Natural Language Processing for the Swiss German Dialect Area

Scherrer, Yves; Rambow, Owen

Natural Language Processing for the Swiss German dialect area

Yves Scherrer
LATL
Université de Genève
Genève, Switzerland
yves.scherrer@unige.ch

Owen Rambow
CCLS
Columbia University
New York, USA
rambow@ccls.columbia.edu

Abstract

This paper discusses work on data collection for Swiss German dialects taking into account the continuous nature of the dialect landscape, and proposes to integrate these data into natural language processing models. We present knowledge-based models for machine translation into any Swiss German dialect, for dialect identification, and for multi-dialectal parsing. In a dialect continuum, rules cannot be applied uniformly, but have restricted validity in well-defined geographic areas. Therefore, the rules are parametrized with probability maps extracted from dialectological atlases.

1 Introduction

Most work in natural language processing is geared towards written, standardized language varieties. This focus is generally justified on practical grounds of data availability and socio-economical relevance, but does not always reflect the linguistic reality. In this paper, we propose to include continuous linguistic variation in existing natural language processing (NLP) models, as it is encountered in various dialect landscapes.

Besides continuous variation on the geographical axis, dialects represent some interesting challenges for NLP. As mostly spoken language varieties, few data are available in written form, and those which exist do not follow binding spelling rules. Moreover, dialect use is often restricted to certain social contexts or modalities (diglossia), reducing further the availability of resources.

In contrast, two facts facilitate the development of NLP models for dialects. First, dialects are generally in a historic and etymological relationship with a standardized language variety for which linguistic resources are more readily accessible. Second, many dialects have been studied systematically by dialectologists, and these results can be exploited in a computational setting. The work presented here is applied to Swiss German dialects; this dialect area is well documented by dialectological research and is among the most vital ones in Europe in terms of social acceptance and media exposure.

This paper introduces ongoing work on a rule-based system that accounts for the differences between Standard German and the Swiss German dialects, using rules that are aware of their geographical application area. The system proposed here transforms morphologically annotated Standard German words into Swiss German words depending on the dialect area. The obvious use case for these components is (word-by-word) machine translation, which will be described in section 5.1. We also present two other applications that indirectly rely on these components, dialect identification (Section 5.2) and dialect parsing (Section 5.3).

We will start by presenting some related work (Section 2) and by giving an overview of the particularities of Swiss German dialects (Section 3). In Section 4, we present original work on data collection and show how probabilistic maps can be extracted from existing dialectological research and incorporated in the rule base. Then, the applications introduced above will be presented, and the paper will conclude with the discussion of some preliminary results.

2 Related work

Several research projects have dealt with dialect machine translation. The most similar work is the thesis by Forst (2002) on machine translation from Standard German to the Zurich Swiss German dialect

within the LFG framework. Delmonte et al. (2009) adapt recent statistical machine translation tools to translate between English and the Italian Veneto dialect, using Standard Italian as a pivot language. In contrast, we are interested in handling a continuum of dialects.

Translation between dialectal variants can be viewed as a case of translation between closely related languages. In this domain, one may cite works on different Slavic languages (Hajic et al., 2003) and on the Romance languages of Spain (Corbí-Bellot et al., 2005).

Dialect parsing models have also been developed in the last years. Chiang et al. (2006) build a synchronous grammar for Modern Standard Arabic and the Levantine Arabic dialect. Their approach is essentially corpus-driven on the Standard Arabic side, but includes manual adaptations on the dialect side. Vaillant (2008) presents a factorized model that covers a group of French-based Creole languages of the West-Atlantic area. His model relies on hand-crafted rules within the TAG framework and uses a numeric parameter to specify a particular dialect.

With the exception of Vaillant (2008), the cited papers only deal with one aspect of dialect NLP, namely the fact that dialects are similar to a related standardized language. They do not address the issue of interdialectal variation. Vaillant's factorized model does deal with several related dialects, but conceives the different dialects as discrete entities which can be clearly distinguished. While this view is probably justified for Caribbean creoles spoken on different islands, we argue that it cannot be maintained for dialect areas lacking major topographical and political borders, such as German-speaking Switzerland.

