Machine translation into multiple dialects: The example of Swiss German

SCHERRER, Yves

Abstract

In this paper, we propose to approach dialects and dialectology from a Natural Language Processing (NLP) point of view. NLP covers a series of computational applications that analyze and/or transform linguistic data, such as machine translation, parsing, or text summarization. For practical reasons, most NLP applications focus on standardized, written language varieties. We argue that non-standard varieties, often also characterized by internal variation, can result in interesting methodological insights in NLP. Our work focuses on Swiss German dialects. Today, dialect represents the default variety of oral communication in the German-speaking part of Switzerland (Standard German is almost exclusively used for writing). Recently, dialect writing has also become popular in electronic media (Siebenhaar 2003). This evolution justifies the development of dialect NLP tools, and at the same time provides us with data to validate them. We present a system that automatically translates (written) Standard German sentences into (written) sentences of any Swiss German dialect. It is based on hand-built transfer rules operating on [...]

Reference


Available at:
http://archive-ouverte.unige.ch/unige:22818

Disclaimer: layout of this document may differ from the published version.
Machine translation into multiple dialects: The example of Swiss German

Yves Scherrer
University of Geneva, Switzerland

SIDG 2012, Vienna
Outline

1 Introduction

2 Digitization of atlas maps

3 Transformation rules

4 Evaluation

5 Conclusion
Computational linguistics and dialectology

Opportunities for dialectology:

- Computational approaches may present the results of existing dialectological research in a more intuitive and dynamic way, and make them better accessible to laymen.
- Computational approaches (e.g., dialectometry) may give new insights into dialectal variation patterns.

Challenges for computational linguistics:

- Dialect data requires an explicit handling of different types of variation.
- Dialect data may challenge current methodological beliefs:
  - More scientifically edited data than (easily accessible) raw text corpora
  - Traditional rule-based approaches are probably more adapted than recent machine-learning approaches.
Computational linguistics and dialectology

Opportunities for dialectology:

- Computational approaches may present the results of existing dialectological research in a more intuitive and dynamic way, and make them better accessible to laymen.
- Computational approaches (e.g., dialectometry) may give new insights into dialectal variation patterns.

Challenges for computational linguistics:

- Dialect data requires an explicit handling of different types of variation.
- Dialect data may challenge current methodological beliefs:
  - More scientifically edited data than (easily accessible) raw text corpora
  - Traditional rule-based approaches are probably more adapted than recent machine-learning approaches.
Swiss German dialects:

- Spoken by about 2/3 of the Swiss population
  - Other major languages: French and Italian
- Diglossia with Standard German
  - Standard German for written use, dialect for spoken use

Variation between Standard German and Swiss German:
- Phonology, morphology, syntax, lexicon

Variation among the different Swiss German dialects:
- Phonetics/phonology, morphology, syntax, lexicon
Machine translation

Goal:
Translate Standard German text into the different Swiss German dialects.

- Morpho-syntactic analysis of the source text with existing tools.
- Transformation rules created on the basis of dialectological research.

These transformation rules have different outcomes depending on the selected target dialect.

- They must be aware of their application area: georeferenced rules.
- The relevant geographic data are extracted from dialectological atlases.
Georeferenced transformation rules

„Generative dialectology“ (Veith 1982):

- Select a reference system $B$
- Derive dialect systems $D_n$ from the reference system by transformation rules:
  - $#\text{Töpfer}_{B} \rightarrow #\text{Häfner}_{D_{3333} - 46999}$

Our work:

- Our reference system is Standard German. This is etymologically not accurate, but practical.
- Rules are not assigned a finite set of discrete dialects, but a probability map of their occurrence:
  - $\text{immer } [\text{StdG}] \rightarrow \text{geng}$

Georeferenced transformation rules

„Generative dialectology“ (Veith 1982):

- Select a reference system $B$
- Derive dialect systems $D_n$ from the reference system by transformation rules:
  - #$Töpfer#_B \rightarrow #Häfner#_{D_{33333-46999}}$

Our work:

- Our reference system is Standard German. This is etymologically not accurate, but practical.
- Rules are not assigned a finite set of discrete dialects, but a probability map of their occurrence:
  - immer [StdG] $\rightarrow$ geng

Georeferenced transformation rules

1. Where do we get these probability maps from?
2. How are these rules implemented?
3. How can we measure the performances of the resulting system?
1. Introduction

2. Digitization of atlas maps

3. Transformation rules

4. Evaluation

5. Conclusion
Data sources

The transformation rules should cover the most important phonetic, lexical, morphological and syntactic changes for the entire range of Swiss German dialects.

