Areal variation in Middle High German: Methodological and quantitative aspects

Oliver Schallert (Munich)
Carsten Becker (Marburg)
Helmut Schmid (Munich)

Digital Diplomatics 2022:
From Digital to Distant Diplomatics?
30 September 2022
My two collaborators...

- Carsten Becker, *Handschriftencensus* (HSC), University of Marburg
- Helmut Schmid, *Centrum für Informations- und Sprachverarbeitung* (CIS), LMU Munich
Overview

What is our talk about?

Charters as data source
  Basic information
  The CAO
  An example

Methodology
  The database
  Classifying and mapping the data
Overview

Case studies
- Contraction phenomena
- Quantitative aspects

Conclusions

Bonus track

References
What is our talk about?

Twitter version:

- **Charters** offer fine-scaled information about areal variation in 
  MHG.
- However, they require some philological and technical skills in 
  handling.
- **Case studies:**
  - Diatopic distribution of weakly suppletive verb forms
    (contractions).
  - Searching for areal patterns in our data (MDS, clustering).
What is our talk about?

**Variation in space and time:**

- Areal patterns are the reflex of differing linguistic systems (Weinreich 1954); in modern terms, one would talk of minimally varying grammars (cf. Seiler 2004; Bresnan, Deo and Sharma 2007).
- From a diachronic perspective, they offer snapshots of the diffusion of linguistic innovations (grammaticalization processes, morphological change, etc.).
- Areal patterns are “frozen diachrony”.
In which form are MHG charters documented and accessible?

- **Corpus der altdeutschen Originalurkunden** “Corpus of Old German Original Charters until 1300” (CAO) (Wilhelm et al. 1932–2004).
  - An electronic version is available online, however, with only rudimentary search functions. (Link)
Charters as data source
Basic information

- Wörterbuch der mittelhochdeutschen Urkundensprache
  “Dictionary of the Language of Middle High German Charters” (WMU) (Kirschstein et al. 1994–2010).
  - 3 Volumes; completed 2010
  - based on file cards excerpted from the CAO
- Facsimiles (partly in digitized form) (Link)
Charters as data source

The CAO

Some background:

- Founded by Friedrich Wilhelm (1882–1932);
- total of 4617 charters; 4289 in German (to a lesser extent also other languages are included, e.g. Latin, Middle Dutch).
- Covers a period from 1200–1300; however, only 0.2 % of the materials date before 1250 (Ganslmayer 2009: 42).

<table>
<thead>
<tr>
<th></th>
<th>until 1279</th>
<th>1280–1289</th>
<th>1290–1299</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>621</td>
<td>1095</td>
<td>2901</td>
<td>4617</td>
</tr>
</tbody>
</table>

Table 1: Extant charters w. r. t. different decades
Charters as data source

Pros and cons:

- Issue date and place are usually mentioned;
- the materials are lexicographically analyzed (i.e. lemmatized) via the *Wörterbuch der mittelhochdeutschen Urkundensprache* (WMU).
- Clear dominance of Upper German regions; see the map on the next slide.
Figure 1: Number of charters per place (background map adapted from Schmidt, Herrgen and Kehrein, 2008–)
Charters as data source
An example

Next slide:

► Archbishop Friedrich II. lends Fridreich of Hafnerburg and his heirs the Zehen farm at Pirk (in the vicinity of Maria Saal, Carinthia). 10 July, 1278. (Link)

☞ Oldest known German charter issued by an archbishop of Salzburg.
Charters as data source

An example
Charters as data source

An example
Methodology

The database

Corpus der altdeutschen
Originalurkunden
bis zum Jahr 1300

(CAO)

(Wilhelm et al., 1932–2004)
Methodology

The database

Corpus der altdeutschen Originalurkunden bis zum Jahr 1300
(Wilhelm et al., 1932–2004)

Digital edition of CAO
(Gärtner, Rapp and Gniffke, 2007)
Methodology

The database

Corpus der altddeutschen Originalurkunden bis zum Jahr 1300
(Wilhelm et al., 1932–2004)

Digital edition of CAO
(Gärtner, Rapp and Gniffke, 2007)

CAO

CAO Online

PostgreSQL database
(cf. Becker and Schallert, 2021, 2022b)

▶ texts
▶ metadata (time, place)
Methodology

The database

*Corpus der altdeutschen Originalurkunden bis zum Jahr 1300*
(Wilhelm et al., 1932–2004)

Digital edition of *CAO*
(Gärtner, Rapp and Gniffke, 2007)

*CAO Online*

RNNTagger-based automatic annotation of *CAO Online* data
(cf. Schmid, 2020)

*DB*

PostgreSQL database
(cf. Becker and Schallert, 2021, 2022b)

