Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German

Pierre-Edouard Honnet[*,1], **Andrei Popescu-Belis**[*,2], **Claudiu Musat**[3], **Michael Baeriswyl**[3]

[1] Telepathy Labs	[2] HEIG-VD / HES-SO	[3] Swisscom (Schweiz) AG
Schützengasse 25	Route de Cheseaux 1, CP 521	Genfergasse 14
CH-8001 Zürich	CH-1401 Yverdon-les-Bains	CH-3011 Bern
Switzerland	Switzerland	Switzerland
pierre-edouard.honnet	andrei.popescu-belis	claudiu.musat@swisscom.com
@telepathy.ai	@heig-vd.ch	michael.baeriswyl@swisscom.com

Abstract

The goal of this work is to design a machine translation (MT) system for a low-resource family of dialects, collectively known as Swiss German, which are widely spoken in Switzerland but seldom written. We collected a significant number of parallel written resources to start with, up to a total of about 60k words. Moreover, we identified several other promising data sources for Swiss German. Then, we designed and compared three strategies for normalizing Swiss German input in order to address the regional diversity. We found that character-based neural MT was the best solution for text normalization. In combination with phrase-based statistical MT, our solution reached 36% BLEU score when translating from the Bernese dialect. This value, however, decreases as the testing data becomes more remote from the training one, geographically and topically. These resources and normalization techniques are a first step towards full MT of Swiss German dialects.

Keywords: machine translation, low-resource languages, spoken dialects, Swiss German, character-based neural MT

1. Introduction

In the era of social media, more and more people make online contributions in their own language. The diversity of these languages is however a barrier to information access or aggregation across languages. Machine translation (MT) can now overcome this limitation with considerable success for well-resourced languages, i.e. language pairs which are endowed with large enough parallel corpora to enable the training of neural or statistical MT systems. This is not the case, though, for many low-resourced languages which have been traditionally considered as oral rather than written means of communication, and which often lack standardized spelling and/or exhibit significant variations across dialects. Such languages have an increasing presence in written communication, especially through social media, while remaining inaccessible to non-speakers.

This paper presents a written MT system for a mostly spoken family of dialects: Swiss German. Although spoken in a technologically developed country by around five million native speakers, Swiss German has never been significantly used in writing – with the exception of folklore or children books – before the advent of social media. Rather, from primary school, speakers of Swiss German are taught to use High German in writing, more precisely a variety known to linguists as Swiss Standard German, which is one of the three official federal languages along with French and Italian. Swiss German is widely used in social media, but foreigners or even Swiss speakers of the other official languages cannot understand it.

In this paper, we describe the first end-to-end MT system from Swiss German to High German. In Section 2, we present the Swiss German dialects and review the scarce monolingual and even scarcer parallel language resources that can be used for training MT. In Section 3, we review previous work on Swiss German and on MT of low-resource languages. In Section 4, we address the major issue of dialectal variation and lack of standard spelling – which affects many other regional and/or spoken languages as well – through three solutions: explicit conversion rules, phonetic representations, and character-based neural MT. These solutions are combined with phrase-based statistical MT to provide a standalone translation system, as explained in Section 5. In Section 6 we present evaluation results. We first find that the similarity between the regions of training vs. test data has a stronger effect on performance than the similarity of text genre. Moreover, the results show that character-based NMT is beneficial for dealing with spelling variation. Our system is thus an initial general purpose MT system making Swiss German accessible to non-speakers, and can serve as a benchmark for future, better-resourced attempts.

2. Collecting Swiss German Resources

2.1. A Heterogeneous Family of Dialects

Definition. Swiss German (Russ, 1990; Christen et al., 2013) is a family of dialects used mainly for spoken communication by about two thirds of the population of Switzerland (i.e. over five million speakers). Swiss German is typically learned at home as a first language, but is substituted starting from primary school by High German for all written forms, as well as for official spoken discourse, for instance in politics or the media. Linguistically, the variety of High German written and spoken in Switzerland is referred to as *Swiss Standard German* (see Russ (1994),

---

* Work conducted while the first and second authors were at the Idiap Research Institute, Martigny, Switzerland.

Chapter 4, p. 76–99) and is almost entirely intelligible to German or Austrian speakers. On the contrary, Swiss German is generally not intelligible outside Switzerland.

