.65%) words. In the *2 LM* setting, the disambiguation of *Ambiguous* words is even more successful (+5.02% compared with the equivalent single-word model). Even more striking is the 9.68% increase for *New* words. Here, the context clearly adds useful information, but only the *2 LM* model is able to take advantage of this information since, by definition, these words do not appear in the original language model.[8]

However, this system still makes proportionally most errors with *New* words. We have found several categories of words to be prone to normalisation errors:

- In 12% of cases, the root is correctly normalised but an erroneous inflectional affix is selected, due to the inflectional ambiguities mentioned in Table 2. Especially for long compound nouns, the context window of 10 characters is not sufficient to disambiguate the candidates: *muurermäischter* 'master mason' is normalised as *maurermeister* where *maurermeistern* would be the correct form, but the relevant case and number information encoded in the preceding determiner is not accessible.

- 9% of errors concern named entities like place or person names: *buechs* is normalised as *buchs* instead of *buochs* (a town name), *riintel* 'Rhine valley' is normalised as *reintal* instead of *rheintal*. These entities are unlikely to occur in the added language model.

- 8% of errors concern abreviations or foreign words, in which the learned normalisation patterns do not apply: *komfiserii* 'confectionary' is normalised as *konfiserei* instead of *confiserie*, *kaazèt* 'concentration camp' is normalised as *kazat* instead of *kz*, *plimut* is normalised as *pleinmut* instead of *plymouth* (a brand name).

[5] We removed punctuation and lowercased the corpus to make it most similar to our normalisation language.

[6] Recall that a segment is about 4–8 seconds long and contains around 8 words on average.

[7] Segments are transformed in the same way as isolated words, using underscores to mark word boundaries. After translation, the segments are split at the underscores in the source for evaluation. This step is not trivial as there may be different numbers of underscores in the source and target due to the differences in word boundaries illustrated in Figure 1.

[8] Similar trends have been observed for Slovene historical texts and user-generated content (Ljubešić et al., 2016), although the improvements are less marked in Slovene because token ambiguity is lower than in our Swiss German data.

- About 2% of mismatches were due to mistakes and typos in the gold normalisations.

5.4 Adding constraints to the segment model

While the segment-level system outperforms the baseline on *Ambiguous* words and has produced significant improvements for *New* words, it still lags behind the baseline by more than 1% regarding *Unique* words. One simple yet effective improvement is to constrain the segment-level system so that the baseline normalisation is chosen for *Unique* words. Moses supports XML annotations to this effect. We used the segment-level *2 LM* system as a basis, retuned it with the annotated development data and tested it on the annotated test data. We have found the *exclusive* strategy to work slightly better than the *constraint* strategy.[9]

The results of this hybrid system are shown in the rightmost column of Table 1 (for reasons of space, we only show results for the *2 LM* system). For the *Unique* words, the accuracy is now very close to the baseline. The remaining errors concern three very long words that are not normalised at all despite the presence of a baseline normalisation; we suspect this to be a bug in Moses.

While it is not surprising that the constrained system outperforms the segment-level system for *Unique* words, it is striking that the accuracies also rise for the *Ambiguous* and *New* words. The contextual information provided by the *Unique* word annotation also positively impacts the adjacent non-unique words. Overall, the constrained system outperforms the pure segment-level system by 0.9%. We assume that, if enough target language data was present in the system (and not the near-target data of Standard German), these constraints would not be necessary.

Given that this is the smallest recorded difference among all the comparisons throughout the paper, we ran three MERT tuning processes on both systems and calculated on each output the approximate randomisation statistical test (Yeh, 2000) with 1000 iterations to measure the probability of observing the difference by chance. The highest p-value measured on any of the three outputs was $p < 0.001$ showing that the observed difference of $\sim$1% on our test set is already highly significant.

[9] See Section 4.8.2 of the Moses manual, consulted at `http://www.statmt.org/moses/manual/manual.pdf` on 2016-06-06. We have also found that adding the word boundary symbol to the baseline normalisations is useful to prevent spurious suffixes from being appended.

6 Conclusion

In this paper, we have shown that character-level machine translation can be used successfully to learn the process of automatically normalising dialect texts with heterogeneous transcriptions. Translation systems operating on isolated words obtain accuracy levels comparable with previous work for *Unique* and *Ambiguous* words, whereas significant improvements are observed for *New* words. Systems operating on entire segments yield accuracy gains for *Ambiguous* words and, when combined with an additional language model, for *New* words. Constraining the translation of *Unique* words allows to further improve overall accuracy by nearly 1%.

However, there is still room for improvement. In particular, several extensions may be envisaged to improve the treatment of ambiguous words and long range dependencies:

- A language model that operates on the word level (instead of the character level) would allow us to keep track of larger context windows. Such an additional language model could be integrated into the translation process using Moses feature functions.

- Adding part-of-speech tags as an additional word-level feature may also be useful to disambiguate words. Samardžić et al. (2016) have showed that respectable tagging performance can be achieved without using the normalised forms, but it is open whether such a tagger is able to reliably resolve the ambiguities mentioned in Table 2.

- Neural language modeling could learn the morphosyntactic regularities and long distance dependencies of the language much better than the surface n-gram language model currently used.

