Word-based dialect identification with georeferenced rules

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

We present a novel approach for (written) dialect identification based on the discriminative potential of entire words. We generate Swiss German dialect words from a Standard German lexicon with the help of hand-crafted phonetic/graphemic rules that are associated with occurrence maps extracted from a linguistic atlas created through extensive empirical fieldwork. In comparison with a character-n-gram approach to dialect identification, our model is more robust to individual spelling differences, which are frequently encountered in non-standardized dialect writing. Moreover, it covers the whole Swiss German dialect continuum, which trained models struggle to achieve due to sparsity of training data.

1 Introduction

Dialect identification (dialect ID) can be viewed as an instance of language identification (language ID) where the different languages are very closely related. Written language ID has been a popular research object in the last few decades, and relatively simple algorithms have proved to be very successful. The central question of language ID is the following: given a segment of text, which one of a predefined set of languages is this segment written in? Language identification is thus a classification problem.

Dialect identification comes in two flavors: spoken dialect ID and written dialect ID. These two tasks are rather different. Spoken dialect ID relies on speech recognition techniques which may not cope well with dialectal diversity. However, the acoustic signal is also available as input. Written dialect ID has to deal with non-standardized spellings that may occult real dialectal differences. Moreover, some phonetic distinctions cannot be expressed in orthographic writing systems and limit the input cues in comparison with spoken dialect ID.

This paper deals with written dialect ID, applied to the Swiss German dialect area. An important aspect of our model is its conception of the dialect area as a continuum without clear-cut borders. Our dialect ID model follows a bag-of-words approach based on the assumption that every dialectal word form is defined by a probability with which it may occur in each geographic area. By combining the cues of all words of a sentence, it should be possible to obtain a fairly reliable geographic localization of that sentence.

The main challenge is to create a lexicon of dialect word forms and their associated probability maps. We start with a Standard German word list and use a set of phonetic, morphological and lexical rules to obtain the Swiss German forms. These rules are manually extracted from a linguistic atlas. This linguistic atlas of Swiss German dialects is the result of decades-long empirical fieldwork.

This paper is organized as follows. We start with an overview of relevant research (Section 2) and present the characteristics of the Swiss German dialect area (Section 3). Section 4 deals with the implementation of word transformation rules and the corresponding extraction of probability maps from the linguistic atlas of German-speaking Switzerland. We present our dialect ID model in Section 5 and discuss its performance in Section 6 by relating it to a baseline n-gram model.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1151–1161,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

2 Related work

Various language identification methods have been proposed in the last three decades. Hughes et al. (2006) and Řehůřek and Kolkus (2009) provide recent overviews of different approaches. One of the simplest and most popular approaches is based on character n-gram sequences (Cavnar and Trenkle, 1994). For each language, a character n-gram language model is learned, and test segments are scored by all available language models and labeled with the best scoring language model. Related approaches involve more sophisticated learning techniques (feature-based models, SVM and other kernel-based methods).

A completely different approach relies on the identification of entire high-frequency words in the test segment (Ingle, 1980). Other models have proposed to use morpho-syntactic information.

Dialect ID has usually been studied from a speech processing point of view. For instance, 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. 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 words, the 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.

As mentioned, the ID model is based on a large Swiss German lexicon. Its derivation from a Standard German lexicon can be viewed as a case of lexicon induction. Lexicon induction methods for closely related languages using phonetic similarity have been proposed by Mann and Yarowsky (2001) and Schafer and Yarowsky (2002), and applied to Swiss German data by Scherrer (2007).

The extraction of digital data from hand-drawn dialectological maps is a time-consuming task. Therefore, the data should be made available for different uses. Our Swiss German raw data is accessible on an interactive web page (Scherrer, 2010), and we have proposed ideas for reusing this data for machine translation and dialect parsing (Scherrer and Rambow, 2010). An overview of digital dialectological maps for other languages is available on `http://www.ericwheeler.ca/atlaslist`.

3 Swiss German dialects

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

In German-speaking Switzerland, dialects are used in speech, while Standard German is used nearly exclusively in written contexts (diglossia). It follows that all (adult) Swiss Germans are bidialectal: they master their local dialect and Standard German. In addition, they usually have no difficulties understanding Swiss German dialects other than their own.

