Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch

Philip N. Garner, David Imseng, Thomas Meyer

Idiap Research Institute, Martigny, Switzerland
{pgarner,dimseng,tmeyer}@idiap.ch

Abstract

Walliserdeutsch is a Swiss German dialect spoken in the south west of Switzerland. To investigate the potential of automatic speech processing of Walliserdeutsch, a small database was collected based mainly on broadcast news from a local radio station. Experiments suggest that automatic speech recognition is feasible: use of another (Swiss German) database shows that the small data size lends itself to bootstrapping from other data; use of Kullback-Leibler HMM suggests that phoneme mapping techniques can compensate for a grapheme-based dictionary. Experiments also indicate that statistical machine translation is feasible; the difficulty of small data size is offset by the close proximity to (high) German.

Index Terms: speech recognition, speech translation, dialect recognition

1. Introduction

Switzerland has four national languages: German (64%), French (20%), Italian (7%) and Rumantsch (<1%); the numbers in brackets indicate the percentage of the population speaking them[1], so French and German account for 84%. Whilst the French (spoken in the west) is simply an accented form of standard French, the German is highly dialectical. Typically, speakers speak a dialect representative of the region. In formal situations, or simply to be understood to visitors, the German speakers will switch to standard "high"-German (*Hochdeutsch*). The term "Swiss German" is taken to mean Swiss-accented high German. The German dialects are not mutually comprehensible. Furthermore, they have no standard written form.

Valais (*German: Wallis*) is a south-western alpine canton of Switzerland, perhaps best defined as encompassing the Rhone valley from the Rhone glacier in the east to lake Geneva in the north west. It borders France to the west and Italy to the south. The southern border extends from the Matterhorn in the east, almost to Mont Blanc[2] in the west. Linguistically, Valais[3] is bilingual: The western, low-Valais (*French: Bas Valais*) is French speaking; the eastern, high-Valais (*German: Oberwallis*) is German speaking. About two thirds of the population speak French, and one third speaks German. Figure 1 illustrates geographic and linguistic information about Valais.

The German dialect spoken in high-Valais is in fact a group of dialects, *Walliserdeutsch*. The numerous local variations can broadly be classified into two major idioms, a western group and eastern group [2], also shown in Figure 1. Walliserdeutsch is regarded as one of the most difficult to understand of the Swiss dialects and is considered to have a special status in terms of intonation [3].

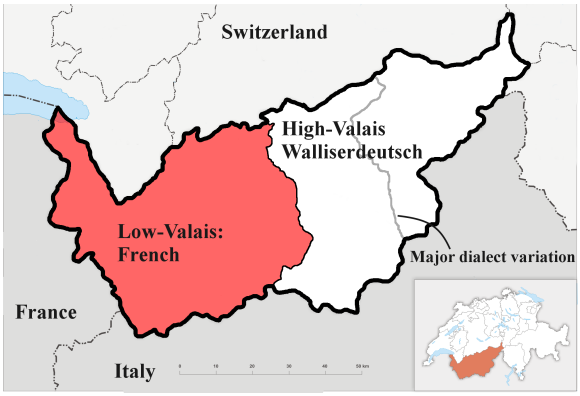

Figure 1: Geographic and linguistic information about Valais, a Swiss canton. The location of Valais within Switzerland is given in the lower right corner. Figure is derivative work from [1].

The fact that there is no written form, coupled with the ease of switching to high-German, means there are few language resources in Swiss German dialects. The purpose of the present study is to evaluate the feasibility of automatic language processing in such dialects. At the outset, we can reasonably state the following expectations:

1. It is unlikely that large amounts of data in dialect can be collected easily. The situation is akin to that of under-resourced language processing [4, 5].
2. Given the close relationship with German, it ought to be possible to "bootstrap" automatic processing using a larger database of high-German.
3. One significant difficulty will be the lack of phonetic dictionary.

In the remainder of the paper, we first describe the data collection and annotation process, and the resulting publicly available[4] database (Section 2). In Section 3, we then describe our automatic speech recognition (ASR) experiments; in Section 4 we give statistical machine translation (SMT) results from Walliserdeutsch to German achieving BLEU scores above 50.

2. Database

This section describes the database as well as the data collection and annotation process.