- Increasing the language model order and/or the maximum phrase length could alleviate the difficulties observed when the normalised form is much longer than the original form. For example, *nüm* 'no more' should be normalised as *nicht mehr*, but incomplete and wrong forms such as *nicht m* or *nicht man* are produced instead by the current models.

Finally, it is expected that normalising additional unseen texts will yield a lot of *New* words that are

	Transcription	Normalisation	
de	jaa *de* het me no gluegt tänkt dasch ez *de* genneraal	ja *dann* hat man noch gelugt gedacht das ist jetzt *der* general	‘yes then one still watched and thought, now this is the General’
das	ich wäiss aber nüme wele leerer *das* *das* gsii isch	ich weiss aber nicht mehr welcher lehrer *dass das* gewesen ist	‘but I don’t know any more which teacher it was’
es	*es* huus *es* isch	*ein* haus *es* ist	‘a house’ ‘it is’
mer	*mer* händ *mer* hät er hät *mer* tanket	*wir* haben *man* hat er hat *mir* gedankt	‘we have’ ‘one has’ ‘he thanked me’
s	*s* huus *s* isch	*das* haus *es* ist	‘the house’ ‘it is’
bi	*bi* frauefeld ich *bi* nöd dehäim gsii	*bei* frauenfeld ich *bin* nicht daheim gewesen	‘near Frauenfeld’ ‘I was not at home’
sii	wüssed *sii* ich han wele schwiizeri *sii*	wissen *sie* ich habe wollen schweizerin *sein*	‘you know’ ‘I wanted to be Swiss’
e	*e* kino *e welotuur*	*ein* kino *eine* velotour	‘a cinema’ ‘a bike tour’
en	si hät *en* gaarte ghaa und dän isch *en* bueb ufgschtande	sie hat *einen* garten gehabt und dann ist *ein* bub aufgestanden	‘she had a garden’ ‘and then a boy stood up’
cho	de händs müse hääi *cho* si isch uf d wält *cho*	dann haben sie müssen heim *kommen* sie ist auf die welt *gekommen*	‘then they had to come home’ ‘she was born’
chliine	er isch en *chliine* gsii im *chliine* chileli	er ist ein *kleiner* gewesen im *kleinen* kirchlein	‘he was a small one’ ‘in the small church’

Table 3: Examples of the normalisation ambiguities shown in Table 2.

named entities. To address this issue, we investigate the inclusion of a lexicon containing toponyms and patronyms of German-speaking Switzerland.

Acknowledgements

Construction and distribution of the ArchiMob corpus was supported by the University of Zurich URPP Language and Space. In particular, we would like to thank Tanja Samardžić, Noëmi Aepli and Fatima Stadler for making the corpus available with the improvements mentioned in Section 4.

The research leading to these results has received funding from the Swiss National Science Foundation grant no. IZ74Z0_160501 (ReLDI).

References

Alistair Baron 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.

Marcel Bollmann. 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)*, pages 3–14, Lisbon, Portugal.

Claudia Bucheli and Elvira Glaser. 2002. The syntactic atlas of Swiss German dialects: Empirical and methodological problems. In Sjef Barbiers, Leoni Cornips, and Susanne van der Kleij, editors, *Syntactic Microvariation*, volume 2, pages 41–73, Amsterdam. Meertens Institute Electronic Publications in Linguistics.

Orphée De Clercq, Bart Desmet, Sarah Schulz, Els Lefever, and Véronique Hoste. 2013. Normalization of Dutch user-generated content. In *Proceedings of RANLP 2013*, pages 179–188, Hissar, Bulgaria.

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

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh.

Nora Hollenstein 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, Ireland. Association for Computational Linguistics.

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

Philipp Koehn, 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*, pages 177–180, Prague, Czech Republic.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of KONVENS 2016*.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2014. Standardizing tweets with character-level machine translation. In *Proceedings of CICLing 2014*, Lecture notes in computer science, pages 164–175, Kathmandu, Nepal. Springer.

Eva Pettersson, Beáta B. Megyesi, and Joakim Nivre. 2013. 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)*, pages 163–79, Oslo, Norway.

Eva Pettersson, 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)*, pages 32–41, Gothenburg, Sweden.

Ulrich Reffle. 2011. Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering*, 17:265–82.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of The 4th Biennial Workshop on Less-Resourced Languages, Seventh Language and Technology Conference*.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob – a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Yves Scherrer and Tomaž Erjavec. 2015. Modernising historical Slovene words. *Natural Language Engineering*, pages 1–25. Available on Cambridge Journals Online.

Beat Siebenhaar. 2005. Die dialektale Verankerung regionaler Chats in der deutschsprachigen Schweiz. In Eckhard Eggers, Jürgen Erich Schmidt, and Dieter Stellmacher, editors, *Moderne Dialekte – Neue Dialektologie*, pages 691 – 717. Steiner, Stuttgart.

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

Friedrich Staub, Ludwig Tobler, Albert Bachmann, Otto Gröger, Hans Wanner, and Peter Dalcher, editors. 1881–. *Schweizerisches Idiotikon: Wörterbuch der schweizerdeutschen Sprache*. Huber, Frauenfeld.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of EAMT 2009*, pages 12–19, Barcelona, Spain.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING 2000*, pages 947–953.