Recovering dialect geography from an unaligned comparable corpus

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

Abstract

This paper proposes a simple metric of dialect distance, based on the ratio between identical word pairs and cognate word pairs occurring in two texts. Different variations of this metric are tested on a corpus containing comparable texts from different Swiss German dialects and evaluated on the basis of spatial autocorrelation measures. The visualization of the results as cluster dendrograms shows that closely related dialects are reliably clustered together, while multidimensional scaling produces graphs that show high agreement with the geographic localization of the original texts.

1 Introduction

In the last few decades, dialectometry has emerged as a field of linguistics that investigates the application of statistical and mathematical methods in dialect research. Also called quantitative dialectology, one of its purposes is to discover the regional distribution of dialect similarities from aggregated data, such as those collected in dialectological surveys.

The work presented here aims to apply dialectometric analysis and visualization techniques to a different type of raw data. We argue that classical dialectological survey data are word-aligned by design, whereas our data set, a comparable multidialectal corpus, has to be word-aligned by automatic algorithms.

We proceed in two steps. First, we present a cognate identification algorithm that allows us to extract cognate word pairs from the corpus. Then, we measure how many of these cognate word pairs are identical. This ratio gives us a measure of dialectal distance between two texts that is then shown to correlate well with geographic distance. The visualization of the resulting data allows us to recover certain characteristics of the Swiss German dialect landscape.

The paper is structured as follows. In Section 2, the multidialectal corpus is presented. We then discuss how this corpus differs from classical dialectological data, and how we can use techniques from machine translation to extract the relevant data (Section 3). In Section 4, we define dialect distance as a function of the number of cognate word pairs and identical word pairs. Both types of word pairs are in turn defined by different thresholds of normalized Levenshtein distance. Section 5 deals with the evaluation and visualization of the resulting data, the latter in terms of clustering and multi-dimensional scaling. We discuss the results and conclude in Section 6.

2 Data: the Archimob corpus

The Archimob corpus used in our experiments is a corpus of transcribed speech, containing texts from multiple Swiss German dialects.

The Archimob project was started in 1998 as an oral history project with the aim of gathering and archiving the people's memory of the Second World War period in Switzerland.[1] 555 surviving witnesses were interviewed in all Swiss language regions. The interviews of the German-speaking witnesses were conducted in their local dialect.

With the goal of obtaining spontaneous dialect data to complement ongoing work on dialect syntax (Bucheli and Glaser, 2002; Friedli, 2006; Steiner, 2006), researchers at the Univer-

[1] Archimob stands for "Archives de la mobilisation"; see www.archimob.ch.

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 63–71,
Avignon, France, April 23 - 24 2012.

BE1142:	*de vatter ìsch lokomitiiffüerer gsìì / de ìsch dispensiert gsìì vom dienscht nattürlech / und / zwo schwöschtere / hani ghaa / wobii ei gsch / eini gschtoorben ìsch u di ander ìsch ìsch ime autersheim / u soo bini ufgwachse ir lenggass / mit em / pruefsleer / mit wiiterbiudig nächheer / (?)*
Translation:	the father has been a train driver / he has been dispensed from military service of course / and / two sisters / I have had / where one / one has died and the other is is in a home for the elderly / this is how I have grown up in the Lenggass / with a / apprenticeship / with further education afterwards / (?)

ZH1270:	*min vatter isch / eh eeh / schlosser hät er gleert / und und isch aber dän schofföör woorde dur en verwante wo bim S. z züri / gschafft hät und dè hät gsait / chum tue doch umsattle bim S. vediensch mee / und dän hät dèè schofföör gleert und das isch doozmaal ja na eener en sältene pruef gsii / dän hät dè das gleert und ich bin schtolz gsii das min / vatter en / pruef ghaa hät wo französischsch töönt hät oder schofföör / ich han gfunde das seig en waansinige pruef*
Translation:	my father has / eh eeh / been a locksmith apprentice / and and has then become a driver through a relative who has worked at S. in Zurich and he said / come and switch jobs, at S. you earn more / and then he was a driver apprentice and this was rather a rare job at that time / so he learned that and I was proud that my / father / had a job which sounded French, you know, *chauffeur* / I found that this was an extraordinary job