Despite the preference for spoken dialect use, written dialect data has been produced in the form of dialect literature and transcriptions of speech recordings made for scientific purposes. More recently, written dialect has been used in electronic media like blogs, SMS, e-mail and chatrooms. The Alemannic Wikipedia contains about 6000 articles, among which many are written in a Swiss German dialect.[2] However, all this data is very heterogeneous in terms of the dialects used, spelling conventions and genre.

4 Georeferenced word transformation rules

The key component of the proposed dialect ID model is an automatically generated list of Swiss German word forms, each of which is associated with a map that specifies its likelihood of occurrence over German-speaking Switzerland. This word list is generated with the help of a set of transformation rules, taking a list of Standard German words as a starting point. In this section, we present the different types of rules and how they can be extracted from a dialectological atlas.

[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.

4.1 Orthography

Our system generates written dialect words according to the Dieth spelling conventions without diacritics (Dieth, 1986).[3] These are characterized by a transparent grapheme-phone correspondence and are widely used by dialect writers. However, they are by no means enforced or even taught.

This lack of standardization is problematic for dialect ID. We have noted two major types of deviations from the Dieth spelling conventions in our data. First, Standard German orthography may unduly influence dialect spelling. For example, *spiele* is modelled after Standard German *spielen* 'to play', although the vowel is a short monophthong in Swiss German and should thus be written *spile* (*ie* represents a diphthong in Dieth spelling). Second, dialect writers do not always distinguish short and long vowels, while the Dieth conventions always use letter doubling to indicate vowel lengthening. Future work will incorporate these fluctuations directly into the dialect ID model.

Because of our focus on written dialect, the following discussion will be based on written representations, but IPA equivalents are added for convenience.

4.2 Phonetic rules

Our work is based on the assumption that many words show predictable phonetic differences between Standard German and the different Swiss German dialects. Hence, in many cases, it is not necessary to explicitly model word-to-word correspondences, but a set of phonetic rules suffices to correctly transform words.

For example, the word-final sequence *nd* [nd̥] (as in Standard German *Hund* 'dog'[4]) is maintained in most Swiss German dialects. However, it has to be transformed to *ng* [ŋ] in Berne dialect, to *nn* [n] in Fribourg dialect, and to *nt* [nt] in Valais and Uri dialects.

This phenomenon is captured in our system by four transformation rules *nd* → *nd*, *nd* → *ng*, *nd* → *nn* and *nd* → *nt*. Each rule is *georeferenced*, i.e. linked to a probability map that specifies its validity in every geographic point. These four *rules* capture one single linguistic *phenomenon*: their left-hand side is the same, and they are geographically complementary.

Some rules apply uniformly to all Swiss German dialects (e.g. the transformation *st* [st] → *scht* [ʃt]). These rules do not immediately contribute to the dialect identification task, but they help to obtain correct Swiss German forms that contain other phonemes with better localization potential.

More information about the creation of the probability maps is given in Sections 4.5 and 4.6.

4.3 Lexical rules

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. Standard German *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. Standard German *immer* 'always' corresponds to *geng*, *immer*, or *all*, depending on the dialect). We currently use lexical rules mainly for function words and irregular verb stems.

4.4 Morphological rules

The transformation process from inflected Standard German word forms to inflected Swiss German word forms is done in two steps. First, the word stem is adapted with phonetic or lexical rules, and then, the affixes are generated according to the morphological features of the word.

Inflection markers also provide dialect discrimination potential. For example, the verbal plural suffixes offer a surprisingly rich (and diachronically stable) interdialectal variation pattern.

4.5 The linguistic atlas SDS

One of the largest research projects in Swiss German dialectology has been the elaboration of the *Sprachatlas der deutschen Schweiz* (SDS), a linguistic atlas that covers phonetic, morphological and lexical differences of Swiss German dialects. Data collection and publication were carried out between 1939 and 1997 (Hotzenköcherle et al., 1962-1997). Linguistic data were collected in about 600 villages (*inquiry points*) of German-speaking Switzerland, and

[3] Of course, these spelling conventions make use of umlauts like in Standard German. There is another variant of the Dieth conventions that uses additional diacritics for finer-grained phonetic distinctions.