One important part of our work deals with bilingual lexicon induction. For closely related languages or dialects, cognate words with high phonetic (or graphemic) similarity play a crucial role. Such methods have been presented in various contexts, e.g. by Mann and Yarowsky (2001), Koehn and Knight (2002), or Kondrak and Sherif (2006). Scherrer (2007) uses similarity models based on learned and hand-crafted rules to induce Standard German – Bern Swiss German word pairs.

Dialect identification has usually been studied from a speech processing point of view. Biadsy et al. (2009) classify speech material from four Arabic dialects plus Modern Standard Arabic. They first run a phone recognizer on the speech input and use the resulting transcription to build a trigram language model. As we are dealing with written dialect data, only the second step is relevant to our work. Classification is done by minimizing the perplexity of the trigram models on the test segment.

An original approach to the identification of Swiss German dialects has been taken by the *Chochichästli-Orakel*.[1] By specifying the pronunciation of ten predefined phonetic and lexical cues, this web site creates a probability map that shows the likelihood of these pronunciations in the Swiss German dialect area. Our model is heavily inspired by this work, but extends the set of cues to the entire lexicon.

Computational methods are also used in dialectometry to assess differences between dialects with objective numerical measures. The most practical approach is to compare words of different dialects with edit distance metrics (Nerbonne and Heeringa, 2001). On the basis of these distance data, dialects can be classified with clustering methods. While the Swiss German data described here provide a valid base for dialect classification, this task is not the object of this paper.

3 Swiss German dialects

The German-speaking area of Switzerland encompasses the Northeastern two thirds of the Swiss territory. Likewise, about two thirds of the Swiss population define (any variety of) German as their first language.

It is usually admitted that the sociolinguistic configuration of German-speaking Switzerland is a model case of diglossia, i.e. an environment in which two linguistic varieties are used complementarily in functionally different contexts. In German-speaking Switzerland, dialects are used in speech, while Standard German is used nearly exclusively in written contexts.

Despite the preference for spoken dialect use, written dialect use has become popular in electronic media like blogs, SMS, e-mail and chatrooms. The Alemannic Wikipedia[2] contains about 6000 articles, among which many are written in a Swiss German dialect. However, all this data is very heterogeneous in terms of the dialects used, spelling conventions and genres. Moreover, parallel corpora are virtually non-existent because need for translation is weak in a diglossic society.

[1] `http://dialects.from.ch`

[2] `http://als.wikipedia.org`; besides Swiss German, the Alemannic dialect group encompasses Alsatian, South-West German Alemannic and Vorarlberg dialects of Austria.

Standard German	Swiss German	Validity Region	Example	
u	*ue*	all	*gut* → *guet*	'good'
au	*uu* [uː]	except Unterwalden	*Haus* → *Huus*	'house'
u	*ü*	South (Alpine)	*(Haus →) Huus* → *Hüüs*	
ü	*i*	South (Alpine), Basel	*müssen* → *miesse*	'must'
k (word-initial)	*ch* [x]	except Basel, Graubünden	*Kind* → *Chind*	'child'
l	*u*	Bern	*alt* → *aut*	'old'
nd (word-final)	*ng* [ŋ]	Bern	*Hund* → *Hung*	'dog'

Table 1: Phonetic transformations occurring in Swiss German dialects. The first column specifies the Standard German graphemes. The second column presents one possible outcome in Swiss German; the area of validity of that outcome is specified in the third column. An example is given in the fourth column.

The classification of Swiss German dialects is commonly based on administrative and topographical criteria. Although these non-linguistic borders have influenced dialects to various degrees, the resulting classification does not always match the linguistic reality. Our model does not presuppose any dialect classification. We conceive of the Swiss German dialect area as a continuum in which certain phenomena show more clear-cut borders than others. The nature of dialect borders is to be inferred from the data.[3]

Swiss German has been subject to dialectological research since the beginning of the 20th century. One of the major contributions is the *Sprachatlas der deutschen Schweiz* (SDS), a linguistic atlas that covers phonetic, morphological and lexical differences. Data collection and publication were carried out between 1939 and 1997 (Hotzenköcherle et al., 1962 1997). The lack of syntactic data in the SDS has led to a follow-up project called *Syntaktischer Atlas der deutschen Schweiz* (SADS), whose results are soon to be published (Bucheli and Glaser, 2002). Besides these large-scale projects, there also exist grammars and lexicons for specific dialects, as well as general presentations of Swiss German.