In fact, Swiss German constitutes a group of heterogeneous dialects, which exhibit strong local variations. Due to their spoken nature, they have no standardized written form: for instance, the word *kleine* (meaning small' in Standard German) could be written as *chlyni*, *chliini*, *chline*, *chli* or *chlii* in Swiss German. Linguistic studies of the Swiss German dialects (see Russ (1990) or Christen et al. (2013)) generally focus on the phonetic, lexical or syntactic variations and their geographical distribution, often concluding that such variations are continuous and non-correlated with each other. Finally, little teaching material in Swiss German is available to foreigners.

Divisions. The areas where each dialect is spoken are influenced both by the administrative divisions (cantons and communes) and by natural borders (topography). Within the large group of Germanic languages, the dialects of Switzerland belong to the *Alemannic* group. However, while a majority of dialects are *High Alemannic* (yellow area on map in Figure 1), those spoken in the city of Basel and in the Canton of Valais belong respectively to the *Low Alemannic* and the *Highest Alemannic* groups. Within the High Alemannic group, a multitude of divisions have been proposed. One of the most consistent ones is the Brunig-Napf-Reuss line between the eastern and western groups (red line in Fig. 1). A fine-grained approach could easily identify one or more dialects for each canton.

For the purpose of this study, we distinguish only two additional sub-groups on each side of the Brunig-Napf-Reuss line, and refer to them using the largest canton in which they are spoken. Westwards, we distinguish the Bernese group from the group spoken around Basel (cantons of Basel-Country, Solothurn and parts of Aargau). Eastwards, we distinguish the Zürich group from the easternmost group around St. Gallen. Therefore, for training and testing MT on various dialects, we consider in what follows six main variants of Swiss German, represented on the map in Figure 1.

Notations. We refer to Swiss German as 'GSW' (abbreviation from ISO 639-2) followed by the indication of the variant: GSW-BS (city of Basel), GSW-BL (regions of Basel, Solothurn, parts of Aargau), GSW-BE (mainly canton of Bern), GSW-ZH (canton of Zurich and neighbors), GSW-SG (St. Gallen and easternmost part of Switzerland), GSW-VS (the German-speaking part of the canton of Valais/Wallis). These groups correspond to the dialect labels used in the Alemannic Wikipedia (see Section 2.2 below): *Basel*, *Baselbieter*, *Bern*, *Zurich*, *Ùndertòggeborg*, and *Wallis* (Valais). In contrast, Swiss Standard German is referred to as 'DE-CH', a qualified abbreviation from IETF. Moreover, below, we will append the genre of the training data to the dialect abbreviation.

Usage and Need for MT. Swiss German is primarily used for spoken communication, but the widespread adoption of social media in Switzerland has significantly increased its written use for informal exchanges on social platforms or in text messages. No standardized spelling has
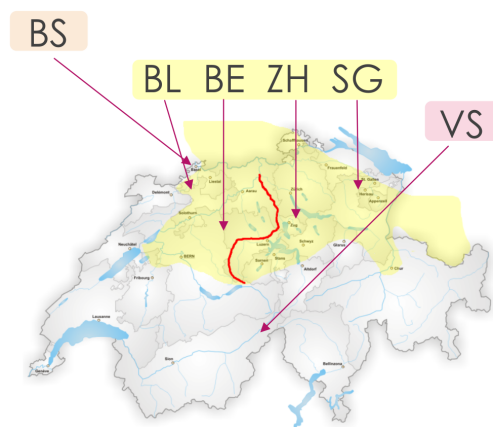

Figure 1: Map of Switzerland, with six main dialects that we identify for our study. The area in yellow indicates the High Alemannic dialects. Image source: https://commons.wikimedia.org/wiki/File:Brunig-Napf-Reuss-Linie.png.

emerged yet, a fact related to the lack of GSW teaching as a second language. GSW is still written partly with reference to High German and partly using a phonetic transcription, also inspired from German pronunciation. Access to such content in social media is nearly impossible to foreigners, and even to speakers of different dialects, e.g. Valaisan content to Bernese speakers. Our goal is to design an MT system translating all varieties of GSW (with their currently observed spelling) towards High German, taking advantage of the relative similarity of these languages. By pivoting through High German, other target languages can then be supported. Moreover, if a speech-to-text system existed for Swiss German (Garner et al., 2014), our system would also enable spoken translation.