[4] Standard German *nd* is always pronounced [nt] following a general final devoicing rule; we neglect that artifact as we rely only on graphemic representations.

resulted in about 1500 published maps (see Figure 1 for an example).

Each map represents a linguistic phenomenon that potentially yields a set of transformation rules. For our experiments, we selected a subset of the maps according to the perceived importance of the described phenomena. There is no one-to-one correspondence between maps and implemented phenomena, for several reasons. First, some SDS maps represent information that is best analyzed as several distinct phenomena. Second, a set of maps may illustrate the same phenomenon with different words and slightly different geographic distributions. Third, some maps describe (especially lexical) phenomena that are becoming obsolete and that we chose to omit.

As a result, our rule base contains about 300 phonetic rules covering 130 phenomena, 540 lexical rules covering 250 phenomena and 130 morphological rules covering 60 phenomena. We believe this coverage to be sufficient for the dialect ID task.

4.6 Map digitization and interpolation

Recall the *nd*-example used to illustrate the phonetic rules above. Figure 1 shows a reproduction of the original, hand-drawn SDS map related to this phenomenon. Different symbols represent different phonetic variants of the phenomenon.[5] We will use this example in this section to explain the preprocessing steps involved in the creation of georeferenced rules.

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. (Many of these small-scale variants likely have disappeared since the data collection in the 1940s.) 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, 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

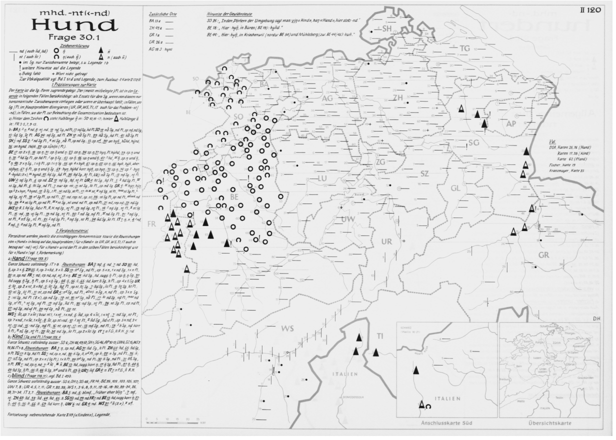

Figure 1: Original SDS map for the transformation of word-final *-nd*. The map contains four major linguistic variants, symbolized by horizontal lines (*-nd*), vertical lines (*-nt*), circles (*-ng*), and triangles (*-nn*) respectively. Minor linguistic variants are symbolized by different types of circles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

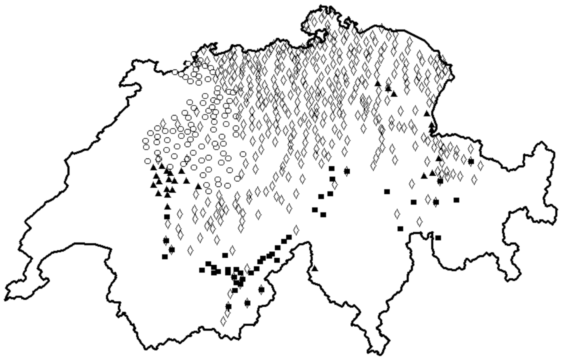

Figure 2: Digitized equivalent of the map in Figure 1.

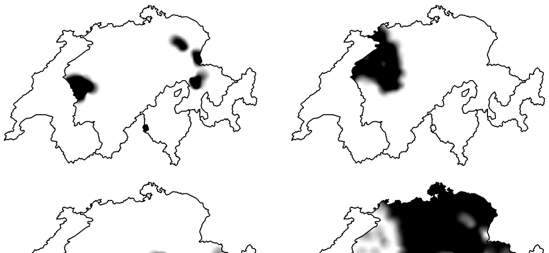

Figure 3: Interpolated surface maps for the variants *-nn* (upper left), *-ng* (upper right), *-nt* (lower left) and *-nd* (lower right). Black areas represent a probability of 1, white areas a probability of 0.