Swiss German dialects differ in many ways from Standard German. In the following sections, some of the differences in phonetics, lexicon, morphology and syntax are presented.

3.1 Phonetic dialect differences

Table 1 shows some of the most frequent phonetic transformations occurring in Swiss German dialects. Note that our system applies to written representations of dialect according to the Dieth spelling conventions (Dieth, 1986). As a consequence, the examples are based on written dialect representations, with IPA symbols added for convenience in ambiguous cases. The Dieth rules are characterized by a transparent grapheme-phone correspondence and are generally quite well respected – implicitly or explicitly – by dialect writers.

The SDS contains two volumes of phonetic data, amounting to about 400 maps.

3.2 Lexical dialect differences

Some differences at the word level cannot be accounted for by pure phonetic alternations. One reason are idiosyncrasies in the phonetic evolution of high frequency words (e.g. *und* 'and' is reduced to *u* in Bern dialect, where the phonetic rules would rather suggest **ung*). Another reason is the use of different lexemes altogether (e.g. *immer* 'always' corresponds to *geng*, *immer*, or *all*, depending on the dialect).

The SDS contains five volumes of lexical data, although large parts of it concern aspects of rural life of the 1940s-1950s and are thus becoming obsolete. The *Wörterbuch der schweizerdeutschen Sprache*[4] contains a much broader spectrum of lexical data, but its contents are difficult to access. Word lists published on the internet by dialect enthusiasts certainly offer smaller coverage and lower quality, but can present an interesting alternative to extend lexical coverage.

[3]Nonetheless, we will refer to political entities for convenience when describing interdialectal differences in the following sections of this paper.

[4]The *Wörterbuch der schweizerdeutschen Sprache* is a major lexicographic research project (Staub et al., 1881). Work started in 1881 and is scheduled to be fully achieved by 2020. Unfortunately, most of this work is not available in digital format, nor with precise geographical references. These issues are currently being addressed for the Austrian dialect lexicon in the project *dbo@ema* (Wandl-Vogt, 2008).

	1st Pl.	2nd Pl.	3rd Pl.
Standard	*-en*	*-t*	*-en*
West	*-e*	*-et*	*-e*
Wallis	*-e*	*-et*	*-end, -und*
East	*-ed*	*-ed*	*-ed*
Central	*-id*	*-id*	*-id*
Graubünden	*-end*	*-end*	*-end*

Table 2: Indicative plural suffixes of regular verbs in different Swiss German dialects. The first row shows the Standard German endings for comparison.

3.3 Morphological and morphosyntactic dialect differences

Swiss German inflectional paradigms are generally reduced with respect to Standard German. Translation into Swiss German requires thus a set of morphosyntactic rules that insert, remove or reorder words in a sentence. For example, the lack of preterite tense in Swiss German requires all preterite sentences to be restructured as present perfect sentences. Similarly, the lack of genitive case gives rise to different syntactic structures to express possession. In contrast, Swiss German has clitic and non-clitic pronouns, a distinction that is not made in written Standard German.

On a purely morphological level, one can mention the verb plural suffixes, which offer surprisingly rich (and diachronically stable) interdialectal variation, as illustrated in Table 2. Minor interdialectal differences also exist in noun and adjective inflection.

In derivational morphology, the most salient dialect difference concerns diminutive suffixes: Swiss German has *-li* (or *-ji* / *-i* in Wallis dialect) instead of Standard German *-chen* and *-lein*.

Volume 3 of the SDS deals with morphology in the form of about 250 maps. Many morphosyntactic features of Swiss German are also investigated in the SADS survey.