Figure 1: Excerpts of two informants' turns in the Archimob corpus. The excerpts contain identical cognate pairs like ⟨*vatter*, *vatter*⟩, and non-identical cognate pairs like ⟨*ìsch*, *isch*⟩.

sity of Zurich selected a subset of the Swiss German Archimob interviews and transcribed them.[2] The selection process ensured that only interviews from non-mobile speakers (speakers that have not spent long periods of their life outside of their native town) were retained, and that the most important dialect areas of German-speaking Switzerland were represented.

As a result, 16 interviews were selected for transcription, amounting to 26 hours of speech. All texts were anonymized. In order to ensure consistency, all texts were transcribed by the same person.

The interviews were transcribed using the spelling system of Dieth (1986). This is an orthographic transcription system which intends to be as phonetically transparent as possible, while remaining readable for readers accustomed to Standard German orthography (see Figure 1 for two examples). For instance, the Dieth guidelines distinguish *ì* (IPA [ɪ]) from *i* (IPA [i]), while Standard German spelling only uses *i*.

In our experiments, we discarded the interviewer's questions and only used the witnesses' turns. The whole corpus contains 183 000 words, with individual interviews ranging from 6 500 to 16 700 words. Excerpts of two interviews are shown in Figure 1. The place of residence of the witness was given in the corpus metadata.

It should be stressed that our data set is very small in comparison with other studies in the field: it contains 16 data points (texts) from 15 different locations. Moreover, some dialect areas are not represented in the sample (e.g. Graubünden in the South-East and Fribourg in the West).[3] Therefore, the goal of the present study cannot be to induce a precise dialect landscape of German-speaking Switzerland. Rather, we aim to find out if geographically close texts can be shown to be linguistically close, and if the most important dialectal divisions of German-speaking Switzerland are reflected in the classification of the texts.

3 Corpora and word alignment

3.1 Comparable corpora

The machine translation community generally distinguishes between parallel and comparable corpora (McEnery and Xiao, 2008). A *parallel corpus* consists of a source text and its translations into other languages. Hence, the different language versions share the same content and the same order of paragraphs and sentences. On the other hand, such corpora have been criticized for containing "translationese", i.e., wording which

[2]The corpus is not yet publicly available, awaiting the completion of further annotation layers.

[3]For an overview of the geographic distribution of the texts, see Figure 3.

is influenced by the grammatical and informational structure of the source text and which is not necessarily representative of the target language. In contrast, a *comparable corpus* is a collection of original texts of different languages that share similar form and content (typically, same genre, same domain and same time period).

The Archimob corpus can be qualified as comparable: all texts deal with the same subject and the same time period (life in Switzerland at the outbreak of the Second World War), and they are collected in the same way, in the form of oral interviews guided by an interviewer.

3.2 Word alignment in dialectology

Dialectological analyses rely on word-aligned data. Traditionally, dialectological data are collected in surveys with the help of questionnaires. A typical question usually intends to elicit the local words or pronunciations of a given concept. The mere fact that two responses are linked to the same question number of the questionnaire suffices to guarantee that they refer to the same concept. This property leads us to consider dialectological survey data as *word-aligned by design*.

In contrast, the Archimob corpus is not aligned. Again, algorithms for aligning words in parallel and comparable corpora have been proposed in the field of machine translation. For large parallel corpora, distributional alignment methods based solely on cooccurrence statistics are sufficient (Och and Ney, 2003; Koehn et al., 2007). For comparable corpora, the order and frequency of occurrence of the words cannot be used as alignment cues. Instead, the phonetic and orthographic structures are used to match similar word pairs (Simard et al., 1992; Koehn and Knight, 2002; Kondrak and Sherif, 2006). Obviously, this approach only works for cognate word pairs – word pairs with a common etymology and similar surface forms. This task is known as *cognate identification*.