[5] We define a *variant* simply as a string that may occur on the right-hand side of a transformation rule.

less sensitive to outliers than simpler linear interpolation methods.[6] 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 of the corresponding transfer rule: $p(r \mid t)$, where r is the rule and t is the geographic location (represented as a pair of longitude and latitude coordinates) situated in German-speaking Switzerland. (We call the set of all points in German-speaking Switzerland *GSS*.) Figure 3 shows the resulting surface maps for each variant. Surface maps are generated with a resolution of one point per square kilometer.

As mentioned above, rules with a common left-hand side are grouped into phenomena, such that at any given point $t \in GSS$, the probabilities of all rules r describing a phenomenon Ph sum up to 1:

$$\underset{t \in GSS}{\forall} \sum_{r \in Ph} p(r \mid t) = 1$$

5 The model

The dialect ID system consists of a Swiss German lexicon that associates word forms with their geographical extension (Section 5.1), and of a testing procedure that splits a sentence into words, looks up their geographical extensions in the lexicon, and condenses the word-level maps into a sentence-level map (Sections 5.2 to 5.4).

5.1 Creating a Swiss German lexicon

The Swiss German word form lexicon is created with the help of the georeferenced transfer rules presented above. These rules require a lemmatized, POS-tagged and morphologically disambiguated Standard German word as an input and generate a set of dialect word/map tuples: each resulting dialect word is associated with a probability map that specifies its likelihood in each geographic point.

To obtain a Standard German word list, we extracted all leaf nodes of the TIGER treebank (Brants et al., 2002), which are lemmatized and morphologically annotated. These data also allowed us to obtain word frequency counts. We discarded words with one single occurrence in the TIGER treebank, as well as forms that contained the genitive case or preterite tense attribute (the corresponding grammatical categories do not exist in Swiss German dialects).

The transfer rules are then applied sequentially on each word of this list. The notation $w_0 \overset{*}{\rightarrow} w_n$ represents an iterative derivation leading from a Standard German word w_0 to a dialectal word form w_n by the application of n transfer rules of the type $w_i \rightarrow w_{i+1}$. The probability of a derivation corresponds to the joint probability of the rules it consists of. Hence, the probability map of a derivation is defined as the **pointwise product** of all rule maps it consists of:

$$\underset{t \in GSS}{\forall} p(w_0 \overset{*}{\rightarrow} w_n \mid t) = \prod_{k=0}^{n-1} p(w_i \rightarrow w_{i+1} \mid t)$$

Note that in dialectological transition zones, there may be several valid outcomes for a given w_0.

The Standard German word list extracted from TIGER contains about 36,000 entries. The derived Swiss German word list contains 560,000 word forms, each of which is associated with a map that specifies its regional distribution.[7] Note that proper nouns and words tagged as “foreign material” were not transformed. Derivations that did not obtain a probability higher than 0.1 anywhere (because of geographically incompatible transformations) were discarded.

5.2 Word lookup and dialect identification

At test time, the goal is to compute a probability map for a text segment of unknown origin.[8] As a preprocessing step, the segment is tokenized, punctuation markers are removed and all words are converted to lower case.

The identification process can be broken down in three levels:

1. The probability map of a text segment depends on the probability maps of the words contained in the segment.

2. The probability map of a word depends on the probability maps of the derivations that yield the word.

[6] A comparison of different interpolation methods will be the object of future work.

[7] Technically, we do not store the probability map, but the sequence of rule variants involved in the derivation. The probability map is restored from this rule sequence at test time.

[8] The model does not require the material to be syntactically well-formed. Although we use complete sentences to test the system, any sequence of words is accepted.

3. The probability map of a derivation depends on the probability maps of the rules it consists of.

In practice, every word of a given text segment is looked up in the lexicon. If this lookup does not succeed (either because its Standard German equivalent did not appear in the TIGER treebank, or because the rule base lacked a relevant rule), the word is skipped. Otherwise, the lookup yields m derivations from m different Standard German words.[9] The lexicon already contains the probability maps of the derivations (see 5.1), so that the third level does not need to be discussed here. Let us thus explain the first two levels in more detail, in reverse order.

5.3 Computing the probability map for a word

A dialectal word form may originate in different Standard German words. For example, the three derivations *sind [VAFIN]* $\xrightarrow{*}$ *si* (valid only in Western dialects), *sein [PPOSAT]* $\xrightarrow{*}$ *si* (in Western and Central dialects), and *sie [PPER]* $\xrightarrow{*}$ *si* (in the majority of Swiss German dialects) all lead to the same dialectal form *si*.