2.2. Parallel Resources

Despite attempts to use comparable corpora or even monolingual data only (reviewed in Section 3), parallel corpora aligned at the sentence level are essential resources for training statistical MT systems. In our case, while written resources in Swiss German are to some extent available (as reviewed in Section 2.3), it is rare to find their translations into High German or vice-versa. When these are available, the two documents are often not available in electronic version, which requires a time-consuming digitization effort to make them usable for MT.[1]

One of our goals is to collect the largest possible set of parallel GSW/DE texts, in a first stage regardless of their licensing status. We include among such resources parallel lexicons ("dictionaries"), and show that they are helpful for training MT. We summarize in Table 1 the results of our data collection effort, providing brief descriptions of each resource with especially their variant of GSW and their domain. We describe in detail each resource hereafter.

---

[1]Many of them are children books, such as *Pitschi* by Hans Fischer, *The Gruffalo* by Julia Donaldson, or *The Little Prince* by Antoine de Saint-Exupéry. Other examples include transcripts of Mani Matter's songs, or several Asterix comics in Bernese.

Dataset	Train	Dev.	Test	Total
GSW-BE-Novel	2,667	218	183	3,251
GSW-BE-Wikipedia	–	180	67	247
GSW-VS-Radio	463	100	50	613
GSW-ZH-Wikipedia	–	45	50	95
GSW-BE-Bible	–	–	126	126
GSW-Archimob	40,159	2,710	2,710	45,579
GSW-ZH-Lexicon1	1,527	–	–	1,527
GSW-BE-Lexicon2	1,224	–	–	1,224

Table 1: GSW/DE parallel datasets partitioned for MT training, tuning and testing, with sizes in numbers of parallel sentences. The lexicons (last two lines) were not used for testing, and 183 additional lines from GSW_BE_Novel are kept apart for future testing.

GSW-BE-Novel. Translations of books from DE into GSW are non-existent. We thus searched for books written originally in GSW and then translated into DE. Among the growing body of literature published in Swiss German, we found only one volume translated into High German and available in electronic form: *Der Goalie bin ig* (in English: *I am the Keeper*), written in Bernese by Pedro Lenz in 2010. The DE translation stays close to the original GSW-BE text, therefore sentence-level alignment was straightforward, resulting in 3,251 pairs of sentences with 37,240 words in GSW-BE and 37,725 words in DE.

GSW-BE-Wikipedia and **GSW-ZH-Wikipedia**. The Alemannic version of Wikipedia[2] appeared initially as a promising source of data. However, its articles are written not only in Swiss German, but also in other Alemannic dialects such as Alsatian, Badisch and Swabian. As its contributors are encouraged to write in their own dialects, only a few articles are homogeneous and have an explicit indication of their dialect, using an Infobox with one of the six labels indicated above. Among them, even fewer have an explicit statement indicating that they have been translated from High German (which would make the useful as parallel texts). We identified two such pages and sentence-aligned them to serve as test data: *"Hans Martin Sutermeister"* translated from DE into GSW-BE and *"Wädenswil"* from DE into GSW-ZH.[3]

GSW-VS-Radio. A small corpus of Valaisan Swiss German (also called *Wallisertiitsch*) has been collected at the Idiap Research Institute (Garner et al., 2014).[4] The corpus consists of transcriptions of a local radio broadcast[5] translated into High German.

GSW-BE-Bible. The Bible has been translated in several GSW dialects, but the only electronic version available to us were online excerpts in Bernese.[6] However, this is not translated from High German but from a Greek text, hence the alignment with any of the German Bibles is problematic.[7] We selected the contemporary *Gute Nachricht Bibel* (1997) for its modern vocabulary, and generated parallel data from four excerpts of the Old and New Testament, while acknowledging their particular style and vocabulary. The following excerpts were aligned: *Üse Vatter*, *D Wienachtsgschicht*, *Der barmhärzig Samaritaner* and *D Wält wird erschaffe*.

GSW-Archimob. Archimob is a corpus of standardized Swiss German (Samardžić et al., 2016), consisting of transcriptions of interviewees speaking Swiss German, with a word-align normalized version in High German.[8] The interviews record memories of WW II, and all areas of Switzerland are represented. In most cases, the normalization provides the corresponding High German word or group of words, but in other cases it is Swiss German with a standardized orthography devised by the annotators. Using a vocabulary of High German, we filtered out all sentences whose normalizations included words outside this vocabulary. In other words, we kept only truly High German sentences, along with their original Swiss German counterparts, resulting in about 45,000 GSW/DE word-aligned sentence pairs.