Two linguistic atlases contain such data:

1. **Sprachatlas der deutschen Schweiz (SDS):**
   - Data collection 1939-1958 (interviews)
   - Phonetics/phonology, morphology, lexicon
   - 8 volumes, 1500 hand-drawn maps

2. **Syntaktischer Atlas der deutschen Schweiz (SADS):**
   - Data collection 2000-2009 (written questionnaires)
   - Syntax
   - 120 questions
   - Not yet published, but access to database
Data sources

**Work flow:**

1. Scan map (SDS only)
2. Digitize map
3. Create surface maps by interpolation

**Digitized SDS data:**
- 60 (of 400) phonetic maps
- 110 (of 250) morphological maps
- 30 (of 800) lexical maps

**Digitized SADS data:**
- 14 syntactic constructions

1 map ~ 1 transformation rule

Yves Scherrer: Machine translation into multiple dialects – The example of Swiss German
Data sources

Work flow:

1. Scan map (SDS only)
2. Digitize map
3. Create surface maps by interpolation

Digitized SDS data:
- 60 (of 400) phonetic maps
- 110 (of 250) morphological maps
- 30 (of 800) lexical maps

Digitized SADS data:
- 14 syntactic constructions

1 map ~ 1 transformation rule
Data sources

Work flow:

1. Scan map (SDS only)
2. Digitize map
3. Create surface maps by interpolation

Digitized SDS data:
- 60 (of 400) phonetic maps
- 110 (of 250) morphological maps
- 30 (of 800) lexical maps

Digitized SADS data:
- 14 syntactic constructions

1 map ~ 1 transformation rule
Original (scanned) SDS map
Digitization

Legend of map II/120:

- 4 major variants
- 2 minor variants of η and n.

Hypotheses:

- Most/all words with final -nd show the same geographical distribution of variants
- The two minor variants are not reliably distinguished in dialect texts → merge them with major variants
- Variant distributions at other locations can be derived from the available data: interpolation
### Interpolation

**Method:** Kernel density estimation (Rumpf et al. 2009)

**Normalization:** The gray scale values on the map should be interpretable as probabilities. Therefore, the probabilities of all variants must sum up to 1 at each point.

---

Outline

1. Introduction
2. Digitization of atlas maps
3. Transformation rules
4. Evaluation
5. Conclusion
A phonetic transformation rule

Rules are written in the XFST formalism (Beesley & Karttunen 2003).

\[
\text{define } \text{ndVoc} [ \\
\text{n d } \rightarrow \text{ [ n d "@U.2-119.nd@" | n g "@U.2-119.ng@" | n n "@U.2-119.nn@" | n "@U.2-119.n@" ]} \\
|| \text{Vowel \_ Vowel }] ;
\]

- This rule is called \text{ndVoc}.
- It transforms \text{nd} into one of the four variants \text{nd}, \text{ng}, \text{nn}, \text{n}:
  - It only applies to intervocalic \text{nd}.
    - \text{gestanden, stood’} \rightarrow \text{gschtande, gschtange, gschtanne, gschtane}
- "@U.2-119.nd@" is a special symbol (a \textbf{flag diacritic}) that specifies the name of the corresponding probability map.

A lexical transformation rule

```ast
define immer [
    {immer} -> [
        {immer} "@U.6-026-immer.immer@" |
        {gäng} "@U.6-026-immer.gäng@" |
        {geng} "@U.6-026-immer.geng@" |
        {gi} "@U.6-026-immer.gi@" |
        {ging} "@U.6-026-immer.ging@" |
        {alewil} "@U.6-026-immer.alewil@" |
        {all} "@U.6-026-immer.all@" |
        {albig} "@U.6-026-immer.albig@" |
        {eisder} "@U.6-026-immer.eisder@" ]
];
```

- This rule is called *immer* and transforms the Standard German word *immer* 'always' into one of 9 variants.
- Entire words are defined in curly braces.
A morphological transformation rule

```
define adj-2-flex [
    ADJA [Nom | Acc] Sg Gender Degree Weak -> [
        0 "@U.3-254-adj-F-Sg.0@" |
        i "@U.3-254-adj-F-Sg.i@"
    ]
];
```