Our system does not take the syntactic context into account and therefore cannot determine which derivation is the correct one. We approximate by choosing the most probable one in each geographic location. The probability map of a Swiss German word w is thus defined as the **pointwise maximum**[10] of all derivations leading to w, starting with different Standard German words $w_0^{(j)}$:

$$\underset{t \in GSS}{\forall} \; p(w \mid t) = \max_j p(w_0^{(j)} \xrightarrow{*} w \mid t)$$

This formula does not take into account the relative frequency of the different derivations of a word. This may lead to unintuitive results. Consider the two derivations *der [ART]* $\xrightarrow{*}$ *dr* (valid only in Western dialects) and *Dr. [NN]* $\xrightarrow{*}$ *dr* (valid in all dialects). The occurrence of the article *dr* in a dialect text is a good indicator for Western Swiss dialects, but it is completely masked by the potential presence of the abreviation *Dr.* in all dialects. We can avoid this by weighting the derivations by the **word frequency** of w_0: the article *der* is much more frequent than the abreviation *Dr.* and is thus given more weight in the identification task. This weighting can be justified on dialectological grounds: frequently used words tend to show higher interdialectal variation than rare words.

Another assumption in the above formula is that each derivation has the same **discriminative potential**. Again, this is not true: a derivation that is valid in only 10% of the Swiss German dialect area is much more informative than a derivation that is valid in 95% of the dialect area. Therefore, we propose to weight each derivation by the proportional size of its validity area. The discriminative potential of a derivation d is defined as follows:[11]

$$DP(d) = 1 - \frac{\sum_{t \in GSS} p(d \mid t)}{|GSS|}$$

The experiments in Section 6 will show the relative impact of these two weighting techniques and of the combination of both with respect to the unweighted map computation.

5.4 Computing the probability map for a segment

The probability of a text segment s can be defined as the joint probability of all words w contained in the segment. Again, we compute the **pointwise product** of all word maps. In contrast to 5.1, we performed some smoothing in order to prevent erroneous word derivations from completely zeroing out the probabilities. We assumed a minimum word probability of $\phi = 0.1$ for all words in all geographic points:

$$\underset{t \in GSS}{\forall} \; p(s \mid t) = \prod_{w \in s} \max(\phi, p(w \mid t))$$

Erroneous derivations were mainly due to non-implemented lexical exceptions.

6 Experiments and results

6.1 Data

In order to evaluate our model, we need texts annotated with their gold dialect. We have chosen to use the Alemannic Wikipedia as a main data source.

[9] Theoretically, two derivations can originate at the same Standard German word and yield the same Swiss German word, but nevertheless use different rules. Our system handles such cases as well, but we are not aware of such cases occurring with the current rule base.

[10] Note that these derivations are alternatives and not joint events. This is thus not a joint probability.

[11] d is a notational abreviation for $w_0 \xrightarrow{*} w_n$.

Wikipedia name	Abbr.	Pop.	Surface
Baseldytsch	BA	8%	1%
Bärndütsch	BE	17%	13%
Seislertütsch	FR	2%	1%
Ostschwizertütsch	OS	14%	8%
Wallisertiitsch	WS	2%	7%
Züritüütsch	ZH	22%	4%

Table 1: The six dialect regions selected for our tests, with their annotation on Wikipedia and our abreviation. We also show the percentage of the German-speaking population living in the regions, and the percentage of the surface of the region relative to the entire country.

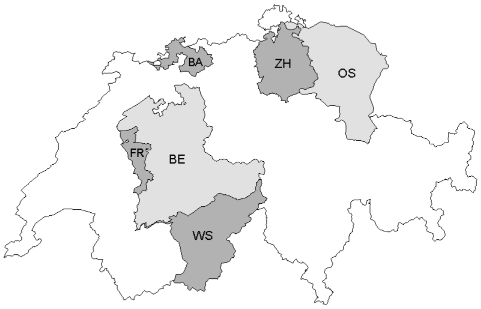

Figure 4: The localization of the six dialect regions used in our study.