4 Georeferenced transfer rules

The system proposed here contains sets of phonetic, lexical, morphological rules as illustrated in the examples above. Some of these rules apply uniformly to all Swiss German dialects, but most of them yield different outcomes (variants) in different dialect regions. For example, the phonetic rule governing the transformation of word-final *-nd* will have four distinct variants *-nd, -ng, -n, -nt* (the *-nd* variant has been mentioned in Table 1). Each variant is linked to a probability map that specifies the areas of its validity. We refer to a rule, its associated variants and probability maps as a *georeferenced transfer rule*.

The maps for the georeferenced rules are extracted from the SDS. Currently, the system contains about 100 phonetic rules based on about 50 SDS maps. This corresponds to a fairly complete coverage. Lexical rules are currently limited to some high-frequency function words that are referenced in the SDS (about 100 rules). Morphological coverage is complete for regular inflection patterns and corresponds to about 60 rules. Some morphosyntactic and syntactic rules using unpublished SADS material have been added for testing purposes, but coverage is so far very limited.

4.1 Map generation

The SDS consists of hand-drawn maps on which different symbols represent different dialectal variants. Figure 1 shows an example of an original SDS map.

In a first preprocessing step, the hand-drawn map is digitized manually with the help of a geographical information system. The result is shown in Figure 2. To speed up this process, variants that are used in less than ten inquiry points are omitted. This can be justified by the observation by Christen (1998) that many small-scale variants in verbal morphology have disappeared since the data collection of the SDS in the 1940s and 1950s, while large-scale variants have not. We also collapse minor phonetic variants which cannot be distinguished in the Dieth spelling system.

The SDS maps, hand-drawn or digitized, are point maps. They only cover the inquiry points (about 600 in the case of the SDS), but do not provide information about the variants used in other locations. Therefore, a further preprocessing step interpolates the digitized point maps to obtain surface maps. We follow Rumpf et al. (2009) to create kernel density estimators for each variant. This method is less sensible to outliers than simpler linear interpolation methods. The resulting surface maps are then normalized such that at each point of the surface, the weights of all variants sum up to 1. These normalized weights can be interpreted as conditional probabilities $p(v \mid t)$, where v is a variant and t is the geographic location (represented as a pair of longitude and latitude coordinates). Figure 3 shows the resulting surface maps for each variant. Surface maps are generated with a resolution of one point per square kilometer.

Formally, the application of a rule is represented

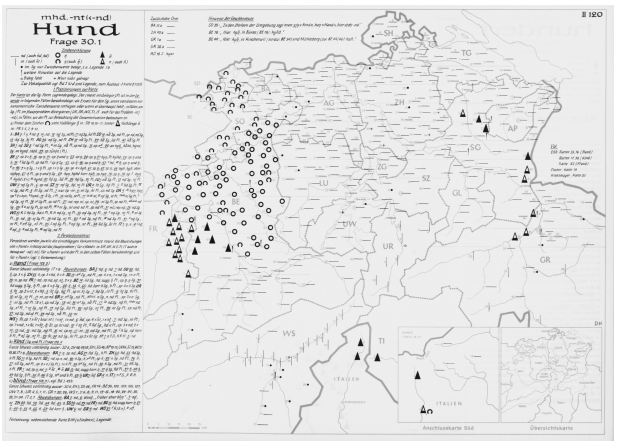

Figure 1: Original SDS map for the transformation of word-final *-nd*. The map contains four major linguistic variants, symbolized by horizontal lines (*-nd*), vertical lines (*-nt*), circles (*-ng*), and triangles (*-n*) respectively. Minor linguistic variants are symbolized by different types of circles and triangles.

Figure 2: Digitized version of the map in Figure 1.

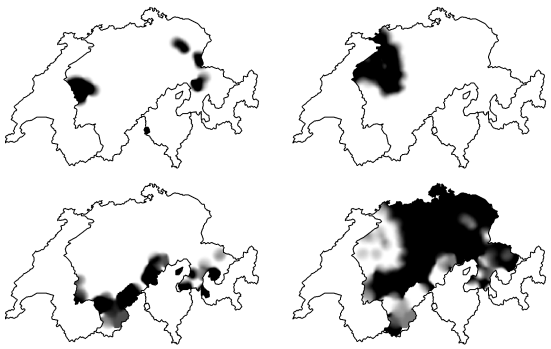

Figure 3: Interpolated surface maps for each variant of the map in Figure 2. Black areas represent a probability of 1, white areas a probability of 0.