- This rule is called \textit{adj-2-flex} and adds an inflectional ending to weak singular adjectives.
- The input words are morphosyntactically tagged. These feature tags are then replaced by the actual dialectal affixes.
- Two variants: no ending (0) and -i:
  - \texttt{die schwarze Katze, the black cat'} $\rightarrow$ \texttt{di schwarz/schwarz\text{\textemdash}i Chatz}

---

Yves Scherrer: Machine translation into multiple dialects – The example of Swiss German
A syntactic transformation rule

Syntactic rules are not implemented in XFST, but with specific Python scripts.

Syntactic rules require parsed input (dependency relations).

This rule potentially changes the word order in a verb cluster: It applies whenever a past participle is followed by a finite auxiliary verb, and the former is syntactically dependent on the latter.

VVPP[head=VAFIN] VAFIN

\[\rightarrow \text{VVPP VAFIN } @1.19.1@,\]
\[\rightarrow \text{VAFIN VVPP } @1.19.2@\]
Outline

1 Introduction
2 Digitization of atlas maps
3 Transformation rules
4 Evaluation
5 Conclusion
To evaluate the performances of the transformation rules, we need:

1. a parallel, multi-dialectal reference corpus
   - **parallel**: Standard German text with dialect translations
   - **multi-dialectal**: sub-corpora for different Swiss German dialects

2. an evaluation measure.

**Reference corpus:**

- Texts from the Swiss German Wikipedia
  - „Clean“ material (entire sentences, few typos, …)
  - Covers several dialects
  - Many articles are annotated with their dialect
- Five dialects: Basel, Bern, Ostschweiz, Wallis, Zürich
- 100 randomly selected sentences for each dialect
- Each dialect sentence was translated back to Standard German

http://als.wikipedia.org/
Evaluation

To evaluate the performances of the transformation rules, we need:

1. a parallel, multi-dialectal reference corpus
   - parallel: Standard German text with dialect translations
   - multi-dialectal: sub-corpora for different Swiss German dialects

2. an evaluation measure.

Reference corpus:

- Texts from the Swiss German Wikipedia
  - „Clean“ material (entire sentences, few typos, …)
  - Covers several dialects
  - Many articles are annotated with their dialect
- Five dialects: Basel, Bern, Ostschweiz, Wallis, Zürich
- 100 randomly selected sentences for each dialect
- Each dialect sentence was translated back to Standard German

http://als.wikipedia.org/
Evaluation

Evaluation measure:

- Common evaluation measures for Machine Translation (BLEU, TER, ...) compare the word overlap between the system output and the reference translation.
  - As soon as a proposed word differs by one letter from the reference, it is considered wrong.
  - Our approach of using phonetic transformations would be heavily penalized by such an approach: A single transformation error is weighted just as heavily as several errors in the same word.

- **Longest Common Subsequence Ratio (LCSR):** count the proportion of identical letters (not words) between the system output and the reference translation.

Experiment:

1. Automatically translate a Standard German text into all five dialects.
2. Compare the resulting texts with the reference translation (written in one of the five dialects).
Evaluation

Evaluation measure:

- Common evaluation measures for Machine Translation (BLEU, TER, …) compare the word overlap between the system output and the reference translation.
  - As soon as a proposed word differs by one letter from the reference, it is considered wrong.
  - Our approach of using phonetic transformations would be heavily penalized by such an approach: A single transformation error is weighted just as heavily as several errors in the same word.

- Longest Common Subsequence Ratio (LCSR): count the proportion of identical letters (not words) between the system output and the reference translation.