The Alemannic Wikipedia allows authors to write articles in any dialect, and to annotate the articles with their dialect. Eight dialect categories contained more than 10 articles; we selected six dialects for our experiments (see Table 1 and Figure 4).

We compiled a test set consisting of 291 sentences, distributed across the six dialects according to their population size. The sentences were taken from different articles. In addition, we created a development set consisting of 550 sentences (100 per dialect, except FR, where only 50 sentences were available). This development set was also used to train the baseline model discussed in section 6.2.

In order to test the robustness of our model, we collected a second set of texts from various web sites other than Wikipedia. The gold dialect of these texts could be identified through metadata.[12] This information was checked for plausibility by the first author. The Web data set contains 144 sentences (again distributed according to population size) and is thus roughly half the size of the Wikipedia test set.

	Wikipedia			Web		
Dialect	P	R	F	P	R	F
BA	34	61	44	27	61	37
BE	78	51	61	51	47	49
FR	28	71	40	10	33	15
OS	63	64	64	50	38	43
WS	58	100	74	14	33	20
ZH	77	62	69	77	41	53
W. Avg.			62			46

Table 2: Performances of the 5-gram model on Wikipedia test data (left) and Web test data (right). The columns refer to precision, recall and F-measure respectively. The average is weighted by the relative population sizes of the dialect regions.

The Wikipedia data contains an average of 17.8 words per sentence, while the Web data shows 14.9 words per sentence on average.

6.2 Baseline: N-gram model

To compare our dialect ID model, we created a baseline system that uses a character-n-gram approach. This approach is fairly common for language ID and has also been successfully applied to dialect ID (Biadsy et al., 2009). However, it requires a certain amount of training data that may not be available for specific dialects, and it is uncertain how it performs with very similar dialects.

We trained 2-gram to 6-gram models for each dialect with the SRILM toolkit (Stolcke, 2002), using the Wikipedia development corpus. We scored each sentence of the Wikipedia test set with each dialect model. The predicted dialect was the one which obtained the lowest perplexity.[13]

The 5-gram model obtained the best overall performance, and results on the Wikipedia test set were surprisingly good (see Table 2, leftmost columns).[14] Note that in practice, 100% accuracy is not always achievable; a sentence may not contain a sufficient localization potential to assign it unambiguously to one dialect.

[12] We mainly chose websites of local sports and music clubs, whose localization allowed to determine the dialect of their content.

[13] We assume that all test sentences are written in one of the six dialects.

[14] All results represent percentage points. We omit decimal places as all values are based on 100 or less data points. We did not perform statistical significance tests on our data.

However, we suspect that these results are due to overfitting. It turns 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 ZH articles are about Swiss politicians, while many OS articles deal with religion and mysticism. Our hypothesis is thus that the n-gram model learns to recognize a specific author and/or topic rather than a dialect. This hypothesis is confirmed on the Web data set: the performances drop by 15 percentage points or more (same table, rightmost columns; the performance drops are similar for $n = [2..6]$).

In all our evaluations, the average F-measures for the different dialects are weighted according to the relative population sizes of the dialect regions because the size of the test corpus is proportional to population size (see Section 6.1).[15]

We acknowledge that a training corpus of only 100 sentences per dialect provides limited insight into the performance of the n-gram approach. We were able to double the training corpus size with additional Wikipedia sentences. With this extended corpus, the 4-gram model performed better than the 5-gram model. It yielded a weighted average F-measure of 79% on Wikipedia test data, but only 43% on Web data. The additional increase on Wikipedia data (+17% absolute with respect to the small training set), together with the decrease on Web data (−3% absolute) confirms our hypothesis of overfitting. An ideal training corpus should thus contain data from several sources per dialect.

To sum up, n-gram models can yield good performance even with similar dialects, but require large amounts of training data from different sources to achieve robust results. For many small-scale dialects, such data may not be available.

6.3 Our model

The n-gram system presented above has no geographic knowledge whatsoever; it just consists of six distinct language models that could be located anywhere. In contrast, our model yields probability maps of German-speaking Switzerland. In order to evaluate its performance, we thus had to determine the geographic localization of the six dialect regions defined by the Wikipedia authors (see Table 1). We defined the regions according to the respective canton boundaries and to the German-French language border in the case of bilingual cantons. The result of this mapping is shown in Figure 4.