▶ texts
▶ metadata (time, place)
Methodology

The database

- **CAO**
  - Corpus der altdeutschen Originalurkunden bis zum Jahr 1300
  - (Wilhelm et al., 1932–2004)

- **CAO Online**
  - Digital edition of CAO
  - (Gärtner, Rapp and Gniffke, 2007)

- **DB**
  - PostgreSQL database
  - (cf. Becker and Schallert, 2021, 2022b)

- **texts**
- **metadata** (time, place)

- **PoS tags**
  - RNNTagger-based automatic annotation of CAO Online data
  - (cf. Schmid, 2020)
  - ▶ lemmatization
  - ▶ parts of speech
  - ▶ morphological annotations
Methodology

The database

▶ File cards (basis for WMU = Kirschstein and Schulze, 1994–2010)

Figure 2: WMU card box and file card for üf ‘on’ referencing de Boor and Haacke (1957: 82.4) = no. 1783 (Berne and Köniz, BE, 21/07/1293)
Methodology
The database

- File cards (basis for WMU = Kirschstein and Schulze, 1994–2010)

Figure 2: WMU card box and file card for ûf ‘on’ referencing de Boor and Haacke (1957: 82.4) = no. 1783 (Berne and Köniz, BE, 21/07/1293)

- Querying PoS tags
Methodology

The database

- File cards (basis for *WMU* = Kirschstein and Schulze, 1994–2010)

*Figure 2: *WMU* card box and file card for ūf ‘on’ referencing de Boor and Haacke (1957: 82.4) = no. 1783 (Berne and Köniz, BE, 21/07/1293)*

- Querying PoS tags
- Querying lines with *regular expressions* using information on spelling variants listed in *WMU*
lâzen st.V.

lassen, unterlassen, aufgeben; zurücklassen, hinterlassen; freilassen, entlassen; überlassen, übertragen, übergeben; belassen; erlassen, nachlassen; veranlassen, bewirken, zulassen; (sich) verlassen

[1235 Bel.]


Figure 3: Excerpt from article lâzen (Kirschstein, Schulze, Ohly et al., 2003: 1095–1096)
Table 2: Attested spelling variants of *lâzen* (INF, uncontracted) ‘to let’ (Kirschstein, Schulze, Ohly et al., 2003: 1095–1096)

<table>
<thead>
<tr>
<th>Onset</th>
<th>Core</th>
<th>Coda</th>
<th>Inflection</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>ae</td>
<td>sch</td>
<td>en</td>
</tr>
<tr>
<td></td>
<td>ai</td>
<td>cz</td>
<td>in</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>hs</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>sc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>ss</td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>sz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>á</td>
<td>zc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>â</td>
<td>zs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>zz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Attested spelling variants of *lâzen* (INF, uncontracted) ‘to let’ (Kirschstein, Schulze, Ohly et al., 2003: 1095–1096)

<table>
<thead>
<tr>
<th>Onset</th>
<th>Core</th>
<th>Coda</th>
<th>Inflection</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>ae</td>
<td>sch</td>
<td>en</td>
</tr>
<tr>
<td></td>
<td>ai</td>
<td>cz</td>
<td>in</td>
</tr>
<tr>
<td></td>
<td>ay</td>
<td>hs</td>
<td>e</td>
</tr>
<tr>
<td>ä</td>
<td>sc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>ss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o</td>
<td>sz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>á</td>
<td>zc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>â</td>
<td>zs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>æ</td>
<td>zz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\bl(?: \ a [eiy] \ |a \ | [aoáâæ])\)\(\)\(?: \ [sf]ch \ |cz \ |h[sf] \ | [sfz][csfz] \ |[sftz]\)\(?: \ [ei]n \ |e \)\(\)\b
Methodology

The database

grep-style querying texts in the database

► line by line
► with regular expressions
Methodology

The database

grep-style querying texts in the database
  ▶ line by line
  ▶ with regular expressions
  ▶ script returns matches in context, metadata
Methodology
Classifying and mapping the data

- raw CSV data of limited use for mapping
- classify data, group by location of issue
Methodology
Classifying and mapping the data

- raw CSV data of limited use for mapping
- classify data, group by location of issue
- script returns counts by feature type, coordinates
Methodology
Classifying and mapping the data

- In our analyses, we only considered charters which are:
  - in German language
  - attributed to a single place of issue
  - attributed to a certain year of issue

- For mapping, we used the open source software QGIS.

❗ This removes approx. 50% of charters.
Case studies
Contraction phenomena

- Already in MHG times, several verbs show contracted forms.
  - e.g. ziehen (INF) → zien ‘pull’; ligest (2SG), liget (3SG) → lîst, lît; lâzen → lân (INF); hâben (INF) → hân, etc.
- Nübling (1995: 136) contrasts regular contractions with intervocalic $h$ and irregular contractions with $b$, $z$, and $m$.
- The latter cases cannot be explained by phonological processes.