[1] http://www.swissinfo.ch/
[2] Mont Blanc is on the French-Italian border
[3] In English, we use the French name
[4] https://www.idiap.ch/dataset/walliserdeutsch

2.1. Sources

Given that Walliserdeutsch speakers will readily switch to Swiss German when necessary, there was no obvious source of recordings in dialect. However, two sources did come to light: The first was from a local radio station, *Radio Rottu Oberwallis*[5] (RRO), based in Visp. RRO is a typical *local* radio station, broadcasting in a mixture of Swiss German and dialect. However, a news bulletin is broadcast daily in dialect. The bulletins are available as *podcasts*, i.e., downloadable recordings in *mp3* format. The second source was from the national broadcaster, *Schweizer Radio und Fernsehen*[6] (SRF). SRF broadcasts predominantly in Swiss German. However, a (paid) search revealed local interest programs comprising recordings of (often older) people speaking in dialect.

2.2. Annotation

The annotation process was carried out by two native Walliserdeutsch speakers over the course of around 3 months in the summer of 2013. The bulk of the data were taken from the RRO news bulletins, as they could be downloaded as required.

The main issue with annotation was orthography. Walliserdeutsch has no standard written form. However, speakers will readily write it. This *de-facto* written form has become more common recently in text messages (SMS), and in social media such as facebook and twitter. Words are written either phonetically, or influenced by the orthography of Swiss German where the words are similar.

Two references were available to us. One was the "scripts", kindly made available by RRO for around two months of news bulletins. The other was the book *Wallissertitschi Weerter* [2]. The annotators were asked to keep to a standard defined by this book. However, the resulting annotations do deviate from this standard quite often. This is indicative of the fact that Walliserdeutsch speakers are simply not used to writing it, or have differing views on how to write it.

Practically, the annotation process was the same as that of [6]. Recordings were broken into segments of around 20 seconds, annotated, then checked twice via a web-based interface. Additionally, to train the SMT, parts of the RRO data were also transcribed in Swiss German.

2.3. Dictionary

For the creation of the Walliserdeutsch dictionary, we took local variants as well as Walliserdeutsch phonetics into account.

2.3.1. Dialect variations

Although Walliserdeutsch is understood over all of high-Valais, many local differences exist. However, the group of dialects can be divided into two main idioms, the western and the eastern variant. This classification originates from the Germanization that happened from east to west in the 8th and 9th century [2].

The main differences that came to light are vowel substitutions and vowel deletions. In the western variant, for example, many nouns end with the vowel *u*, whereas they end with the vowel *a* in the eastern variant. Furthermore, in the eastern variant, vowel deletions in the first syllable of verbs in the perfect tense are common.

[5] http://rro.ch
[6] http://www.srf.ch

German	Walliserdeutsch	Phonetic transcription
Schrank	**Sch**aft	**/sch/** /a/ /f/ /t/
Ein**k**auf	**Ch**öüf	/i/ **/ch/** /oe/ /ue/ /f/
ni**chts**	ni**x**	/n/ /i/ **/k/ /s/**

Table 1: Mappings of exemplar words to phonemes. The standard German writing, the de-facto Walliserdeutsch writing as well as the phonetic transcription is provided.

Set	# sent.	Amount of data
Training	1,502	6.8 h
Development	45	10 min
Testing	463	1.3 h

Table 2: The partitioning of the RRO dataset. The training set contains all the utterances for which we only have Walliserdeutsch transcripts. The remainder is randomly split into testing and development sets.

2.3.2. Phonetics

Given that one of our objectives was speech recognition and/or synthesis, a phonetic dictionary was necessary. One *easy* way to make such a dictionary was to simply map orthography to phonetics; the language is written phonetically if at all, so this is likely to be close to actual phonetics. The dictionary comprised a one-to-one letter to phoneme mapping. However, the German influence leads to certain substitutions, depicted in Table 1.

2.3.3. Dictionary Creation

The phonetics described above leads to a phonetic dictionary that can be discerned automatically from the orthography. However, speech recognition should (ideally) be dialect independent. In this sense, dialect variations should map to the same orthography. This issue of dialectical variation led to the creation of a more involved dictionary with three fields:

1. A canonical entry corresponding to the eastern variant.
2. A list of dialectical variants (most often just the eastern and western variant).
3. The high-German translation.