The predicted dialect region of a sentence s is defined as the region in which the most probable point has a higher value than the most probable point in any other region:

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

Experiments were carried out for the four combinations of the two derivation-weighting techniques presented in Section 5.3 and for the two test sets (Wikipedia and Web). Results are displayed in Tables 3 to 6. The majority of FR sentences were misclassified as BE, which reflects the geographic and linguistic proximity of these regions.

The tables show that frequency weighting helps on both corpora: the discriminative potential only slightly improves performance on the web corpus. Crucially, the two techniques are additive, so in combination, they yield the best overall results. In comparison with the baseline model, there is a performance drop of about 16 percent absolute on Wikipedia data. In contrast, our model is very robust and outperforms the baseline model on the Web test set by about 7 percent absolute.

These results seem to confirm what we suggested above: that the n-gram model overfitted on the small Wikipedia training corpus. Nevertheless, it is still surprising that our model has a lower performance on Wikipedia than on Web data. The reason for this discrepancy probably lies in the spelling conventions assumed in the transformation rules: it seems that Web writers are closer to these (implicit) spelling conventions than Wikipedia authors. This may be explained by the fact that many Wikipedia articles are translations of existing Standard German articles, and that some words are not completely adapted to their dialectal form. Another reason could be that Wikipedia articles use a proportionally larger amount of proper nouns and low-frequency words which can-

[15] Roughly, this weighting can be viewed as a prior (the probability of the text being constant):

$$p(dialect \mid text) = p(text \mid dialect) * p(dialect)$$

	Wikipedia			Web		
Dialect	P	R	F	P	R	F
BA	41	19	26	80	22	35
BE	42	62	50	48	76	59
FR	0	0	0	17	33	22
OS	36	41	38	45	41	43
WS	3	14	5	8	33	13
ZH	65	33	44	62	37	46
W. Avg.			40			46

Table 3: Performances of the word-based model using unweighted derivation maps.

	Wikipedia			Web		
Dialect	P	R	F	P	R	F
BA	50	33	40	57	22	32
BE	47	60	53	60	79	68
FR	0	0	0	0	0	0
OS	29	31	30	46	50	48
WS	11	29	15	17	33	22
ZH	60	47	53	65	53	58
W. Avg.			44			**53**

Table 4: Performances of the word-based model using derivation maps weighted by word frequency.

	Wikipedia			Web		
Dialect	P	R	F	P	R	F
BA	34	31	32	38	17	23
BE	46	47	47	54	76	63
FR	11	14	13	20	33	25
OS	34	50	40	53	59	56
WS	5	14	7	0	0	0
ZH	47	27	34	75	43	55
W. Avg.			37			51

Table 5: Performances of the word-based model using derivation maps weighted by their discriminative potential.

	Wikipedia			Web		
Dialect	P	R	F	P	R	F
BA	46	28	35	33	11	17
BE	47	62	54	58	84	69
FR	0	0	0	20	33	25
OS	35	31	33	47	47	47
WS	8	29	13	14	33	20
ZH	63	53	58	66	51	58
W. Avg.			**46**			52

Table 6: Performances using derivation maps weighted by word frequency and discriminative potential.

not be found in the lexicon and which therefore reduce the localization potential of a sentence.

However, one should note that the word-based dialect ID model is not limited on the six dialect regions used for evaluation here. It can be used with any size and number of dialect regions of German-speaking Switzerland. This contrasts with the n-gram model which has to be trained specifically on every dialect region; in this case, the Swiss German Wikipedia only contains two additional dialect regions with an equivalent amount of data.

6.4 Variations

In the previous section, we have defined the predicted dialect region as the one in which the most probable point (**maximum**) has a higher probability than the most probable point of any other region. The results suggest that this metric penalizes small regions (BA, FR, ZH). In these cases, it is likely that the most probable point is slightly outside the region, but that the largest part of the probability mass is still inside the correct region. Therefore, we tested another approach: we defined the predicted dialect region as the one in which the **average** probability is higher than the average probability in any other region:

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

This metric effectively boosts the performance on the smaller regions, but comes at a cost for larger regions (Table 7). We also combined the two metrics by using the **maximum** metric for the three larger regions and the **average** metric for the three smaller ones (the cutoff lies at 5% of the Swiss territory). This combined metric further improves the performance of our system while relying on an objective measure of region surface.