as follows:

$$R_{ij}(w_k) = w_{k+1}$$

where R_i represents the rule which addresses the ith phenomenon, and R_{ij} represents the jth variant of rule R_i. The result of applying R_{ij} to the word form w_k is w_{k+1}. The maps define probability distributions over rule variants at each geographic point t situated in German-speaking Switzerland (we call this set of points *GSS*), such that at any given point $t \in GSS$, the probabilities of all variants sum up to 1:

$$\forall i \underset{t \in GSS}{\forall} \sum_j p(R_{ij} \mid t) = 1$$

5 Three applications

The phonetic, lexical and morphological rules presented above allow to transform Standard German words into words of a specific Swiss German dialect. This rule base can be utilized in several NLP applications. The following sections will discuss the three tasks machine translation, dialect identification and dialect parsing.

5.1 Machine translation

Machine translation of a Standard German sentence begins with a syntactic and morphological analysis. Every word of the sentence is lemmatized (including compound word splitting), part-of-speech tagged and annotated with morphological features. The goal of this preprocessing is to take advantage of existing Standard German analysis tools to reduce ambiguity and to resolve some specific issues of German grammar like noun composition.[5]

Then, each annotated word is translated. Starting with the base form of the Standard German word, lexical rules are used to build a new Swiss German base form. If no lexical rule applies, the phonetic rules are used instead.

For example, the Standard German word *nichts* 'nothing' triggers a lexical rule; one variant of this rule, valid in the Northeast, yields the form *nünt*. In contrast, no lexical rule applies to the Standard German word *suchen*-VVFIN-3.Pl.Pres.Ind 'they search', which therefore triggers the following phonetic rules in Graubünden dialect:

[5]For the time being, we perform this analysis simply by looking up word forms in a Standard German lexicon extracted from the Tiger treebank. Work is underway to merge the output of parsers like BitPar (Schmid, 2004) or Fips (Wehrli, 2007), part-of-speech taggers like TnT (Brants, 2000), and morphological analyzers like Morphisto (Zielinski and Simon, 2008) in order to provide accurate and complete annotation.

– $u \rightarrow u$ (not $\ddot{u}$)
– $u \rightarrow ue$
– $e \rightarrow a$ (in diphthong)

and results in the stem *suach-*.

The georeferenced morphological rules represent a morphological generator for Swiss German: given a Swiss German base form and a set of morphological features, it creates an inflected dialectal form. In the above example, the Graubünden dialect suffix *-end* is attached, resulting in the inflected form *suachend*.

This approach of analyzing and recreating word forms may sound overly complicated, but allows generalization to (the morphological part of) morphosyntactic restructuring like the transformation of preterite tense verbs into past participles. Similarly, it is easy to account for the fact that more Swiss German nouns build their plural with an *umlaut* than in Standard German.

The target dialect is fixed by the user by selecting the coordinates of a point t situated in German-speaking Switzerland.[6] As illustrated above, the rules are applied sequentially, such that a Standard German word w_0 yields an intermediate form w_1 after the first transformation, and the final Swiss German form w_n after n transformations.

The probability resulting from the application of one rule variant R_{ij} transforming string w_k to w_{k+1} is read off the associated variant map at that point t:

$$p(w_k \rightarrow w_{k+1} \mid t) = p(R_{ij} \mid t) \quad s.t. \quad w_{k+1} = R_{ij}(w_k)$$

A derivation from w_0 to w_n, using n transfer rules, yields the following probability:

$$p(w_0 \stackrel{*}{\rightarrow} w_n \mid t) = \prod_{k=0}^{n-1} p(w_k \rightarrow w_{k+1} \mid t)$$

The number n of rules in a derivation is not known in advance and depends on the structure of the word.

Note however that in transition zones, several variants of the same rule may apply. All rule applications are thus potentially ambiguous and lead to multiple derivations.[7] Among multiple derivations, we choose the one that maximizes the probability.

[6] Points are specified in the Swiss Coordinate System, either numerically or through a web interface based on Google Maps. The *nichts* example above assumed a point in the Northeast, while the *suchen* example assumed a point in the Southeast (Graubünden).

