The distribution of aggregated syntactic construction types compared with other linguistic levels - A dialectometrical analysis of Swiss German dialects

KELLERHALS, Sandra, et al.

Abstract

Although there is a good availability of Swiss German dialect data, very few works have looked at them from an aggregated perspective (e.g. Kelle 2001) and none has considered syntactic data in comparison with data from other linguistic levels. As an interdisciplinary working group, composed of linguists and geographers, we apply dialectometrical methods to two different data sets of Swiss German dialects. The first is a digitized subset of the Swiss German Dialect Atlas (Sprachatlas der deutschen Schweiz SDS), which consists of phonetical, morphological and lexical data. The second originates from the Syntactic Atlas of Swiss German Dialects (SADS), collected in the early 2000s by the University of Zurich. Our contribution focuses on the following research questions: 1) How do the areal distributions of aggregated phonetical, morphological, lexical and syntactical phenomena of German-speaking Switzerland differ? 2) Which linguistic levels mainly determine the dialectal landscape of German-speaking Switzerland? Are there individual phenomena that correlate particularly well with the aggregated maps? In the presentation, […]

Reference

KELLERHALS, Sandra, et al. The distribution of aggregated syntactic construction types compared with other linguistic levels - A dialectometrical analysis of Swiss German dialects. In: Methods in Dialectology XV, Groningen (Netherlands), 7-12 Aug, 2014

Available at: http://archive-ouverte.unige.ch/unige:39956

Disclaimer: layout of this document may differ from the published version.
The distribution of aggregated syntactic construction types compared with other linguistic levels

A dialectometrical analysis of Swiss German dialects

Sandra Kellerhals¹, Yves Scherrer², Elvira Glaser³, Robert Weibel¹

¹Department of Geography, University of Zurich
²Department of Linguistics / Computer Science Center, University of Geneva
³Department of German Studies, University of Zurich

Methods in Dialectology XV, Groningen
Swiss German dialects:

- Two large-scale atlas projects, but few dialectometrical studies:
  
  **SDS** Sprachatlas der deutschen Schweiz  
  (1939–1958, Phonology, Morphology, Lexicon)  
  - Kelle (2001)  
  - Scherrer (PhD 2012); Goebl, Scherrer & Smečka (2013)

  **SADS** Syntaktischer Atlas der deutschen Schweiz  
  - Sibler et al. (2012); Jeszenszky et al. (here); Stöckle et al. (here)

This work:

- Comparison of SDS and SADS data with methods from dialectometry
**Swiss German dialects:**

- Two large-scale atlas projects, but few dialectometrical studies:
  - **SDS** Sprachatlas der deutschen Schweiz
    - (1939–1958, Phonology, Morphology, Lexicon)
    - *Kelle* (2001)
    - *Scherrerr* (PhD 2012); *Goebel, Scherrer & Smečka* (2013)
  - **SADS** Syntaktischer Atlas der deutschen Schweiz
    - *Sibler et al.* (2012); *Jeszczyszky et al.* (here); *Stöckle et al.* (here)

**This work:**

- Comparison of SDS and SADS data with methods from dialectometry
Swiss German dialects:

- Two large-scale atlas projects, but few dialectometrical studies:
  - **SDS** Sprachatlas der deutschen Schweiz
    (1939–1958, Phonology, Morphology, Lexicon)
    - Kelle (2001)
    - Scherrer (PhD 2012); Goebel, Scherrer & Smečka (2013)
  - **SADS** Syntaktischer Atlas der deutschen Schweiz
    - Sibler et al. (2012); Jeszenszky et al. (here); Stöckle et al. (here)

This work:

- Comparison of SDS and SADS data with methods from dialectometry
Two main goals

1. Do the different linguistic levels determine different spatial variation patterns?
   - We have data from four linguistic levels (phonology, morphology, lexicon, syntax) and two atlases (SDS and SADS).

2. Are there specific linguistic variables that predict dialect classification particularly well? From which linguistic level(s) do they stem?

   We search cluster determinants following the method of Prokić et al. (2012), implemented in Gabmap.

Conversion of the existing SDS and SADS data sets to make them comparable
Visual comparison of the data sets: parameter maps, MDS maps, cluster maps
Numerical comparison of the data sets: different correlation measures
We use VDM (Goebl 2004) and Gabmap (Nerbonne et al. 2011) for the analyses.
Two main goals

1. Do the different linguistic levels determine different spatial variation patterns?
   - We have data from four linguistic levels (phonology, morphology, lexicon, syntax) and two atlases (SDS and SADS).