In the next section, we detail how cognate identification is used to compute the distance between different dialect versions of a comparable corpus.

4 Computing the linguistic similarity of two comparable texts

The hypothesis put forward in this paper is that the linguistic similarity of two comparable texts can be approximated by the degree of similarity of the cognate word pairs occurring in the texts. Computing the similarity of two texts amounts to the following two tasks:

1. Given two texts, extract the set of word pairs that are considered cognates. This corresponds to the *cognate identification* task presented above.

2. Given a set of cognate word pairs, determine the proportion of word pairs that are considered identical.

The underlying intuition is that identically pronounced cognate words account for evidence that the two dialects are closely related, whereas differently pronounced cognate words are evidence that the two dialects are distant. Word pairs that are not cognates are not relevant for our similarity measure.

Let us illustrate the idea with an example:

(1) *es schtòòt nìd*

(2) *wil si nìd schtoot*

Intuitively, two cognate word pairs can be found in the texts (1) and (2): $\langle$*schtòòt*, *schtoot*$\rangle$ and $\langle$*nìd*, *nìd*$\rangle$.[4] The words *es, wil, si* do not have cognate equivalents in the other text. As a result, the two texts have a similarity of $\frac{1}{2}$, one of the two cognate pairs consisting of identical words.

In the example above, we have assumed informal meanings of *cognate word pair* and *identical word pair*. In the following sections, we define these concepts more precisely.

4.1 Identifying cognate word pairs

Most recently proposed cognate identification algorithms are based on variants of Levenshtein distance, or string edit distance (Levenshtein, 1966; Heeringa et al., 2006; Kondrak and Sherif, 2006). Levenshtein distance is defined as the smallest number of insertion, deletion and substitution operations required to transform one string into another.

(3)

	b	i	i	s	c	h	p	i	i	u
	b	i		s	c	h	p	i		l
	0	0	1	0	0	0	0	0	1	1

[4] Accented and unaccented characters are considered as different. See footnote 5.

Example (3) shows two words and the associated operation costs. There are two deletion operations and one substitution operation, hence Levenshtein distance between *biischpiiu* and *bischpil* is 3.[5]

Among other proposals, Heeringa et al. (2006) suggest normalizing Levenshtein distance by the length of the alignment. The underlying idea is that a Levenshtein distance of 2 for two long words does not mean the same as a Levenshtein distance of 2 for two very short words. In example (3), the length of the alignment is 10 (in this case, it is equal to the length of the longer word). Normalized Levenshtein distance is $\frac{3}{10} = 0.3$.

A cognate identification algorithm based on normalized Levenshtein distance requires a threshold such that only those word pairs whose distance is below the threshold are considered cognates. In order to identify sensible values for this threshold, we classified all word pairs of the corpus according to their distance. We evaluated nine thresholds between 0.05 and 0.4 to see if they effectively discriminate cognate pairs from non-cognate pairs. The evaluation was done on the basis of 100 randomly selected word pairs with a normalized Levenshtein distance lower or equal than the respective threshold.

In this evaluation, we distinguish between *form cognates* – words that represent the same inflected forms of the same lemma –, and *lemma cognates* – words that represent different inflected forms of the same lemma. Example (4) is a form cognate pair: it shows two dialectally different realizations of the singular form of the Standard German lemma *Gemeinde* 'municipality'. Example (5) is only a lemma cognate pair: one of the word contains the plural ending *-e*, while the other word is a singular form.

(4) gmeind — gmaind

(5) gmeind — gmainde

Table 1 shows the results of this evaluation. As the distance threshold increases, the proportion of cognates drops while the proportion of non-cognates rises. With thresholds higher than 0.25, the number of non-cognates surpasses the number of cognates. We therefore expect the cognate detection algorithm to work best below this threshold.

Let us conclude this section by some additional remarks about the evaluation:

- The distinction between form cognates and lemma cognates cannot be easily operationalized with an automatic approach. For instance, the correspondance *u* – *ü* may be a phonological one and distinguish two identical forms of different dialects. But it may also be a morphological correspondence that distinguishes singular from plural forms independently of the dialect. In the following experiments, we treat both types of cognate pairs in the same way.