[7] We did not encounter cases where multiple derivations lead from the same Standard German word to the same Swiss German word. In that case, we would have to sum the probabilities of the different derivations.

The translation model presented here does not account for morphosyntactic adaptations and word reordering. While this word-by-word approach is sufficient in many cases, there are some important (morpho-)syntactic differences between Standard German and Swiss German (see section 3.3). Therefore, additional syntactic rules will provide context-dependent morphological and phonetic adaptations as well as word reordering in future versions of our system.

5.2 Dialect identification

Dialect identification or, more generally, language identification is commonly based on distributions of letters or letter n-grams. While these approaches have worked very well for many languages, they may be unable to distinguish related dialects with very similar phoneme and grapheme inventories. Moreover, they require training corpora for all dialects, which may not be available.

As an alternative, we propose to identify entire words in a text and find out in which regions these particular forms occur. This approach is similar to the *Chochichästli-Orakel*, but instead of using a small predefined set of cues, we consider as cues all dialect words that can be generated from Standard German words with the help of the transfer rules presented above. To do this, we first generate a list of Swiss German word forms, and then match the words occurring in the test segment with this list.

We obtained a list of lemmatized and morphologically annotated Standard German words by extracting all leaf nodes of the Tiger Treebank (Brants et al., 2002). Word forms that appeared only once in the corpus were eliminated. These Standard German words were then translated with our system. In contrast to the machine translation task, the target dialect was not specified. All potentially occurring dialect forms were generated and stored together with their validity maps.

For example, the *suchen* example yielded one single form *suachend* when restricted to a point in the Graubünden dialect area (for the translation task), but 27 forms when the target dialect was not specified (for the dialect identification task).

At test time, the test segment is tokenized, and each word of the segment is looked up in the Swiss German lexicon. (If the lookup fails, the word is skipped.) We then produce a probability map of each Swiss German word w_n by pointwise multiplication of all variant maps that contributed to generating it from Standard German word w_0, in the same way as

in the machine translation task illustrated above.

Note that a dialect form can be the result of more than one derivation. For example, the three derivations *sind*-VAFIN $\stackrel{*}{\rightarrow}$ *si* (valid only in Western dialects), *sein*-PPOSAT $\stackrel{*}{\rightarrow}$ *si* (in Western and Central dialects), and *sie*-PPER $\stackrel{*}{\rightarrow}$ *si* (in the majority of Swiss German dialects) lead to the same dialectal form *si*. In these cases, we take the pointwise maximum probability of all derivations $D(w)$ that lead to a Swiss German word form *w*:

$$\underset{t \in GSS}{\forall} \; p(w \mid t) = \max_{d \in D(w)} p(d \mid t)$$

Once we have obtained a map for each word of the segment, we merge them according to the following formula: The probability map of a segment *s* corresponds to the pointwise average of the probabilities of the words *w* contained in the sequence:

$$p(s \mid t) = \frac{\sum_{w \in s} p(w \mid t)}{|s|}$$

This is thus essentially a bag-of-words approach to dialect identification that does not include any notion of syntax.

5.3 Dialect parsing

A multidialectal parser can be defined in the following way: a source text, not annotated with its dialect, is to be analyzed syntactically. The goal is to jointly optimize the quality of the syntactic analysis and the dialect region the text comes from.

The exact implementation of dialect parsing is an object of future research. However, some key elements of this approach can already be specified.

Constituent parsers commonly consist of a grammar and of a lexicon. In a multidialectal parsing setting, the grammar rules as well as the lexicon entries have to be linked to probability maps that specify their area of validity. The lexicon built for the dialect identification task can be reused for parsing without further modifications. For the grammar however, more work is needed. A Swiss German grammar can be built by extracting a Standard German grammar from a treebank and manually modifying it to match the syntactic particularities of Swiss German (Chiang et al., 2006). In this process, the syntactic machine translation rules may serve as a guideline.

Instead of directly annotating each syntactic rule with a dialect parameter (Vaillant, 2008), we indirectly annotate it with a map containing its probability distribution over the dialect area.