The annotation was then defined to be the dialectical variation. Given the dictionary, this variation could be mapped to a canonical form if necessary. The final dictionary contained just over 18,500 pronunciations.

3. Automatic Speech Recognition

For the automatic speech recognition (ASR) experiments in this paper, we only used the RRO data. To perform the experiments, we first split the data into training, development and testing set as shown in Table 2. The training set contains all the utterances for which we only have Walliserdeutsch transcripts. The utterances, which are transcribed in Walliserdeutsch and Swiss German, are randomly split into testing (90%) and development (10%) sets.

3.1. Language model

It must be stressed that we have no extra data for language modelling. At the outset, a bigram model was built using *all* the RRO data. This produced a model with a perplexity of 613 on

the test set (463 sentences). This is high, especially for a model trained on testing data as well. However, it enables evaluation of ASR acoustic models without the difficulties of out of vocabulary words. All ASR results below should be regarded as optimistic; only relative results are meaningful.

3.2. Hypotheses under test

Walliserdeutsch is an extremely low resourced Swiss dialect with no standard written form. Aside from the basic feasibility test, these peculiarities lead to two hypotheses for the investigation.

3.2.1. Small amount of training data

Low resourced languages have been investigated in many recent studies [4, 5, 7]. One possibility to improve ASR of low resourced languages is the exploitation of foreign data, as for example done in [4]. In this paper, we train the DNN on 20 hours of Swiss German MediaParl (MP) data from the bilingual Valais parliament [6], then adapt to RRO, to test the hypothesis that such data ought to boost the performance of the Walliserdeutsch speech recogniser.

3.2.2. Uncertainty about the phone set

Since there is no standard written form of Walliserdeutsch, we simply mapped orthography to phonetics. Hence, the dictionary comprises a one-to-one letter to phoneme mapping. However, the phone set may be sub-optimal and lead to confusion. Recently, we investigated acoustic phone space transformation to map source to destination phonemes [8]. The underlying technique, Kullback-Leibler divergence based Hidden Markov models (KL-HMM), has been shown to work well in mismatched phone set scenarios [9]. This leads to a second hypothesis: if KL-HMM leads to an improvement for Walliserdeutsch then the phone set is probably sub-optimal.

3.3. Acoustic modelling

All the acoustic modelling techniques use standard 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC) including deltas and double deltas, and are trained using the Kaldi speech recognition toolkit [10]. Decoding is done using a graph that is built from the bigram language model described above.

HMM/GMM The HMM/GMM system serves as a standard baseline system. The tree has 1,567 leaves.

HMM/DNN The HMM/DNN system is based on a deep neural network that has three hidden layers with 2,000 units each. The output layer has 1,567 units trained on alignments created with the HMM/GMM system. The neural network considers a temporal context of 9 frames (4 preceding and following frames). The DNN is pre-trained using the Restricted Boltzmann Machines (RBM) technique.

KL-HMM The KL-HMM acoustic modeling technique is trained by minimizing the Kullback–Leibler divergence between the posterior outputs of the DNN and the state distributions of the HMM. The KL-HMM system uses 10,000 HMM states.

3.4. Results

The results of the ASR experiments are shown in Table 3. The HMM/DNN system outperforms the HMM/GMM baseline; this is in line with the current trend in the state of the art. Applying KL-HMM on top of the HMM/DNN yields improvement; this demonstrates the hypothesis that the (grapheme-based) phoneme set is probably sub-optimal. The exploitation of Swiss German MP data also yields improvement compared to the HMM/DNN system; this demonstrates the hypothesis that Walliserdeutsch lends itself to bootstrapping. The KL-HMM and foreign data exploitation are complementary; applying both techniques yields the best performing ASR system.

System	WER
HMM/GMM	23.6 %
HMM/DNN	20.3 %
+KL	19.6 %
+MP	19.8 %
+MP+KL	**19.4** %

Table 3: ASR results on the RRO database. The best performing system exploits MediaParl data and uses KL-HMM acoustic modeling.