- In practice, the reported figures are measures of precision. Recall may be estimated by the number of cognates situated above a given threshold. While we have not evaluated the entire distance interval, the given figures suggest that many true cognates are indeed found at high distance levels. This issue may be addressed by improving the string distance metric.

- Ambiguous words were not disambiguated according to the syntactic context and the dialect. As a result, all identical word pairs (threshold 0.00) are considered form cognates, although some of them may be false friends.

4.2 Identifying identical words

In common understanding, an *identical word pair* is a pair of words whose Levenshtein distance is 0. In some of the following experiments, we adopt this assumption.

However, we found it useful to relax this definition in order to avoid minor inconsistencies in the transcription and to neglect the smallest dialect differences. Therefore, we also carried out experiments where identical word pairs were defined as having a normalized Levenshtein distance of 0.10 or lower.

4.3 Experiments

Recall that we propose to measure the linguistic similarity of two texts by the ratio of identical word pairs among the cognate word pairs.

[5] Note that we treat all characters in the same way: replacing *o* by *k* yields the same cost as replacing it by *u* or by *ò*. This simple approach may not be the optimal solution when dealing with similar dialects. This issue will be addressed in future work.

Distance threshold	Word pairs	Form cognates	Lemma cognates	All cognates	Non-cognates	Non-words
0.00	5230	100%	0%	100%	0%	0%
0.05	5244	98%	0%	98%	0%	2%
0.10	6611	94%	4%	98%	1%	1%
0.15	10674	79%	16%	95%	4%	1%
0.20	18582	55%	16%	71%	29%	0%
0.25	27383	48%	13%	61%	38%	1%
0.30	36002	40%	12%	52%	47%	1%
0.35	49011	29%	10%	39%	61%	0%
0.40	65955	20%	13%	33%	67%	0%

Table 1: Manual evaluation of the cognate identification task. Percentages are based on a random sample of 100 word pairs with a normalized Levenshtein distance below or equal to the given threshold. Form cognate and lemma cognate counts are summed up in the 'All cognates' column. The interviewees sometimes made false starts and stopped in the middle of the word; these incomplete words, together with obvious typing errors in the transcription, are counted in the last column.

Cognate pairs as well as identical word pairs are characterized by different thresholds of normalized Levenshtein distance. We experiment with thresholds of 0.20, 0.25, 0.30, 0.35 and 0.40 for cognate word pairs, and with thresholds of 0 and 0.10 for identical word pairs.

4.4 Normalization by text length

A major issue of using comparable corpora is the large variation in text length and vocabulary use. This has to be accounted for in our experiments. First, all counts refer to types of word pairs, not tokens. We argue that the frequency of a word in a given text depends too much on the content of the text and is not truly representative of its dialect. Second, if few identical words are found, this does not necessarily mean that the two texts are dialectally distant, but may also be because one text is much shorter than the other. Hence, the proportion of identical words is normalized by the number of cognate words contained in the shorter of the two texts.

5 Evaluation and visualisation

By computing the linguistic distance for all pairs of texts in our corpus, we obtain a two-dimensional distance matrix. Recent dialectometric tradition provides several techniques to evaluate and visualize the data encoded in this matrix.

First, one can measure how well the linguistic distances correlate with geographic distances (Section 5.1). Second, one can group the texts into maximally homogeneous clusters (Section 5.2). Third, one can plot the texts as data points on a two-dimensional graph and visually compare this graph with the geographical locations of the texts (Section 5.3).

5.1 Numerical measures of spatial autocorrelation

A general postulate of spatial analysis is that "on average, values at points close together in space are more likely to be similar than points further apart" (Burrough and McDonnell, 1998, 100). This idea that the distance of attribute values correlates with their geographical distance is known as *spatial autocorrelation*. The same idea has been coined the *fundamental dialectological postulate* by Nerbonne and Kleiweg (2005, 10): "Geographically proximate varieties tend to be more similar than distant ones."