Experiment:

1. Automatically translate a Standard German text into all five dialects.
2. Compare the resulting texts with the reference translation (written in one of the five dialects).
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>BA</th>
<th>BE</th>
<th>OS</th>
<th>WS</th>
<th>ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>80.60%</td>
<td>82.77%</td>
<td>80.68%</td>
<td>77.61%</td>
<td>82.53%</td>
</tr>
<tr>
<td>BA</td>
<td>80.94%</td>
<td>81.21%</td>
<td>79.78%</td>
<td>75.88%</td>
<td>81.33%</td>
</tr>
<tr>
<td>BE</td>
<td>76.82%</td>
<td>82.56%</td>
<td>78.89%</td>
<td>72.50%</td>
<td>77.97%</td>
</tr>
<tr>
<td>OS</td>
<td>77.40%</td>
<td>82.29%</td>
<td><strong>81.23%</strong></td>
<td>72.78%</td>
<td>78.97%</td>
</tr>
<tr>
<td>WS</td>
<td>76.57%</td>
<td>80.88%</td>
<td>77.14%</td>
<td>75.82%</td>
<td>78.16%</td>
</tr>
<tr>
<td>ZH</td>
<td>80.71%</td>
<td>82.43%</td>
<td><strong>81.26%</strong></td>
<td><strong>76.54%</strong></td>
<td><strong>83.30%</strong></td>
</tr>
</tbody>
</table>

1. Is the dialectal system output more similar to the reference (of the same dialect) than the Standard German system input?
   - True for three out of five tested dialects.

2. Is the system output for the original dialect more similar to the reference than the system output for other dialects?
   - True for three out of five tested dialects.
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BA</td>
</tr>
<tr>
<td>Standard</td>
<td>80.60%</td>
</tr>
<tr>
<td>BA</td>
<td><strong>80.94%</strong></td>
</tr>
<tr>
<td>BE</td>
<td>76.82%</td>
</tr>
<tr>
<td>OS</td>
<td>77.40%</td>
</tr>
<tr>
<td>WS</td>
<td>76.57%</td>
</tr>
<tr>
<td>ZH</td>
<td>80.71%</td>
</tr>
</tbody>
</table>

1. Is the dialectal system output more similar to the reference (of the same dialect) than the Standard German system input?
   - True for three out of five tested dialects.

2. Is the system output for the original dialect more similar to the reference than the system output for other dialects?
   - True for three out of five tested dialects.
# Results

<table>
<thead>
<tr>
<th>System</th>
<th>BA</th>
<th>BE</th>
<th>OS</th>
<th>WS</th>
<th>ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>80.60%</td>
<td><strong>82.77%</strong></td>
<td>80.68%</td>
<td><strong>77.61%</strong></td>
<td>82.53%</td>
</tr>
<tr>
<td>BA</td>
<td><strong>80.94%</strong></td>
<td>81.21%</td>
<td>79.78%</td>
<td>75.88%</td>
<td>81.33%</td>
</tr>
<tr>
<td>BE</td>
<td>76.82%</td>
<td><strong>82.56%</strong></td>
<td>78.89%</td>
<td>72.50%</td>
<td>77.97%</td>
</tr>
<tr>
<td>OS</td>
<td>77.40%</td>
<td>82.29%</td>
<td><strong>81.23%</strong></td>
<td>72.78%</td>
<td>78.97%</td>
</tr>
<tr>
<td>WS</td>
<td>76.57%</td>
<td>80.88%</td>
<td>77.14%</td>
<td>75.82%</td>
<td>78.16%</td>
</tr>
<tr>
<td>ZH</td>
<td>80.71%</td>
<td>82.43%</td>
<td><strong>81.26%</strong></td>
<td><strong>76.54%</strong></td>
<td><strong>83.30%</strong></td>
</tr>
</tbody>
</table>

1. Is the dialectal system output more similar to the reference (of the same dialect) than the Standard German system input?
   - True for three out of five tested dialects.

2. Is the system output for the original dialect more similar to the reference than the system output for other dialects?
   - True for three out of five tested dialects.
Discussion

These results are promising, but they also show that the transformation rules are not yet robust enough. Several issues have been identified:

- Syntactic rules are not fully integrated yet.
- The morphosyntactic analysis of the Standard German source text is error-prone.
- Many dialect authors do not reflect all phonetic changes in their spelling, so that they produce less "dialectal looking" text than our system.
- The dialect annotation of the Wikipedia texts is not precise enough. *Bärndütsch* refers to a large, heterogeneous dialect area.
- Some dialectal variants have disappeared since the SDS data collection in the 1940s.
Outline

1. Introduction
2. Digitization of atlas maps
3. Transformation rules
4. Evaluation
5. Conclusion
Conclusion

Machine translation into multiple dialects:

- Georeferenced transformation rules, inspired by generative dialectology.
- Manually created rules in order to advantage of earlier dialectological fieldwork (SDS and SADS atlases).