4. Statistical Machine Translation

Once the database and dictionary for the Walliserdeutsch (WD) / German (DE) were in place, we built a so-called phrase-based statistical machine translation (SMT) model that is able to translate any text transcribed in WD dialect to DE standard language. In purely statistical phrase-based SMT, no (manually crafted) translation or syntactical rules are applied. Rather, such a model learns, from sentence- and word aligned parallel texts, the chunks or "phrases" that correspond to each other in source and target language. The phrases thereby are arbitrary and not necessarily linguistically motivated. A translation probability is calculated over the frequency of a phrase pair appearing in a parallel corpus [11].

There is very little work on automatic translation of Swiss German dialects. Scherrer [12] has built a hybrid, rule-based and statistical MT system that is able to go from standard DE to WD dialect, with the supplementary help of geolocation data (indexed maps), to find the right dialectal variant and to produce translation output that is a) closer to the dialect in question than standard German and b) closer to the dialect in question than to the other four dialects considered. These experiments are therefore not directly comparable to ours, as we did not integrate any rules nor any map data into our system and considered the opposite translation direction: to translate from a dialect (WD) to standard German (DE).

4.1. Data and hypothesis

Since we only have WD/DE translations for the development and testing sets of Table 2, we used these two database parts for the translations experiments. The "ASR development set" was used as the "SMT tuning set" and the "ASR testing set" was split into two folds for "SMT training and testing". and we run a two-fold evaluation procedure. The statistics of the datasets used for SMT are given in Table 4.

Due to the tedious and costly procedure of human translating such texts, our database is very limited for the SMT task, where normally hundreds of thousands of sentences are needed in order to reach reasonable translation quality. We could however attach our rather large dictionary of WD/DE word correspondences directly to the already parallel training data in order to increase at least vocabulary coverage. The size and distribu-

Stage	Model	Size (# sent./# words)
Training	Fold 1	230 sent. 43,694 words
	Fold 2	233 sent. 43,694 words
Tuning	–	45 sent.
Testing	Fold 1	230 sent.
	Fold 2	233 sent.

Table 4: Sizes and distributions of data to build different SMT models for WD/DE translation.

tion of the final training set is again shown in Table 4.

The data sparsity leads to a hypothesis for the SMT experiment: At the outset, we would expect translation between two dialects to work very well. However, if the performance is not good, it can be attributed to data sparsity.

4.2. SMT models

All DE data was then used to build a language model with the IRSTLM toolkit [13] as an additional feature component of the SMT system. For the translation model, we word-aligned our corpus with Giza++ [14] and built the phrase table with the baseline Moses SMT toolkit pipeline [15]. For tuning we used the Minimum Error Rate Training (MERT) algorithm [16].

We built two translation models, referred to as fold 1 and 2 in order to cross-validate their performance by testing them on fold 2 and 1, respectively. The tuning set and the language model were the same in both settings. These translation tests allowed us to test for the upper bound of translation performance as in fold 1 and 2 manual gold transcriptions of the WD dialectal text were available.

A third and fourth testing scenario for the models was to have them decode the two same folds, but as they were directly from an ASR system, without gold transcription. For these experiments we used two different ASR systems, the *Baseline system* (HMM/GMM in Table 3) and the *Best system* (HMM/DNN+KL+MP in Table 3). The translation performance was likely to go down in this setting, as the texts stemming from ASR are likely to contain errors that are propagated to SMT and lead e.g. to out-of-vocabulary and therefore untranslated words, when not the full or correct word forms are output by ASR. To directly compare the influence of the ASR performance on the translation quality, we evaluated the baseline and the best system.

4.3. Evaluation

Translation quality assessment was carried out by using the BLEU metric [17] which is normally used for evaluating SMT systems. The fully automatic metric compares a system (or candidate) translation against one human reference translation (or several if at hand) and is based on matching n-gram counts, normalised over the document length. Its values (the higher the better) range from 0 to 100, and usually lie between 20 to 35 points for state-of-the-art systems in European language pair settings. Table 5 provides the scores for all model and testing settings. Our BLEU scores are very high compared to usual values, especially for the gold transcript test sets. This can be explained by the fact that WD/DE are relatively closely related languages and our texts used for training/tuning/testing were relatively similar in terms of genre and style (news articles). Further, this demonstrates the hypothesis that the closeness of the languages outweighs the data sparsity. When translating ASR output directly, there is a huge drop (about 20 BLEU points) for each of the folds, which can be explained by the errors, inconsistencies and unrecognised words associated with ASR output. Nevertheless, when testing on the best ASR output, around 5 BLEU points can be recovered, which is significant in terms of translation quality and shows the direct influence of the correctness of the input conveyed to the SMT system.