2. Conversion of the existing SDS and SADS data sets to make them comparable
3. Visual comparison of the data sets: parameter maps, MDS maps, cluster maps
4. Numerical comparison of the data sets: different correlation measures
   - We use VDM (Goebl 2004) and Gabmap (Nerbonne et al. 2011) for the analyses.

2. Are there specific linguistic variables that predict dialect classification particularly well? From which linguistic level(s) do they stem?
   - We search cluster determinants following the method of Prokić et al. (2012), implemented in Gabmap.
Two main goals

1. Do the different linguistic levels determine different spatial variation patterns?
   - We have data from four linguistic levels (phonology, morphology, lexicon, syntax) and two atlases (SDS and SADS).

Conversion of the existing SDS and SADS data sets to make them comparable
Visual comparison of the data sets: parameter maps, MDS maps, cluster maps
Numerical comparison of the data sets: different correlation measures
We use VDM (Goebl 2004) and Gabmap (Nerbonne et al. 2011) for the analyses.

2. Are there specific linguistic variables that predict dialect classification particularly well? From which linguistic level(s) do they stem?

We search cluster determinants following the method of Prokić et al. (2012), implemented in Gabmap.
Two main goals

1. Do the different linguistic levels determine different spatial variation patterns?
   - We have data from four linguistic levels (phonology, morphology, lexicon, syntax) and two atlases (SDS and SADS).

Conversion of the existing SDS and SADS data sets to make them comparable

Visual comparison of the data sets: parameter maps, MDS maps, cluster maps

Numerical comparison of the data sets: different correlation measures

We use VDM (Goebl 2004) and Gabmap (Nerbonne et al. 2011) for the analyses.

2. Are there specific linguistic variables that predict dialect classification particularly well? From which linguistic level(s) do they stem?

We search cluster determinants following the method of Prokić et al. (2012), implemented in Gabmap.
Outline

1. Introduction
2. Creating comparable datasets
3. Visual analysis
4. Numerical analysis
5. Cluster determinants
6. Conclusions
## SDS vs SADS

<table>
<thead>
<tr>
<th></th>
<th>SDS</th>
<th>SADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic levels</td>
<td>Phonology, morphology, lexicon</td>
<td>(Morpho-)Syntax</td>
</tr>
<tr>
<td>Data format</td>
<td>Paper maps in 8 volumes</td>
<td>Internal database</td>
</tr>
<tr>
<td>Data availability</td>
<td>235 digitized working maps (out of 1548 paper maps)</td>
<td>118 questions on 52 phenomena</td>
</tr>
<tr>
<td>Method of data collection</td>
<td>On-site interview with trained fieldworker, direct transcription</td>
<td>4 written questionnaires, three question types</td>
</tr>
<tr>
<td>Informants</td>
<td>1–4 per location, NORMs</td>
<td>3–26 per location, all ages, genders, professions</td>
</tr>
<tr>
<td>Inquiry points</td>
<td>565 (+8 in Italy, not used)</td>
<td>383</td>
</tr>
</tbody>
</table>
## SDS vs SADS

<table>
<thead>
<tr>
<th></th>
<th>SDS</th>
<th>SADS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic levels</strong></td>
<td>Phonology, morphology, lexicon</td>
<td>(Morpho-)Syntax</td>
</tr>
<tr>
<td><strong>Data format</strong></td>
<td>Paper maps in 8 volumes</td>
<td>Internal database</td>
</tr>
<tr>
<td><strong>Data availability</strong></td>
<td>235 digitized working maps (out of 1548 paper maps)</td>
<td>118 questions on 52 phenomena</td>
</tr>
<tr>
<td><strong>Method of data collection</strong></td>
<td>trained fieldworker, direct transcription</td>
<td>three question types</td>
</tr>
<tr>
<td><strong>Informants</strong></td>
<td>1–4 per location, NORMs</td>
<td>3–26 per location, all ages, genders, professions</td>
</tr>
<tr>
<td><strong>Inquiry points</strong></td>
<td>565 (+8 in Italy, not used)</td>
<td>383</td>
</tr>
</tbody>
</table>