Here, we use this postulate to evaluate the different threshold combinations of our dialect similarity measure: the higher a threshold combination correlates with geographic distance (i.e., places of residence of the interviewees), the better it is able to discriminate the dialects. Here, the results obtained with two correlation measures are reported.

Local incoherence has been proposed by Nerbonne and Kleiweg (2005). The idea of this measure is that the correlation between linguistic and geographic distances is local and does not need to hold over larger geographical distances. In practice, for every data point, the 8 linguistically most

similar points[6] are inspected according to their linguistic distance value. Then, the geographic distance of these pairs of points is measured and summed up. This means that high incoherence values represent poor measurements, while lower values stand for better results.

The **Mantel-Test** (Sokal and Rohlf, 1995, 813-819) is a general statistical test which applies to data expressed as dissimilarities. It is often used in evolutionary biology and ecology, for example, to correlate genetic distances of animal populations with the geographic distances of their range. The Mantel coefficient Z is computed by computing the Hadamard product of the two matrices. The statistical significance of this coefficient is obtained by a randomization test. A sample of permutations is created, whereby the elements of one matrix are randomly rearranged. The correlation level depends on the proportion of samples whose Z-value is higher than the Z-value of the reference matrix. All experiments were carried out with a sample size of 999 permutations, which corresponds to a simulated p-value of 0.001.

Table 2 shows the results of both correlation measures for all experiments. These results are in line with the manual evaluation of Table 1. At first, increasing the cognate pair threshold leads to more data, and in consequence, to better results. Above 0.35 however, the added data is essentially noise (i.e., non-cognate pairs), and the results drop again.

According to local incoherence, the best threshold combination is $\langle 0.10, 0.35 \rangle$. In terms of Mantel test correlation, the $\langle 0.10, 0.25 \rangle$ threshold performs slightly better. Adopting an identical pair threshold of 0.00 results in slightly inferior correlations.

Distance thresholds		Local	Mantel Test	
Identical	Cognate	inc.	r	p
0.00	0.20	0.59	0.56	0.001
	0.25	0.47	0.68	0.001
	0.30	0.49	0.66	0.001
	0.35	**0.41**	**0.70**	0.001
	0.40	0.46	0.65	0.001
0.10	0.20	0.55	0.65	0.001
	0.25	0.41	**0.73**	0.001
	0.30	0.43	0.70	0.001
	0.35	**0.37**	0.72	0.001
	0.40	0.43	0.67	0.001

Table 2: Correlation values for the different experiments. The first and second columns define each experiment in terms of two Levenshtein distance thresholds. For local incoherence, lower values are better. For the Mantel test figures, we report the correlation coefficient r as well as the significance level p.

5.2 Clustering

The distance matrix can also be used as input to a clustering algorithm. Clustering has become one of the major data analysis techniques in dialectometry (Mucha and Haimerl, 2005), but has also been used with plain text data in order to improve information retrieval (Yoo and Hu, 2006).

Hierarchical clustering results in a *dendrogram* which represents the distances between every two data points as a tree. However, clustering is known to be unreliable: small changes in the distance matrix may result in completely different dendrograms. To counter this issue, *noisy clustering* has been proposed (Nerbonne et al., 2008): clustering is repeated 100 times, and at each run, random amounts of noise are added to the different cells of the distance matrix. This gives an indication of the reliability of the resulting clusters. Figure 2 shows a dendrogram obtained with noisy clustering. We used both group average and weighted average clustering algorithms, and a noise level of 0.2.[7] Figure 3 localizes the data points on a geographical map. All clusters show a reliability score of 92% or above.