Model	Test	BLEU
Fold 1	Fold 2	75.10
	Fold 2 – Baseline ASR	55.93
	Fold 2 – Best ASR	60.96
Fold 2	Fold 1	71.07
	Fold 1 – Baseline ASR	53.92
	Fold 1 – Best ASR	58.17

Table 5: BLEU scores for the translation models evaluated in different settings and on different inputs from gold transcripts and ASR experiments.

System	WER WD	WER DE
Baseline ASR (HMM/GMM)	23.6%	53.0%
Best ASR (HMM/DNN+KL+MP)	19.4%	49.8%

Table 6: Word error rate (WER) scores for the translated ASR output.

We also scored the translated ASR outputs in terms of WER. Averaged results over both folds are given in Table 6.

5. Conclusion

It seems reasonable to conclude that automatic language processing of Walliserdeutsch is feasible, even using the rather small database described above. The ASR results suggest that data sparsity issues can be addressed by bootstrapping a dialect system from Swiss German; it follows that high German may also be appropriate. The KL-HMM result suggests that the grapheme-based phonetic dictionary is not optimal. Nevertheless, the ASR produces reasonable results, and the KL-HMM can be thought of as mitigating the dictionary problem. It remains unclear how much improvement a bespoke dictionary would yield. Certainly, language modelling remains a problem; bootstrapping from German being one obvious potential solution. Translation results are promising, and demonstrate that the data sparsity issue is offset to a large extent by the closeness of the two languages. Not surprisingly, translation accuracy is directly dependent upon ASR accuracy. The authors have no reason to believe that these results do not generalise to other Swiss German dialects.

6. Acknowledgements

The authors are grateful to Radio Rottu in Visp for initial discussions and for providing scripts for some broadcasts. We are also grateful to Raphael Ullmann for help with the initial stages of the project.

This work was supported by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)", and by armasuisse, the Competence Center for Procurement, Technology, Real Estate and Geodata within the Federal Department of Defence, Civil Protection and Sport.

7. References

[1] Tschubby, "Karte Kanton Wallis Bezirke 2010," http://upload.wikimedia.org/wikipedia/commons/e/e6/Karte_Kanton_Wallis_Bezirke_2010.png.

[2] A. Grichting, *Wallissertitschi Weerter*. Radio Rottu Oberwallis und Walliser Bote, 2011.

[3] A. Leemann and L. Zuberbühler, "Declarative sentence intonation patterns in 8 Swiss German dialects," in *Proc. of Interspeech*, 2010, pp. 1768–1771.

[4] D. Imseng, H. Bourlard, and P. N. Garner, "Boosting under-resourced speech recognizers by exploiting out of language data - case study on Afrikaans," in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages*, 2012, pp. 60–67.

[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource LVCSR systems." in *Prof. of Interspeech*, 2010, pp. 877–880.

[6] D. Imseng, H. Bourlard, H. Caesar, P. N. Garner, G. Lecorvé, and A. Nanchen, "MediaParl: Bilingual mixed language accented speech database," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 263–268.

[7] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of ICASSP*, 2013.

[8] D. Imseng, H. Bourlard, J. Dines, P. N. Garner, and M. Magimai-Doss, "Applying multi- and cross-lingual stochastic phone space transformations to non-native speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, 2013.

[9] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. of ICASSP*, 2012, pp. 4869–4872.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011.

[11] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, Cambridge UK, 2010.

[12] Y. Scherrer, "Machine translation into multiple dialects: The example of swiss german," in *7th SIDG Congress - Dialect 2.0*, Vienna, Austria, 2012.

[13] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Brisbane, Australia, 2008.

[14] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbs, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, Prague, Czech Republic, 2007, pp. 177–180.

[16] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 2003, pp. 160–167.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of Machine Translation," in *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, 2002, pp. 311–318.