More on digitization in my talk on Friday!
## SDS vs SADS

<table>
<thead>
<tr>
<th></th>
<th>SDS</th>
<th>SADS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic levels</strong></td>
<td>Phonology, morphology, lexicon</td>
<td>(Morpho-)Syntax</td>
</tr>
<tr>
<td><strong>Data format</strong></td>
<td>Paper maps in 8 volumes</td>
<td>Internal database</td>
</tr>
<tr>
<td><strong>Data availability</strong></td>
<td>235 digitized working maps (out of 1548 paper maps)</td>
<td>118 questions on 52 phenomena</td>
</tr>
<tr>
<td><strong>Method of data</strong></td>
<td>/-site interview with trained fieldworker, direct transcription</td>
<td>Questionnaires, written questionnaires, three question types</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Informants</strong></td>
<td>1–4 per location, NORMs</td>
<td>3–26 per location, all ages, genders, professions</td>
</tr>
<tr>
<td><strong>Inquiry points</strong></td>
<td>565 (+8 in Italy, not used)</td>
<td>383</td>
</tr>
</tbody>
</table>

**Goal:** Create data sets with common characteristics
Unify inquiry point network
Unify inquiry point network

1. Keep matching yellow points
2. Merge red (SADS) points with nearby blue (SDS) points
3. Discard remaining red and blue points

378 common inquiry points
Unify choice of informants and phenomena

- SDS: generally one informant per inquiry point
  - Non-mobile, older, agricultural or manual job
- SADS: several informants per inquiry point
- Solution: Choose one SADS informant per inquiry point that best satisfies the SDS criteria
  - Two inquiry points did not satisfy these criteria and were discarded.

- SADS questions are not yet edited:
  - Some questions did not show dialectal variation
  - Some questions yielded too many null-values
- Solution: Discard 10 of 118 questions
Unify choice of informants and phenomena

- **SDS**: generally one informant per inquiry point
  - Non-mobile, older, agricultural or manual job
- **SADS**: several informants per inquiry point
- **Solution**: Choose one SADS informant per inquiry point that best satisfies the SDS criteria
  - Two inquiry points did not satisfy these criteria and were discarded.

- **SADS questions are not yet edited**:  
  - Some questions did not show dialectal variation  
  - Some questions yielded too many null-values
- **Solution**: Discard 10 of 118 questions
Datasets

**Resulting datasets:**
- Phonetics
  - 75 variables, from SDS
- Morphology
  - 120 variables, from SDS
- Lexicon
  - 36 variables, from SDS
- Syntax
  - 108 variables from SADS
  + 4 variables from SDS

**Remaining differences between datasets:**
- Number of variables
- Number of variants per variable
  - Between 2.9 (phonetics) and 6.9 (lexicon) on average
- Methods of data collection
  - Oral vs. written
  - On-line vs. off-line
- Periods of data collection
  - 50 years difference
Datasets

Resulting datasets:
- Phonetics
  - 75 variables, from SDS
- Morphology
  - 120 variables, from SDS
- Lexicon
  - 36 variables, from SDS
- Syntax
  - 108 variables from SADS
  + 4 variables from SDS

Remaining differences between datasets:
- Number of variables
- Number of variants per variable
  - Between 2.9 (phonetics) and 6.9 (lexicon) on average
- Methods of data collection
  - Oral vs. written
  - On-line vs. off-line
- Periods of data collection
  - 50 years difference
Outline

1. Introduction
2. Creating comparable datasets
3. Visual analysis
4. Numerical analysis
5. Cluster determinants
6. Conclusions
Mean parameter maps (VDM)

**Mittelwerte Lexik**

**Mittelwerte Morphologie**

**Mittelwerte Phonetik**

**Mittelwerte Syntax**

---

S. Kellermals, Y. Scherrer, E. Glaser, R. Weibel: A dialectometrical analysis of Swiss German dialects
Mean parameter maps (VDM)

Interpretation (Goebl 1984)

Cell color represents mean linguistic distance between given cell and all other cells.