We believe that region surface as such is not so crucial for the metrics discussed above, but rather serves as a proxy for linguistic heterogeneity. Geographically large regions like BE tend to have internal dialect variation, and averaging over all dialects in the region leads to low figures. In contrast, small regions show a quite homogeneous dialect landscape that may protrude over adjacent regions. In this case, the probability peak is less relevant than the average probability in the entire region. Future work will attempt to come up with more fine-grained measures of

	Wikipedia			Web		
Dialect	Max	Avg	Cmb	Max	Avg	Cmb
BA	35	32	32	17	43	43
BE	54	39	54	69	54	69
FR	0	7	7	25	11	11
OS	33	23	33	47	49	47
WS	13	13	13	20	31	20
ZH	58	60	60	58	68	68
W. Avg.	46	40	**47**	52	55	**58**

Table 7: Comparison of different evaluation metrics. All values refer to F-measures obtained with frequency and discriminative potential-weighted derivation maps. Max refers to the Maximum metric as used in Table 6. Avg refers to the average metric, and Cmb is the combination of both metrics depending on region surfaces. The underlined values in the Avg and Max columns represent those used for the Cmb metric.

linguistic heterogeneity in order to test these claims.

7 Future work

In our experiments, the word-based dialect identification model skipped about one third of all words (34% on the Wikipedia test set, 39% on the Web test set) because they could not be found in the lexicon. While our model does not require complete lexical coverage, this figure shows that the system can be improved. We see two main possibilities of improvement. First, the rule base can be extended to better account for lexical exceptions, orthographic variation and irregular morphology. Second, a mixed approach could combine the benefits of the word-based model with the n-gram model. This would require a larger, more heterogeneous set of training material for the latter in order to avoid overfitting. Additional training data could be extracted from the web and automatically annotated with the current model in a semi-supervised approach.

In the evaluation presented above, the task consisted of identifying the dialect of single sentences. However, one often has access to longer text segments, which makes our evaluation setup harder than necessary. This is especially important in situations where a single sentence may not always contain enough discriminative material to assign it to a unique dialect. Testing our dialect identification system on the paragraph or document level could thus provide more realistic results.

8 Conclusion

In this paper, we have compared two empirical methods for the task of dialect identification. The n-gram method is based on the approach most commonly used in NLP: it is a supervised machine learning approach where training data of the type we need to process is annotated with the desired outcome of the processing.

Our second approach – the main contribution of this paper – is quite different. The empirical component consists in a collection of data (the SDS atlas) which is not of the type we want to process, but rather embodies some features of the data we ultimately want to process. We therefore analyze this data in order to extract empirically grounded knowledge for more general use (the creation of the georeferenced rules), and then use this knowledge to perform the dialect ID task in conjunction with an unrelated data source (the Standard German corpus).

Our choice of method was of course related to the fact that few corpora, annotated or not, were available for our task. But beyond this constraint, we think it may be well worthwhile for NLP tasks in general to move away from a narrow machine learning paradigm (supervised or not) and to consider a broader set of empirical resources, sometimes requiring methods which are quite different from the prevalent ones.

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 *EACL 2009 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.

W. B. Cavnar and J. M. Trenkle. 1994. N-gram based text categorization. In *Proceedings of SDAIR'94*, Las Vegas.

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

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

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of LREC'06*, Genoa.

N. Ingle. 1980. A language identification table. *Technical Translation International*.

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

Radim Řehůřek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. In *Computational Linguistics and Intelligent Text Processing – Proceedings of CICLing 2009*, pages 357–368, Mexico. Springer.

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).

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of CoNLL'02*, pages 146–152, Taipei.

Yves Scherrer and Owen Rambow. 2010. Natural language processing for the Swiss German dialect area. In *Proceedings of KONVENS'10*, Saarbrücken.

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

Yves Scherrer. 2010. Des cartes dialectologiques numérisées pour le TALN. In *Proceedings of TALN'10*, Montréal.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP'02*, pages 901–904, Denver.