GSW-ZH-Lexicon and **GSW-BE-Lexicon**. The last two parallel resources are vocabularies, i.e. lists of GSW words with their DE translation. As such, they are useful for training our research systems, but not for testing them. The first one is based on *Hoi Zäme*, a manual of Zürich Swiss German intended for High German speakers. The data was obtained by scanning the printed version, performing OCR[9] and manually aligning the result. Although the book contains also parallel sentences, only the bilingual dictionary was used in our study, resulting in 1,527 words with their translations. A similar dictionary for Bernese (GSW-BE vs. DE) was found online[10] with 1,224 words for which we checked and corrected the alignments.

2.3. Monolingual Resources

The Phonolex dictionary, a phonetic dictionary of High German,[11] was used for training our grapheme-to-phoneme converter (see Section 4.2). It contains High German words with their phonetic transcriptions.

---

[2] http://als.wikipedia.org

[3] These pages are respectively available at https://de.wikipedia.org/wiki/Hans_Martin_Sutermeister (High German), https://als.wikipedia.org/wiki/Hans_Martin_Sutermeister (Bernese), https://de.wikipedia.org/wiki/W%E4denswil (High German), and https://als.wikipedia.org/wiki/W%E4denswil (Zurich Swiss German).

[4] www.idiap.ch/dataset/walliserdeutsch

[5] Radio Rottu, http://www.rro.ch.

[6] www.edimuster.ch/baernduetsch/bibel.htm

[7] www.die-bibel.de/bibeln/online-bibeln/

[8] http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html

[9] Tesseract: https://github.com/tesseract-ocr/

[10] www.edimuster.ch/baernduetsch/woerterbuechli.htm

[11] www.bas.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html. We also use it to find OOV words.

About 75 pages from the Alemannic Wikipedia mentioned above have been collected and used to derive orthographic normalization rules in Section 4.1. To build language models (see Section 5) we used the News Crawls 2007–2015 from the Workshop on MT.[12]

3. Previous Work on Swiss German and the MT of Low-Resource Languages

The variability of Swiss German dialects has been investigated in a number of studies, such as those by Russ (1990), Scherer (2012a), and Christen et al. (2013). This variability was illustrated in a system for generating Swiss German text, with fine-grained parameters for each region on a map (Scherrer, 2012b).

Language resources for Swiss German are extremely rare. The ArchiMob corpus (Samardžić et al., 2016) is quite unique, as it provides transcripts of spoken GSW narratives, along with their normalization, as presented above (Samardžić et al., 2015). First performed manually – thus generating ground-truth data – the normalization was then performed automatically using character-based statistical MT (Scherrer and Ljubešić, 2016).

Initial attempts for MT of GSW include the above-mentioned system for generating GSW texts from DE (Scherrer, 2012a), and a system combining ASR and MT of Swiss German from Valais (Garner et al., 2014). A normalization attempt for MT, on a different Germanic dialect, has been proposed for Viennese (Hildenbrandt et al., 2013). The MT of low-resource languages or dialects has been studied on many other important cases, in particular for Arabic dialects which are also predominantly used for spoken communication (Zbib et al., 2012). The lack of a normalized spelling of dialects has for instance an impact on training and evaluation of automatic speech recognition: a solution is to address spelling variation by mining text from social networks (Ali et al., 2017). Other strategies are the crowdsourcing of additional parallel data, or the use of large monolingual and comparable corpora to perform bilingual lexicon induction before training an MT system (Klementiev et al., 2012; Irvine and Callison-Burch, 2013; Irvine and Callison-Burch, 2016). The METIS-II EU project replaced the need for parallel corpora by using linguistic pre-processing and statistics from target-language corpora only (Carl et al., 2008). In a recent study applied to Afrikaans-to-Dutch translation, the authors use a character-based "cipher model" and a word-based language model to design a decoder for the low-resourced input language (Pourdamghani and Knight, 2017).

The Workshops on Statistical MT have proposed translation tasks for low-resourced languages to/from English, such as Hindi in 2014 (Bojar et al., 2014), Finnish in 2015, or Latvian in 2017. However, these languages are clearly not as low-resourced as Swiss German, and possess at least a normalized version with a unified spelling. In 2011, the featured translation task aimed at translating text messages from Haitian Creole into English, with a parallel corpus of similar size as ours (ca. 35k words on each side, plus a Bible

translation). The original system built in the wake of the 2010 Haiti earthquake leveraged a phonetic mapping from French to Haitian Creole to obtain a large bilingual lexicon (Lewis, 2010; Lewis et al., 2011).