	Word-based	Trigram
Paragraphs (Wikipedia)	52.2%	86.7%
Sentences (Wikipedia)	31.3%	67.8%
Sentences (Non-Wiki.)	41.4%	44.4%

Table 3: F-measure values averaged over all six dialects.

6 Evaluation

6.1 Dialect identification

In terms of annotated resources for evaluation, dialect identification is the least demanding task: it requires texts that are annotated with their respective dialect. Such data can be extracted from the Alemannic Wikipedia, where many Swiss German articles are annotated with their author's dialect.

We extracted about ten paragraphs of text for six dialect regions: Basel, Bern, Eastern Switzerland, Fribourg, Wallis and Zurich. The paragraphs amount to a total of 100 sentences per region.[8] The surfaces of these six regions were defined using political (canton) boundaries and the German-French language border.

The dialect identification system scored each paragraph *s* with a probability map. We calculated the average probability value for each of the six regions and annotated the paragraph with the region obtaining the highest value:

$$Region(s) = \arg\max_{Region} \left(\frac{\sum_{t \in Region} p(s \mid t)}{|Region|} \right)$$

We tested entire paragraphs and single sentences, and repeated both experiments with a simple trigram model trained on Wikipedia data of similar size. The results of these tests are summarized in Table 3 (first two rows).

We suspected that the outstanding results of the trigram model were due to some kind of overfitting. It turned out that the number of Swiss German Wikipedia authors is very low (typically, one or two active writers per dialect), and that every author uses distinctive spelling conventions and writes about specific subjects. For instance, most Zurich German articles are about Swiss politicians, while many Eastern Swiss German articles are about religious subjects. Our hypothesis was thus that the

[8] The choice of the dialects and the size of the corpus was largely determined by the data available. The average sentence length was 17.8 words per sentence.

n-gram model learned to recognize a specific author and/or topic rather than a dialect.

In order to confirm this hypothesis, we collected another small data set from various web resources (not from Wikipedia, 50 sentences per dialect).[9] Table 3 (last row) indeed confirms our suspicion. The performance of the trigram model dropped by more than 20 percent (absolute), while the word-based model surprisingly performed better on the second test set than on the Wikipedia data. One possible explanation is the influence of Standard German spelling on the Wikipedia data, given that many Swiss German articles are translations of their Standard German counterparts. However, we have not thoroughly verified this claim.

While our dialect identification model does not outperform the trigram model, recent adaptations show promising results. First, the dialect annotation based on average probability values penalizes large and heterogeneous regions, where a high-probability sub-region would be cancelled out by a low-probability sub-region. Using maximum instead of average could improve the dialect annotation. Second, not all derivations are equally relevant; for example, word frequency information can provide a crucial clue to weighting derivations.

6.2 Machine translation and parsing

For the evaluation of the machine translation task, we might again resort to data from Wikipedia. As said above, many articles are translations from Standard German and can serve as a small parallel (or at least comparable) corpus. In addition, we plan to extract Swiss German text from other sources and have it translated it into Standard German.

Current translation evaluation metrics like BLEU or TER only use binary measures of word match. Given the importance of phonetic transformations in our approach, and given the problems arising from lacking spelling conventions, finer-grained metrics might be needed in order to account for different degrees of word similarity.

While the machine translation system has not been evaluated yet, a prototype version is accessible on the Web.[10]

For parsing, the data requirements are even more demanding. Syntactically annotated Swiss German dialect texts do not currently exist to our knowledge, so that a small evaluation tree bank would have to be created from scratch.

[9] The gold dialect of these texts could be identified through metadata (URL of the website, name and address of the author, etc.) in all but one case; this information was checked for plausibility by the authors.

[10] `http://latlcui.unige.ch/~yves/`

7 Conclusion

We have presented an approach to natural language processing that takes into account the specificities of Swiss German dialects. Dialects have too often been viewed as homogeneous entities clearly distinguishable from neighbouring dialects. This assumption is difficult to maintain in many dialect areas. Rather, each dialect is defined as a unique combination of variants; some variants may be shared with adjacent dialects, others may act as discriminating features (isoglosses). Our approach reflects this point of view by modelling an entire dialect continuum.