Low (blue) values: linguistically interconnected areas

High (red) values: linguistically isolated areas
**General findings**

- Good interconnectedness of central Plateau regions: Aargau, Lucerne, Zurich
- Isolated regions: Fribourg, Valais, Grisons, partially Basel
Mean parameter maps (VDM)
Mean parameter maps (VDM)

Lexicon vs Morphology

Similar picture, but higher distance values in lexicon data
Lexicon maps have more variants per variable (6.9 vs 3.6 on average)
Mean parameter maps (VDM)

Morphology vs Phonetics
Similar value ranges
Phonetic variables differentiate less in the Northeast

S. Kellerhals, Y. Scherrer, E. Glaser, R. Weibel: A dialectometrical analysis of Swiss German dialects
Syntax

Low distance values, no structure in map
Syntactic variables are least able to differentiate dialects
Multidimensional scaling (Gabmap)
Multidimensional scaling (Gabmap)

Interpretation (Nerbonne et al. 1999)

Linguistic distance matrix is “condensed” onto three color dimensions

Visualize main dialect areas, different types of dialect borders
Multidimensional scaling (Gabmap)

- MDS Lexik
  (36 Variables, 378 Ortepunkte)

- MDS Morphologie
  (36 Variables, 378 Ortepunkte)

- MDS Phonetik
  (75 Variables, 378 Ortepunkte)

- MDS Syntax
  (112 Variables, 378 Ortepunkte)

S. Kellerhals, Y. Scherrer, E. Glaser, R. Weibel: A dialectometrical analysis of Swiss German dialects
**Multidimensional scaling (Gabmap)**

**General findings**

- Most well-known dialect borders can be identified.
- Different linguistic levels show different areal patterns.
Multidimensional scaling (Gabmap)

Syntax
Fragmented map, difficult to identify dialect borders
Outline

1. Introduction
2. Creating comparable datasets
3. Visual analysis
4. Numerical analysis
5. Cluster determinants
6. Conclusions
Hierarchical cluster analysis (VDM)

10 clusters obtained with the Ward algorithm on the entire dataset (SDS+SADS):

How well do the cluster analyses of the four subsets match?
Hierarchical cluster analysis (VDM)
Hierarchical cluster analysis (VDM)

78.3% overlap

83.9% overlap

76.2% overlap

66.9% overlap
The **Pearson product-moment correlation coefficient** measures the linear correlation (dependence) between the distance matrices of two linguistic levels.

Syntax data have much lower internal consistency (local incoherence, Cronbach’s alpha).
The Pearson product-moment correlation coefficient measures the linear correlation (dependence) between the distance matrices of two linguistic levels.

Syntax data have much lower internal consistency (local incoherence, Cronbach’s alpha).
Discussion

Our syntactic data correlate less well with the other linguistic levels, and are less homogeneous.

Löffler (2003, 109): Dialectal syntax is difficult to distinguish and isolate.

Kortmann (2010, 846): “To detect and determine the areal reach of syntactic phenomena, [...] you must look harder, since [...] syntactic variation is much subtler and less salient, less categorical, and in many cases a matter of statistical frequency rather than the presence or absence of a feature [...]”

Spruit (2008) on Dutch data: best correlation found between syntax and phonetics; his syntax data have higher internal consistency than ours.
Cluster determinants

We use the method of Prokić et al. (2012):

1. Perform hierarchical cluster analysis on the entire data set
   Ward, 10 clusters

2. For each cluster, “we seek those features which differ little within the [cluster] in question and a great deal outside that [cluster].”
   1 feature per cluster → 9 unique features
Cluster determinants

- SDS 1/134: t[ie]f (aobd. iu)
- SDS 5/112: greeting
- SDS 4/176: nur
- SDS 3/154: Vögel[chen] (diminutive)
- SDS 6/027: jeweils (time adverb)
- SDS 3/075: (an)fangen (2./3. Sg. Pres.)
- SDS 3/047: haben (Pl. Ind. Pres.)
- SDS 2/138: ge[rn] (mhd. -rn)
- SDS 6/027: jeweils (time adverb)
- SDS 3/262: gehen lassen (word order)

- Selected variables: 2 phonetics, 3 morphology, 3 lexicon, 1 syntax.
- Some, but not all of these variables are known to be characteristic to their corresponding areas.
- Explain 56% of the variance of the whole dataset (a random set of 9 variables explains 42% of the variance).
Cluster determinants

- SDS 1/134: t [ieʃ] (aobl. iu)
- SDS 5/112: greeting
- SDS 4/176: nur
- SDS 3/154: Vögel[chen] (diminutive)
- SDS 6/027: jeweils (time adverb)
- SDS 3/075: (an)fangen (2./3. Sg. Pres.)
- SDS 3/047: haben (Pl. Ind. Pres.)
- SDS 2/138: ge[rn] (mhd. -rn)
- SDS 6/027: jeweils (time adverb)
- SDS 3/262: gehen lassen (word order)

- **Selected variables:** 2 phonetics, 3 morphology, 3 lexicon, 1 syntax.
- Some, but not all of these variables are known to be characteristic to their corresponding areas.
- Explain 56% of the variance of the whole dataset (a random set of 9 variables explains 42% of the variance).
Cluster determinants

+ Selected variables: 2 phonetics, 3 morphology, 3 lexicon, 1 syntax.