4. Normalizing Swiss German for MT

Three issues must be addressed when translating Swiss German into High German, which all contribute to a large number of out-of-vocabulary (OOV) words (i.e. previously unseen during training) in the source language:

1. The scarcity of parallel GSW/DE data for training (see Section 2.2), which cannot be easily addressed by the strategies seen in Section 3.
2. The variability of dialects across training and testing data, which increases dialect-specific scarcity.
3. The lack of a standard spelling, which introduces intra-dialect and intra-speaker variability.

There are several ways to address these issues. The most principled one is the normalization of all GSW input using unified spelling conventions, coupled with the design of a GSW/DE MT system for normalized input. However, such a goal is far too ambitious for our scope. Instead, we propose here to normalize Swiss German input for the concrete perspective of MT by converting unknown GSW words either to known GSW ones or to High German ones, which are preserved by the GSW/DE MT system and increase the number of correctly translated words.[13]

This procedure, summarized below, rests on the assumption that many OOV GSW words are close to DE words, but with a slightly different pronunciation and spelling (see examples in the third column of Table 2). Each of the three strategies follow the same procedure:

1. For each OOV word w, apply the normalization strategy. If it changes w into w' then go to (2), if not to (4).
2. If w' is a known GSW word then replace w with w' and proceed to (4), if not, go to (3).
3. If w' is a known DE word then replace w with w'. If not, leave w unchanged and go to (4).
4. Translate the resulting GSW text.

This normalization method has two possible chances to help MT, by converting OOV words either into a known GSW word, or into a correct DE word which is no longer processed by MT. We describe below three strategies to normalize GSW text input before GSW/DE MT.

4.1. Explicit Spelling Conversion Rules

The first strategy is based on explicit conversion rules for every OOV word w, which is changed into w' by applying in sequence several spelling conversion rules, keeping the result if it is a GSW or a DE word, as explained above. The orthographic rules implemented in our system are shown in Table 2, with possible conversion examples.

4.2. Using Phonetic Representations

The second approach is based on the assumption that despite spelling differences, variants of the same word will

---

[12]http://www.statmt.org/wmt17/translation-task.html

[13]The MT system is specifically built so that OOV words are copied in the target sentence, rather than deleted.

Spelling	Convert to	Example
.*scht.*	.*st.*	Angscht → Angst
.*schp.*	.*sp.*	Schprache → Sprache
^gäge.*	^gegen.*	Gägesatz → Gegensatz
CäC	CeC	Präsident → President
^gm.*	^gem	Gmeinde → Gemeinde
^gf.*	^gef	gfunde → gefunde(n)
^gw.*	^gew	gwählt → gewählt
^aa.*	^an.*	Aafang → Anfang
.*ig$	.*ung$	Regierig → Regierung
^ii.*	^ein.*	Iiwohner → Einwohner

Table 2: Orthographic conversion rules using meta-characters ^ and $ for the beginning and end of a word, .* for any sequence of characters, and C for any consonant.

have the same pronunciation. Thus, converting an out-of-vocabulary (OOV) word to its phonetic transcription may allow finding the equivalent word which is present in the vocabulary. In this case, substituting the OOV word with a known word with the same pronunciation should help MT, assuming the same meaning.

For this, a grapheme-to-phoneme (G2P) converter is needed. It consists of an algorithm which is able to convert sequences of characters into phonetic sequences, or go from the written form of a word to its pronunciation. The idea is to build it on High German, as we expect Swiss German to be written in a phonetic way, which means that the G2P conversion should be close to High German pronunciation rules. In our experiments, a G2P converter was trained on the Phonolex dictionary, which contains High German words with their phonetic transcriptions. A GSW phonetic dictionary was created by using this system. To translate a new OOV word, we convert the word to its pronunciation, and check whether the resulting pronunciation exists either in the phonetic GSW dictionary or in the phonetic DE dictionary, following the procedure explained at the beginning of this section.

4.3. Character-based Neural MT

Mainstream neural MT systems are typically trained using recurrent neural networks (RNNs) to encode a source sentence, and then decode its representation into the target language (Cho et al., 2014). The RNNs are often augmented with an attention mechanism to the source sentence (Bahdanau et al., 2014). However, training an NMT is not feasible for GSW/DE, as the size of our resources is several orders of magnitude below NMT requirements. However, several recent approaches have explored a new strategy: the translation system is trained at the character level (Ling et al., 2015; Costa-jussà and Fonollosa, 2016; Chung et al., 2016; Bradbury et al., 2017; Lee et al., 2017), or at least character-level techniques such as byte-pair encoding are used to translate OOV words (Sennrich et al., 2016).