Evaluation:

- A multi-dialectal parallel corpus based on Wikipedia texts.
- An evaluation measure that accounts for character-level differences.
- Two hypotheses that have been partially verified.

Future work:

- Integrate syntactic rules and use better source analysis tools.
- Dialectometric analyses with the newly digitized data sets.

Demo: http://latlntic.unige.ch/~scherrey/
Conclusion

Machine translation into multiple dialects:

- Georeferenced transformation rules, inspired by generative dialectology.
- Manually created rules in order to advantage of earlier dialectological fieldwork (SDS and SADS atlases).

Evaluation:

- A multi-dialectal parallel corpus based on Wikipedia texts.
- An evaluation measure that accounts for character-level differences.
- Two hypotheses that have been partially verified.

Future work:

- Integrate syntactic rules and use better source analysis tools.
- Dialectometric analyses with the newly digitized data sets.

Demo: http://latlntic.unige.ch/~scherrey/
Conclusion

Machine translation into multiple dialects:

- Georeferenced transformation rules, inspired by generative dialectology.
- Manually created rules in order to advantage of earlier dialectological fieldwork (SDS and SADS atlases).

Evaluation:

- A multi-dialectal parallel corpus based on Wikipedia texts.
- An evaluation measure that accounts for character-level differences.
- Two hypotheses that have been partially verified.

Future work:

- Integrate syntactic rules and use better source analysis tools.
- Dialectometric analyses with the newly digitized data sets.

Demo: http://latIntic.unige.ch/~scherrey/
Conclusion

Machine translation into multiple dialects:

- Georeferenced transformation rules, inspired by generative dialectology.
- Manually created rules in order to advantage of earlier dialectological fieldwork (SDS and SADS atlases).

Evaluation:

- A multi-dialectal parallel corpus based on Wikipedia texts.
- An evaluation measure that accounts for character-level differences.
- Two hypotheses that have been partially verified.

Future work:

- Integrate syntactic rules and use better source analysis tools.
- Dialectometric analyses with the newly digitized data sets.

Demo: [http://latIntic.unige.ch/~scherrey/](http://latIntic.unige.ch/~scherrey/)
Dialectometry

The scanned SDS maps can be reused for dialectometric studies.

Hierarchical clustering is a popular method of data analysis in dialectometry, but it has several drawbacks:

- The number of target clusters (groups) has to be defined arbitrarily.
- No general agreement about the appropriateness of the different clustering algorithms.
- It is unstable: small differences in the data may yield completely different results.

A possible answer to these criticisms is noisy clustering (Nerbonne et al. 2008):

- Repeat clustering 100 times, alternating between two clustering algorithms (Weighted Average and Group Average).
- Each time, add random amounts of noise to the distance values.
- Define color value with multidimensional scaling.
Dialectometry

The scanned SDS maps can be reused for dialectometric studies. **Hierarchical clustering** is a popular method of data analysis in dialectometry, but it has several drawbacks:

- The number of target clusters (groups) has to be defined arbitrarily.
- No general agreement about the appropriateness of the different clustering algorithms.
- It is unstable: small differences in the data may yield completely different results.

A possible answer to these criticisms is **noisy clustering** (Nerbonne et al. 2008):

- Repeat clustering 100 times, alternating between two clustering algorithms (Weighted Average and Group Average).
- Each time, add random amounts of noise to the distance values.
- Define color value with multidimensional scaling.
Dialectometry

The scanned SDS maps can be reused for dialectometric studies.

Hierarchical clustering is a popular method of data analysis in dialectometry, but it has several drawbacks:

- The number of target clusters (groups) has to be defined arbitrarily.
- No general agreement about the appropriateness of the different clustering algorithms.
- It is unstable: small differences in the data may yield completely different results.

A possible answer to these criticisms is noisy clustering (Nerbonne et al. 2008):

- Repeat clustering 100 times, alternating between two clustering algorithms (Weighted Average and Group Average).
- Each time, add random amounts of noise to the distance values.
- Define color value with multidimensional scaling.
Noisy clustering with SDS data

(196 features: phonetics, morphology, lexicon)
Noisy clustering with SDS data

(196 features: phonetics, morphology, lexicon)
Noisy clustering with SADS data

(118 features: syntax)
Noisy clustering with SADS data

(118 features: syntax)