Clustering allows us to recover certain characteristics of the Swiss German dialect landscape. First, texts from the same canton (whose IDs contain the same two-letter abreviation) are grouped together with high reliability. Second, the dendrogram shows – albeit with lower reliability scores – a three-fold East-West stratification with blue regions in the West (BE), green regions in Central Switzerland (AG, LU) and yellow areas in the East (ZH, SZ, GL). The border between Western and Central dialects roughly corresponds to the so-called Brünig-Napf line. The border between Central and Eastern varieties is also confirmed by former dialectological research (Haas, 1982; Hotzenköcherle, 1984). Third, three dialects are

[6] The restriction to 8 points is the key of the local component of this measure. The exact value of this parameter has been determined empirically by the authors of the measure.

[7] These are the default settings of the Gabmap program (Nerbonne et al., 2011).

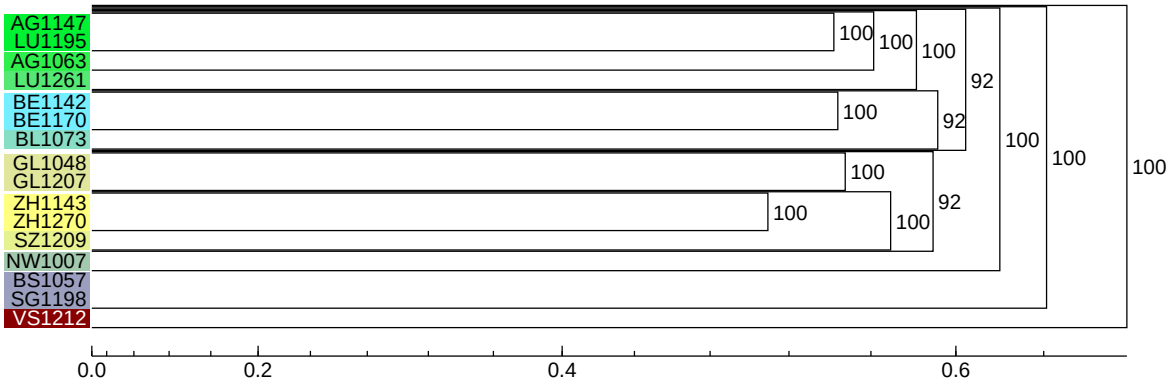

Figure 2: Dendrogram obtained with a threshold setting of $\langle 0.10, 0.35 \rangle$. The scale at the bottom of the graphics represents the distance of the clusters, while the numbers on the vertical lines represent the reliability of the clusters (i.e. in how many of the 100 runs a cluster has been found).

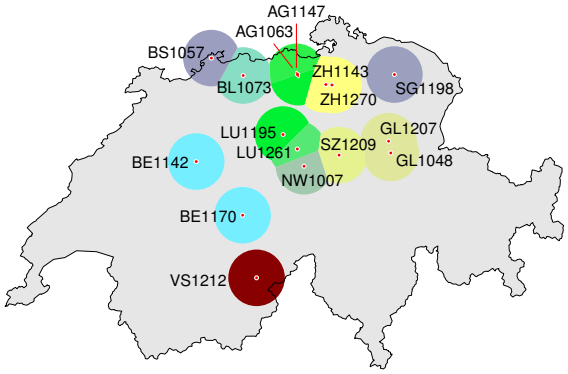

Figure 3: Geographic localization of the Archimob texts, according to the place of residence of the interviewed persons. The colors represent the linguistic distance between texts; they correspond to the colors used in the dendrogram of Figure 2.

clearly considered as outliers: the Northwestern dialect of Basel (BS1057), the Northeastern dialect of St. Gallen (SG1198), and most of all the Southwestern Wallis dialect (VS1212). Again, these observations are in line with common dialectological knowledge.

5.3 Multidimensional scaling

The Swiss German dialect landscape has been known to feature major East-West divisions (see above) as well as several levels of stratification on the North-South axis. Our hypothesis is that the linguistic distances represented in the distance matrix should be able to recover this mainly two-dimensional organization of Swiss German dialects. Since the distance matrix defines a multi-dimensional space in which all data points (texts) are placed, this space has to be reduced to two dimensions. For this purpose, we use multidimensional scaling. If the linguistic distances are correctly defined and the multidimensional scaling algorithm truly extracts the two main dimensions of variation, the resulting two-dimensional graph should be comparable with a geographic map.