As the available data is limited, one possible approach is to combine a PBSMT and a CBNMT system: the former translates known words, while the latter translates OOV ones. The CBNMT system has two main advantages: it can translate unseen words based on the spelling regularities observed in the training data, and it can be trained with

smaller amounts of data compared to the requirements of standard NMT methods.

Among the CBNMT approaches, we use here Quasi Recurrent Neural Networks (QRNNs) (Bradbury et al., 2017), which take advantage of both convolutional and recurrent layers. The increased parallelism introduced by the use of convolutional layers allows to speed up both training and testing of translation models. There are two advantages to use CBNMT for OOV translation only. First, the training data may be sufficient to capture spelling conversion better than hand-crafted rules such as those in Table 2. Second, we can use smaller recurrent layers, as the character sequences to translate for OOV words are much shorter than sentences.

We built a CBNMT system for OOV words based on open source scripts for TensorFlow available online, using the implementation of the QRNN architecture proposed by Kyubyong Park,[14] with the following modifications:

1. We added a "start of word" symbol to avoid mistakes on the first letter of the word. This was done outside the translation scripts, by adding the ':' symbol to each word before the first letter, and removing it after translation.

2. We modified the QRNN translation script to allow the translation of input texts without scoring the translation (for the production mode, when no reference is available).

3. We added the possibility to translate an incomplete minibatch, by padding the last incomplete batch with empty symbols (0).[15]

4. We set the following hyper-parameters: the maximum number of characters is 40, as no longer words were found in our GSW vocabulary. The minibatch size was kept to 16, and the number of hidden units was kept to 320, as in the default implementation.

We trained the CBNMT model using unique word pairs from the Archimob corpus (see 2.2 above), i.e. a Swiss German word and its normalized version, with a training set of 40,789 word pairs and a development set of 2,780 word pairs.

5. Integration with Machine Translation

We use phrase-based statistical MT for the core of our system, as data was not sufficient to train neural MT. We experimented indeed with two NMT systems[16], which are typically trained on at least one million sentences, and tuned on 100k. In our case, the available data did not allow NMT to outperform PBSMT, which is used below.

Using the Moses toolkit[17] to build a PBSMT system (Koehn et al., 2003), we used various subsets of the parallel GSW/DE data presented in Section 2.2 above to learn

---

[14] https://github.com/Kyubyong/quasi-rnn

[15] Originally, if the size of the minibatch is n, and the number of sentences modulo n is y (i.e. there are $n * x + y$ sentences), then only the $n * x$ first sentences were translated by the system, which ignored the y last ones.

[16] The DL4MT toolkit https://github.com/nyu-dl/dl4mt-tutorial and the OpenNMT-py one https://github.com/OpenNMT/OpenNMT-py.

[17] http://www.statmt.org/moses/

translation models. As for the target language model, we trained a tri-gram model using IRSTLM[18] over ca. 1.3 billion words in High German, and tuned the system using the development data indicated above in Section 2.3. As explained above, the normalization strategies are used to attempt to change GSW OOV words into either GSW or even DE words that are in the vocabulary. As we will see, we have combined two strategies in several experiments. We will use the most common metric for automatic MT evaluation, i.e. the BLEU score (Papineni et al., 2002).

6. Results and Discussion

6.1. Effects of Genre and Dialect

The first system to translate from Swiss German into High German was built using Moses trained on the Bernese novel corpus (GSW-BE-Novel in Table 1 above), with character-based NMT for OOV words. Table 3 shows the BLEU scores obtained when testing this system on test sets from different regions or topics. Moreover, we also vary the tuning sets, including ones closer to the target test domains to assess their impact. The best BLEU score is around 35% which can be compared, for instance, with Google's NMT scores of 41% for EN/FR and 26% for EN/DE, trained on tens of millions of sentences and nearly one hundred processors (Wu et al., 2016). In our case, our modest resources enable us to reach quite a high score thanks to the normalization strategy and the relative similarity of GSW to DE.