The data for our model come from dialectological research. Dialects may be among the few language varieties where linguistically processed material is not significatively costlier to obtain than raw textual data. Indeed, data-driven approaches would have to deal with data sparseness and dialectal diversity at the same time. While processing dialectological data is tedious, we have proposed several tasks that allow the data to be reused.

This paper reflects the current status of ongoing work; while data collection is fairly complete, evaluating and tuning the proposed models will be a high priority in the near future.

Besides presenting a novel approach to NLP tasks, we argue that dialectological research can also profit from this work. Dialectological research has traditionally suffered from lack of dissemination among laymen: dialect atlases and lexicons are complex pieces of work and often difficult to access. Dynamic models of dialect use could bring dialectological research closer to a large audience, especially if they are freely accessible on the internet.

Acknowledgements

Part of this work was carried out during the first author's stay at Columbia University, New York, funded by the Swiss National Science Foundation (grant PBGEP1-125929).

References

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL'09 Workshop on Computational Approaches to Semitic Languages*, Athens.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of NAACL 2000*, Seattle, USA.

Claudia Bucheli and Elvira Glaser. 2002. The Syntactic Atlas of Swiss German dialects: empirical and methodological problems. In Sjef Barbiers, Leonie Cornips, and Susanne van der Kleij, editors, *Syntactic Microvariation*, volume II. Meertens Institute Electronic Publications in Linguistics, Amsterdam.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of EACL'06*, pages 369–376, Trento.

Helen Christen. 1998. *Dialekt im Alltag: eine empirische Untersuchung zur lokalen Komponente heutiger schweizerdeutscher Varietäten*. Niemeyer, Tübingen.

Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor, and Kepa Sarasola. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of EAMT'05*, pages 79–86, Budapest.

Rodolfo Delmonte, Antonella Bristot, Sara Tonelli, and Emanuele Pianta. 2009. English/Veneto resource poor machine translation with STILVEN. In *Proceedings of ISMTCL*, volume 33 of *Bulag*, pages 82–89, Besançon.

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

Martin Forst. 2002. La traduction automatique dans le cadre formel de la LFG – Un système de traduction entre l'allemand standard et le zurichois. In *Publications du CTL*, volume 41. Université de Lausanne.

Jan Hajic, Petr Homola, and Vladislav Kubon. 2003. A simple multilingual machine translation system. In *Proceedings of the Machine Translation Summit XI*, pages 157–164, New Orleans.

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

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon*, pages 9–16, Philadelphia.

Grzegorz Kondrak and Tarek Sherif. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the ACL Workshop on Linguistic Distances*, pages 43–50, Sydney.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL'01*, Pittsburgh.

John Nerbonne and Wilbert Heeringa. 2001. Computational comparison and classification of dialects. In *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*, number 9, pages 69–83. Edizioni dell'Orso, Alessandria.

Jonas Rumpf, Simon Pickl, Stephan Elspaß, Werner König, and Volker Schmidt. 2009. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, 76(3).

Yves Scherrer. 2007. Adaptive string distance measures for bilingual dialect lexicon induction. In *Proceedings of the ACL'07 Student Research Workshop*, pages 55–60, Prague.

Helmut Schmid. 2004. Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proceedings of COLING'04*, Geneva, Switzerland.

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

Pascal Vaillant. 2008. A layered grammar model: Using tree-adjoining grammars to build a common syntactic kernel for related dialects. In *Proceedings of TAG+9 2008*, pages 157–164, Tübingen.

Eveline Wandl-Vogt. 2008. An der Schnittstelle von Dialektwörterbuch und Sprachatlas: Das Projekt "Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)". In Stephan Elspaß and Werner König, editors, *Germanistische Linguistik 190-191. Sprachgeographie digital. Die neue Generation der Sprachatlanten*, pages 197–212. Olms, Hildesheim.

Éric Wehrli. 2007. Fips, a "deep" linguistic multilingual parser. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pages 120–127, Prague.

Andrea Zielinski and Christian Simon. 2008. Morphisto – an open-source morphological analyzer for German. In *Proceedings of FSMNLP'08*, Ispra, Italy.