Figure 4 shows the resulting graph for one experiment. Figures 5 and 6 show the values of each data point in grey levels for the two first dimensions obtained by multi-dimensional scaling.

One observes that the localization of data points in Figure 4 closely corresponds to their geographic location (as illustrated in Figure 3): the major North-South divisions as well as some East-West divisions are clearly recovered.

More surprisingly, the two main dimensions of multidimensional scaling correspond to diagonals in geographic terms. The first dimension (Figure 5) allows to distinguish Northwestern from Southeastern variants, while the second dimension (Figure 6) distinguishes Northeastern from Southwestern variants. Instead of +-shaped dialect divisions put forward by traditional dialectology, our approach rather finds X-shaped dialect divisions.

6 Discussion and future work

We have proposed a simple measure that approximates the linguistic distance between two texts according to the ratio of identical words among the cognate word pairs. The definitions of *identical word pair* and *cognate word pair* are operationalized with fixed thresholds of normalized

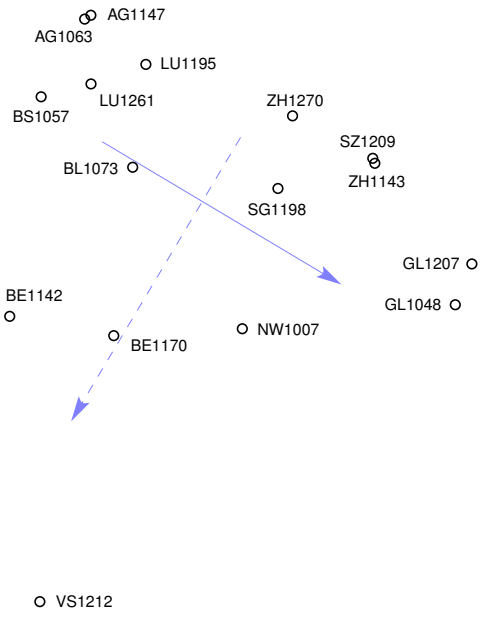

Figure 4: Plot representing the first two dimensions of multi-dimensional scaling applied to the experiment with $\langle 0.10, 0.35 \rangle$ thresholds.

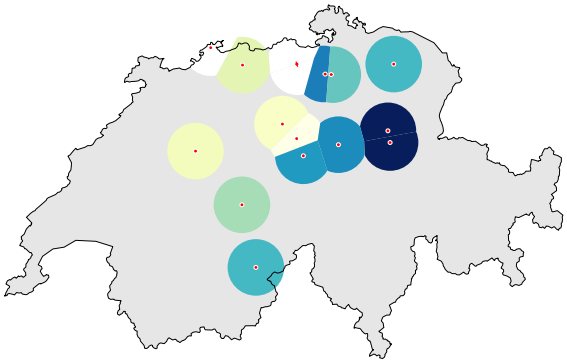

Figure 5: Map representing the first dimension of multi-dimensional scaling (same experiment as Fig. 4).

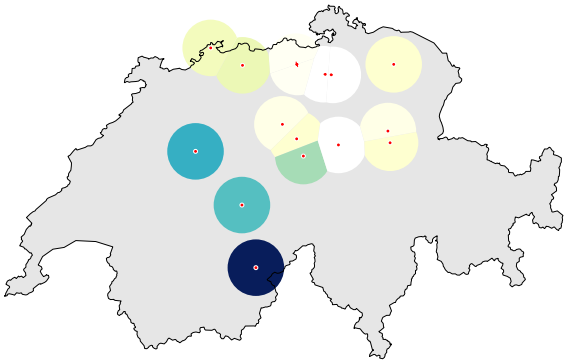

Figure 6: Map representing the second dimension of multi-dimensional scaling (same experiment as Fig. 4).

Levenshtein distance. The resulting distance matrix has been analyzed with correlation measures, and visualized with clustering and multidimensional scaling techniques. The visualizations represent the main characteristics of the Swiss German dialect landscape in a surprisingly faithful way.