- Short forms as cases of (weak) suppletion are highly relevant from a theoretical perspective (Nübling 2000; Corbett 2007).
Case studies
Contraction phenomena

Influencing factors on contraction (Klein, Solms and Wegera 2018: 948–949; Nübling 1995; Becker and Schallert 2022b):

- Phonotactics
- Paradigmatic position
  - token frequency
- Inflectional class
  - strong vs. weak verbs
Case studies
Contraction phenomena

▶ Spatio-temporal dimension
  ▶ increase at the end of the MHG period
▶ Polysemy
  ▶ Alleged functional differences between contracted and uncontracted forms, e.g. with haben ‘have’ (Szczepaniak 2011: ch. 6.3)

It is possible to compare our findings with the new Middle High German Grammar (Klein, Solms and Wegera 2018) and several of the modern regional atlases.
Contractions with lâzen ‘let’:

- With this verb, areal differences w.r.t. contracted and non-contracted forms can be seen already in the MHG era (Paul et al. 2007: 280–284).
- As with haben – hân ‘have’, it constitutes an irregular case of contraction.
- In modern dialects, a lot of morphological variation is found; governing factors are paradigm structure and, partly, syntax.
Case studies

Contraction phenomena
Contraction phenomena

Contractions with *haben* ‘have’:

- The MHG verb *haben* goes back to OHG *habēn* ‘to possess’ and belongs to class 3 of the weak verbs (*ēn*-verbs (Braune and Heidermanns 2018: 353 [§ 302]).
- It has become increasingly irregular in the course of its historical development and forms a weakly suppletive paradigm in contemporary German (Nübling, 2000: ch 2.1, 2001).
  - Connection with the grammaticalization of the periphrastic perfect in German (Grønvik 1986 and others)
  - Layering of possessive *haben* (‘own’) and uses as a temporal auxiliary (‘have’).
Case studies
Contraction phenomena

▶ At the end of the OHG linguistic period, forms with elision of the root consonant appear (Braune and Heidermanns, 2018: 419 [§ 368, note 4]) that cannot be phonologically conditioned.

▶ haben → hân, but not graben ‘dig’ → *grân

▶ In MG, short forms appear in more token-frequent positions, full forms in less-frequent ones.
Case studies
Contraction phenomena

Paradigmatic position:

- A gradual increase in short forms by person and number can be observed, culminating in the 3SG (cf. de Boor 1976: 121).
- In the past tense, they appear without exception (cf. de Boor 1976: 131).
- This amounts to the implicational scale in (1), with the proportion of contracted forms increasing from left to right.

\[
(1) \text{ INF (43.3\%) > 1PL.PRES (47.6\%) > 3PL.PRES (57.7\%) > 1SG.PRES (94.1\%) > 3SG.PRES (100\%) > SG/PL.PST (100\%)}
\]
Maps:

- With the 1PL.PRES, contracted forms are found in Alemannic, East Franconian and occasionally in Ripuarian (also cf. Kirschstein, Schulze, Ohly et al., 1994: 777–778).

- Regarding the 1SG.PRES, the picture is reversed, i.e. full forms are clearly in the minority and only occur in the northern and eastern regions of the Alemannic area (cf. de Boor, 1976: 122), occasionally also in southern Bavarian.
Case studies
Contraction phenomena

Figure 4: Variants of MHG *haben* (1PL.PRES) in the CAO data
Case studies

Contraction phenomena

Figure 5: Variants of MHG *haben* (1SG.PRES) in the CAO data
Case studies
Contraction phenomena

Full verb vs. auxiliary:

- No significant differences in areal distribution between the two variants (see Table 3).
- In both uses, contracted forms predominate over full forms at three quarters to one quarter of occurrences.

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Occurrences</th>
<th>Ratio contracted forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>full verb</td>
<td>211</td>
<td>156 (73.9%)</td>
</tr>
<tr>
<td>auxiliary</td>
<td>498</td>
<td>375 (75.3%)</td>
</tr>
</tbody>
</table>

Table 3: Contracted forms of MHG haben ‘to have’ regarding verb types
Comparison with the *Sprachatlas des deutschen Reichs* (SdR):

- Based on a large survey initiated by the German linguist Georg Wenker (1852–1911).
- Conducted between the 1880s and 1930s, gradually extended to other German-speaking regions (Austria, Switzerland, etc.).
- Questionnaire with 40 model sentences (sent to all school locations) in Standard German, to be translated into the local dialect
- More than 50,000 completed questionnaires altogether, collected in Marburg; basis for the maps of the *Sprachatlas*.  
  (Chambers and Trudgill 1998: 15–16)