+ Some, but not all of these variables are known to be characteristic to their corresponding areas.

+ Explain 56% of the variance of the whole dataset
  (a random set of 9 variables explains 42% of the variance).

SDS 1/134: t[ie]f (aobd. iu)
SDS 5/112: greeting
SDS 4/176: nur
SDS 3/154: Vögel[chen] (diminutive)
SDS 6/027: jeweils (time adverb)
SDS 3/075: (an)fangen (2./3. Sg. Pres.)
SDS 3/047: haben (Pl. Ind. Pres.)
SDS 2/138: ge[rn] (mhd. -rn)
SDS 6/027: jeweils (time adverb)
SDS 3/262: gehen lassen (word order)
Cluster determinants

Selected variables: 2 phonetics, 3 morphology, 3 lexicon, 1 syntax.

Some, but not all of these variables are known to be characteristic to their corresponding areas.

Explain 56% of the variance of the whole dataset (a random set of 9 variables explains 42% of the variance).

SDS 1/134: t[je]f (aobd. iu)
SDS 5/112: greeting
SDS 4/176: nur
SDS 3/154: Vögel[chen] (diminutive)
SDS 6/027: jeweils (time adverb)
SDS 3/075: (an)fangen (2./3. Sg. Pres.)
SDS 3/047: haben (Pl. Ind. Pres.)
SDS 2/138: ge[rn] (mhd. -rn)
SDS 6/027: jeweils (time adverb)
SDS 3/262: gehen lassen (word order)
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?
   - Phonetics, morphology and lexicon: different but comparable variation patterns.
   - Syntax: less consistent data, less clear variation patterns.
     - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
   - Morphological features correlate best with the global distance matrices.
   - Lexical features predict spatial patterns suprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?
   - We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
   - This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?
   - Phonetics, morphology and lexicon: different but comparable variation patterns.
   - Syntax: less consistent data, less clear variation patterns.
     - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
   - Morphological features correlate best with the global distance matrices.
   - Lexical features predict spatial patterns suprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?
   - We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
   - This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?
   - Phonetics, morphology and lexicon: different but comparable variation patterns.
   - Syntax: less consistent data, less clear variation patterns.
     - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
   - Morphological features correlate best with the global distance matrices.
   - Lexical features predict spatial patterns surprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?
   - We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
   - This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?

- Phonetics, morphology and lexicon: different but comparable variation patterns.
- Syntax: less consistent data, less clear variation patterns.
  - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
- Morphological features correlate best with the global distance matrices.
  - Lexical features predict spatial patterns surprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?

- We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
- This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

Do the different linguistic levels determine different spatial variation patterns?

- Phonetics, morphology and lexicon: different but comparable variation patterns.
- Syntax: less consistent data, less clear variation patterns.

  - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?

- Morphological features correlate best with the global distance matrices.
- Lexical features predict spatial patterns suprisingly well, given the small size of the data set.

Are there specific linguistic variables that predict dialect classification particularly well?

- We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
- This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?

   - Phonetics, morphology and lexicon: different but comparable variation patterns.
   - Syntax: less consistent data, less clear variation patterns.
     - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
   - Morphological features correlate best with the global distance matrices.
   - Lexical features predict spatial patterns suprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?

   - We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
   - This depends on the underlying dialect classification and on the number of desired variables.
Conclusions

1. Do the different linguistic levels determine different spatial variation patterns?
   - Phonetics, morphology and lexicon: different but comparable variation patterns.
   - Syntax: less consistent data, less clear variation patterns.
     - Is this a genuine property of syntax, or particular effect of SADS inquiry methods and data selection?
   - Morphological features correlate best with the global distance matrices.
   - Lexical features predict spatial patterns surprisingly well, given the small size of the data set.

2. Are there specific linguistic variables that predict dialect classification particularly well?
   - We have found 9 variables – from all linguistic levels – that predict the dialect regions better than random variables.
   - This depends on the underlying dialect classification and on the number of desired variables.