The close relation obtained among texts from the same canton may suggest that the distance measure is biased towards proper nouns. For example, two Zurich German texts are more likely to use toponyms from the Zurich region than a Bernese German text. If there are many of these (likely identically pronounced) toponyms, the similarity value will increase. However, manual inspection of the relevant texts did not show such an effect. Region-specific toponyms are rare.

The results suggest that a more fine-grained variant of Levenshtein distance might be useful. In the following paragraphs, we present several improvements for future work.

The results suggest that a more fine-grained variant of Levenshtein distance might improve the precision and recall of the cognate detection algorithm. Notably, it has been found that vowels change more readily than consonant in closely related language varieties. In consequence, changing one vowel by another should be penalized less than changing a vowel by a consonant (Mann and Yarowsky, 2001). The same holds for accented vs. non-accented characters. Complex graphemes representing a single phoneme appear rather frequently in the Dieth transcription system (e.g. for long vowels) and should also be treated separately.

We should also mention that the proposed method likely faces a problem of scale. Indeed, each word of each text has to be compared with each word of each text. This is only manageable with a small corpus like ours.

We conclude by pointing out a limitation of this approach: the automatic alignment process based on the concept of cognate pairs obviously only works for phonetically related word pairs. This contrasts with other dialectometric approaches based on lexical differences, in whose data sets different lemmas have been aligned. Future work on the Archimob corpus shall add normalization and lemmatization layers. This information could be useful to improve word alignment beyond cognate pairs.

Acknowledgments

The author wishes to thank Prof. Elvira Glaser, Alexandra Bünzli, Anne Göhring and Agnes Kolmer (University of Zurich) for granting access to the Archimob corpus and giving detailed information about its constitution. Furthermore, the anonymous reviewers are thanked for most helpful remarks.

References

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.

Peter A. Burrough and Rachael A. McDonnell. 1998. *Principles of Geographical Information Systems*. Oxford University Press, Oxford.

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

Matthias Friedli. 2006. Der Komparativanschluss im Schweizerdeutschen – ein raumbildendes Phänomen. In Hubert Klausmann, editor, *Raumstrukturen im Alemannischen*, pages 103–108. Neugebauer, Graz/Feldkirch.

Walter Haas, 1982. *Die deutschsprachige Schweiz*, pages 71–160. Benziger, Zürich.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of string distance algorithms for dialectology. In *Proceedings of the ACL 2006 Workshop on Linguistic Distances*, pages 51–62, Sydney, Australia.

Rudolf Hotzenköcherle. 1984. *Die Sprachlandschaften der deutschen Schweiz*. Sauerländer, Aarau.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16, Philadelphia, PA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 demonstration session*, Prague, Czech Republic.

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

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

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

Tony McEnery and Richard Xiao. 2008. Parallel and comparable corpora: What is happening? In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: The Linguist and the Translator*, chapter 2, pages 18–31. Multilingual Matters, Clevedon.

Hans-Joachim Mucha and Edgar Haimerl. 2005. Automatic validation of hierarchical cluster analysis with application in dialectometry. In C. Weihs and W. Gaul, editors, *Classification – the Ubiquitous Challenge*, pages 513–520. Springer, Berlin.

John Nerbonne and Peter Kleiweg. 2005. Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, 5.

John Nerbonne, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker, editors, *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, pages 647–654. Springer, Berlin.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia, Special Issue*, II:65–89.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*, pages 67–81, Montréal, Canada.

Robert R. Sokal and F. James Rohlf. 1995. *Biometry: the principles and practice of statistics in biological research*. W.H. Freeman, New York, 3rd edition.

Janine Steiner. 2006. Syntaktische Variation in der Nominalphrase – ein Fall für die Dialektgeographin oder den Soziolinguisten? In Hubert Klausmann, editor, *Raumstrukturen im Alemannischen*, pages 109–115. Neugebauer, Graz/Feldkirch.

Illhoi Yoo and Xiaohua Hu. 2006. A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 220–229, Chapel Hill, NC, USA.