Accessible via *Regionalsprache.de* (Link)
Case studies

Contraction phenomena

- **1PL.PRES (SdR, Map 338 [Test Sentence 23]):**
  - The crucial difference between both time stages is that contraction has spread to larger parts of Bavarian.
  - Syncope and progressive assimilation of the alveolar nasal to \( b \) as trigger: \( honen > hobn > hobm > hom \)

- In the 1SG.PRES (SdR, Map 117 [Test Sentence 8]):
  - Situation roughly corresponds to the one found in modern dialects.
Case studies
Contraction phenomena

Figure 6: Variants of MHG haben (1PL.PRES) in the CAO data
Case studies

Contraction phenomena

Figure 7: Variants of NHG *haben* (1PL.PRES) in the SdR
Case studies

Contraction phenomena

Figure 8: Variants of NHG *haben* (1PL.PRES) in the SdR
The CAO-materials from a quantitative perspective:

- Analysis of trigram tokens for all charters with unique issue location.
- We compared the respective frequency profiles via cosine distance and used two techniques for visualizing/analyzing the ensuing distance matrix (Birkenes 2019):
  - Multidimensional scaling (Heeringa 2004)
  - Hierarchical cluster analysis (Ward’s method with two clusters)

Thanks to Magnus Birkenes (Oslo) who provided us with his results.
The cosine similarity between two vectors $\vec{x}, \vec{y}$ corresponds to the cosine of the angle $\theta$ between them (Manning and Schütze 1999: 300):

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

The value ranges between $cos(180^\circ) = -1$ (opposite) und $cos(0^\circ) = 1$ (proportional); orthogonal vectors have a similarity of $0^\circ$. 

The cosine similarity between two vectors $\vec{x}, \vec{y}$ corresponds to the cosine of the angle $\theta$ between them (Manning and Schütze 1999: 300):

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

The value ranges between $cos(180^\circ) = -1$ (opposite) und $cos(0^\circ) = 1$ (proportional); orthogonal vectors have a similarity of $0^\circ$. 
Case studies

Quantitative aspects

Figure 9: CAO-data (MDS)
Case studies
Quantitative aspects

Figure 10: CAO-data (Hierarchical cluster analysis)
Case studies
Quantitative aspects

- Multidimensional scaling yields a tripartite division that reflects the traditional difference between West and East Upper German quite clearly, whereas the central and northern regions show more overlaps.
- This west-east divide in the southern regions is corroborated by a hierarchical cluster analysis.
Conclusions

► The “Corpus” is an interesting base for investigations into areal variation in Middle High German.
  ► Some limitations in areal coverage (the bulk of the materials is from the Upper German regions).
  ► Different means of comparison and validation of our findings, in particular the SdA (and regional atlases).

► Areal structures from a qualitative (contractions) and quantitative perspective (MDS, clustering).

☞ Several additional phenomena are discussed in Becker and Schallert (2021, 2022a,b).
Thanks for your attention!
Diphthongization:

- Affects the MHG high long vowels î (≈ iː/), iu (≈ /yː/), and û (≈ /uː/)
- In larger parts of the language area, they develop into the diphthongs NHG /aɪ̯/, /ɔɪ̯/ und /aʊ̯/.
  - Mnemonic: MHG *mîn niuwez hûs > nhd. mein neues Haus*
    - ‘my new house’
  - First traces already in the 12th century CE, mostly in Carinthia and South Tyrol (Paul et al. 2007: 74–77 [§ L 17]) Not affected areas: Alemannic, parts of Thuringian, North Hessian, and Ripuarian
Figure 11: Areal diffusion of diphthongization (König, Elspaß and Möller 2019: 146)
In the course of the 16th century, the new digraphs (e.g. ⟨au⟩) were adopted in general written usage.

Modern Standard German shows merger of the new diphthongs with the old lowered diphthongs of MHG:

- *ei, oü, ou* → /aɪ̯/, /ɔɪ̯/, /aʊ̯/
- In the dialects, these two rows are usually kept distinct, e.g. Bavarian *hoaß* ‘hot’ [= MHG *ei*], *Freid* ‘joy’ [= MHG *öu*], *Bam* ‘tree’ [= MHG *ou*].
Bonus track

Maps (based on the automatically tagged charter data):
- wîn(garte)
- liute
- hûs
For checking the reliability/validity, the figures given by Lindgren (1961: 18) can be used (based on Vol. 1–2 of the CAO).

- Bavarian texts already show an amount of 65% diphthongized forms.
- In the adjacent regions (East Franconian, Swabian, etc.), the rates are considerably lower and range by about 5–10%
References I


References III


References IV


Gärtner, Kurt, Andrea Rapp and Andreas Gniffke, eds. (2007). *Corpus der altdeutschen Originalurkunden bis zum Jahr 1300*. University of Trier. URL: http://tcdh01.uni-trier.de/cgi-bin/iCorpus/CorpusIndex.tcl (visited on 22/02/2022).


References VI


References VIII