Test set	Tuning (dev) set	BLEU
GSW-BE-Novel	GSW-BE-Novel	35.3
GSW-BE-Wikipedia	GSW-BE-Novel	21.9
same	GSW-BE-Wikipedia	21.7
GSW-ZH-Wikipedia	GSW-BE-Novel	16.2
same	GSW-ZH-Wikipedia	15.3
GSW-VS-Radio	GSW-BE-Novel	9.7

Table 3: BLEU scores for various tuning and test sets for the baseline system trained on GSW-BE-Novel. Performance decreases significantly as the dialect and domain are more remote from the training/tuning data.

A typical output of this system is:

GSW-BE source: *fasch wiwenernolänger hätt wöueküsse . oder hanisächtnumegmeint?*

DE MT: *fast wie wenn er noch länger hätte wollten küsse . oder hab ich es wohl nur gemeint ?*

Human DE reference: *als hätte er noch länger küssen wollen . oder etwa nicht ?*

The scores in Table 3 show the following trends:

1. When testing on similar data, i.e. the same dialect and same domain, the scores are the highest, and in the same range as state of the art EN-DE or EN-FR systems.
2. When changing domain (testing on Wikipedia data in the same dialect), the scores are decreasing.
3. When testing on different dialects, the scores decrease more. This is true both for GSW-ZH and GSW-VS. As the dialect and domain are further from the data

used to train the system, the score gets lower. GSW-VS is known to be very different from any other GSW dialect, and radio broadcast data is expected to be very different from the novel used at training time.

6.2. Effect of the Size of Training Data and Language Model

To evaluate first the effect of using more training data, with larger vocabularies, a new system was trained using the same data as in the previous experiments, complemented with the two bilingual lexicons presented in Section 2.2. Table 4, second column, presents the resulting BLEU scores, which increase in all cases by about 1 BLEU point. As expected, using more training data in the form of bilingual lexicons yields more reliable translation models. To build more robust systems, we also used a larger target language model built on the NewsCrawl 2007-2015 data from WMT (see Section 2.3) instead of only the DE side of our parallel data, which is still used for training the translation models, as above. Table 4, third column, gives the BLEU scores on the same test sets, using the larger target language model. We observe that the scores decrease slightly for the Bernese test sets, and hypothesize that this is due to the different domains of the language model and the test set. However, as the larger language model is trained on more diverse data, we will keep using it below for its robustness.

	BLEU	
Test set	Small LM	Large LM
GSW-BE-Novel	36.2	34.1
GSW-BE-Wikipedia	23.6	22.7
GSW-ZH-Wikipedia	17.3	17.7
GSW-VS-Radio	10.0	10.7

Table 4: BLEU scores for various test sets (Bern, Zurich, Valais dialects) for a Moses-based system trained over data *including the two GSW/DE dictionaries* with two language models (LM).

6.3. Out-of-Vocabulary GSW Words

The three approaches proposed for normalization were evaluated on the same datasets as the previous systems. Additionally, two other approaches combining, on one side orthographic and phonetic based conversions, and on the other side CBNMT and phonetic conversion, were evaluated. Table 5 summarizes the results for the baseline system and the proposed approaches.

Baseline1 corresponds to the system with a language model trained only on the parallel GSW-DE data, while Baseline2 is using a larger language model, described in Sec. 6.2. We can make the following observations:

- In all the cases except GSW-BE-Novel, the orthographic approach improves the BLEU score of the baseline system, and the improvement is bigger for more remote dialects and domains.
- The phonetic approach improves the score in 4 out of 6 cases. In the remaining cases, we suppose that some words did not require pre-processing, and that the pre-processing may have converted the word to a false pos-

Test set	Baseline1	Baseline2	Phon.	Orth.	Orth. & Phon.	CBNMT & Phon.	CBNMT
GSW-Archimob	10.9	10.8	10.8	11.2	13.9	27.9	**32.9**
GSW-BE-Novel	**36.2**	34.1	34.3	34.4	34.4	35.6	35.4
GSW-BE-Wikipedia	23.6	22.7	23.2	23.6	23.7	20.5	**24.0**
GSW-ZH-Wikipedia	17.3	17.7	17.1	18.9	18.2	22.0	**22.1**
GSW-VS-Radio	10.0	10.7	11.0	12.2	12.0	8.7	**22.9**
GSW-BE-Bible	5.7	5.8	6.2	6.1	**6.4**	6.3	6.3

Table 5: BLEU scores for several test sets and normalization strategies (orthographic, phonetic, character-based NMT).

itive (i.e. the algorithm found a matching word, but it was not the correct one for translation).

- Combining both approaches always results in better scores than the baseline, but in the case for which the phonetic approach score deteriorated, orthographic conversion only performs better.

- In all the cases, combining CBNMT with the baseline PBSMT works the best. The highest improvement is brought when dialect or domain are different (except for the Bible), because more data was used to train the CBNMT models. This is especially true for the GSW-Archimob test set, which has similar data as the one used to train the CBNMT models.

- Baseline1 performs better than all the systems for GSW-BE-Novel test set. This is expected as the training data is both from the same dialect and the same domain. Additionally, the language model is trained on this same data.

7. Conclusion

In this paper, we proposed solutions for the machine translation of a family of dialects, Swiss German, for which parallel corpora are scarce. Our efforts on resource collection and MT design have yielded:

- a small Swiss German / High German parallel corpus of about 60k words;
- a larger list of resources which await digitization and alignment;
- three solutions for input normalization, to address variability of region and spelling;
- a baseline GSW-to-DE MT system reaching 36 BLEU points.

Among the three normalization strategies, we found that character-based neural MT was the most promising one. Moreover, we found that MT quality depended more strongly on the regional rather than topical similarity of test vs. training data.

These findings will be helpful to design MT systems for spoken dialects without standardized spellings, such as numerous regional languages across Africa or Asia, which are natural means of communication in social media.

8. Acknowledgments

9. Bibliographical References

Ali, A. M., Nakov, P., Bell, P., and Renals, S. (2017). WERd: Using social text spelling variants for evaluating dialectal speech recognition. *arXiv preprint arXiv:1709.07484*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pages 12–58, Baltimore, MD, USA.

Bradbury, J., Merity, S., Xiong, C., and Socher, R. (2017). Quasi-recurrent neural networks. In *Proceedings of the Int. Conf. on Learning Representations (ICLR)*, Toulon, France.

Carl, M., Melero, M., Badia, T., Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., and Yannoutsou, O. (2008). METIS-II: low resource machine translation. *Machine Translation*, 22(1):67–99.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Christen, H., Glaser, E., Friedli, M., and Renn, M. (2013). *Kleiner Sprachatlas der deutschen Schweiz*. Verlag Huber, Frauenfeld, Switzerland.

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Costa-jussà, M. R. and Fonollosa, J. A. R. (2016). Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 357–361, Berlin, Germany.

Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch. In *Proceedings of Interspeech*, Singapore.

Hildenbrandt, T., Moosmüller, S., and Neubarth, F. (2013). Orthographic encoding of the Viennese dialect for machine translation. In *Proceedings of the 6th Language & Technology Conference (LTC 2013)*, pages 7–9, Poznan, Poland.

Irvine, A. and Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource ma-

chine translation. In *Proceedings of the 8th Workshop on Statistical MT*, pages 262–270, Sofia, Bulgaria.

Irvine, A. and Callison-Burch, C. (2016). End-to-end statistical machine translation with zero or small parallel texts. *Natural Language Engineering*, 22(4):517–548.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 130–140, Avignon, France.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the ACL*, pages 48–54, Edmonton, Canada.

Lee, J., Cho, K., and Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics (TACL)*, 5:365–378.

Lewis, W. D., Munro, R., and Vogel, S. (2011). Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511, Edinburgh, UK.

Lewis, W. D. (2010). Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, and 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*, Saint-Raphaël, France.

Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015). Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA.

Pourdamghani, N. and Knight, K. (2017). Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2513–2518, Copenhagen, Denmark.

Russ, C. V. J. (1990). High Alemannic. In Charles V. J. Russ, editor, *The Dialects of Modern German. A linguistic survey*, pages 364–393. Routledge, London, UK.

Russ, C. V. J. (1994). *The German language today. A linguistic introduction*. Routledge, London, UK.

Samardžić, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference (LTC)*, Poznan, Poland.

Samardžić, T., Scherrer, Y., and Glaser, E. (2016). Archi-Mob – a corpus of spoken Swiss German. In *Proceedings of the 10th Int. Conf. on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Scherrer, Y. and Ljubešić, N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, Bochum, Germany.

Scherrer, Y. (2012a). *Generating Swiss German sentences from Standard German: a multi-dialectal approach*. Ph.D. thesis, University of Geneva, Switzerland.

Scherrer, Y. (2012b). Machine translation into multiple dialects: The example of Swiss German. In *7th SIDG Congress - Dialect 2.0*, Vienna, Austria.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1715–1725, Berlin, Germany.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the ACL (NAACL-HLT)*, pages 49–59, Montreal, Canada.