Annotating a multiregional diachronic corpus of Early New High German handwritten texts

Lisa Dücker, Stefan Hartmann, Renata Szczepaniak (Universität Hamburg)
39. Jahrestagung der DGfS, Saarbrücken
09.03.2017
AG 4 Encoding language and linguistic information in historical corpora
Outline

1. Our corpus – protocols of witch trials (16th-17th centuries)
2. Annotation
   1. Tokenisation – (graphic and syntactic tokens)
   2. Sentence boundaries
   3. Animacy
   4. Semantic roles
3. Summary and outlook
The corpus – protocols of witch trials

• (semi-) spontaneously produced, handwritten texts (based on Macha et al. 2005)
• 56 protocols of witch trials from 16th – 17th centuries
• number of tokens: 61,870 (average length: 1,105 per protocol)
• multi-layer annotation in GATE (gate.ac.uk)

• the main goal of the SiGS-project: a multifactorial analysis of the usage and spread of sentence-internal capitalization in Early New High German
Sentence-internal capitalization

• in the standard orthography of modern German: sentence-internal capitalization of nouns and nominalisations (=head of a noun phrase), e.g. das groß-e Haus ‘the big house’
  
  das groß-e Aber ‘the big but’

• the 16th and 17th centuries – the crucial period for the development of the sentence-internal capitalization

• factors supporting the capitalization in previous research:
  • pragmatic factors (reverence),
  • syntactic factors (part of speech, majuscule as a noun marker)
  • semantic factors (animacy-driven spread: humans > concretes > abstracts)
Annotation: Tokenisation

• two-level annotation = distinction between graphic and syntactic tokens

example: clitics (contraction of preposition and articles), compounds

\[
\begin{align*}
auff=m & \quad \text{Teufel-ß} & \quad \text{dantz} \\
at=[\text{the}]\text{DAT} & \quad \text{devil-LE} & \quad \text{dance[DAT]}
\end{align*}
\]

at the devil’s dance
(Alme 1630)
Annotation of sentence boundaries

• the sentence boundary detection is an essential precondition for the analysis of sentence-internal capitalization

• standard detection means, like punctuation marks, sentence-initial capitalization or finite verb forms are not reliable means of identifying sentence boundaries in historical texts
Annotation of sentence boundaries

• the sentence boundary detection is an essential precondition for the analysis of sentence-internal capitalization

• standard detection means, like punctuation marks, sentence-initial capitalization or finite verb forms are not reliable means of identifying sentence boundaries in historical texts

• „minimal sentence“: a clause with a lexical verb and its subordinate clauses

[daruff Clägerin geantworttet, weiln ihr kein geldt geben worden],
[darbey es Also verblieben].
Animacy

- superhuman (positive)
- superhuman (devil)
- human (collective)
- human
- animal
- concrete (collective)
- concrete (body part)
- concrete (place)
- concrete
- abstract (measure)
- abstract

- superhuman
- human
- animal
- concrete
- abstract

- custom annotation scheme
- coded by two trained annotators
- high inter-annotator agreement (F = 0.99)
Animacy

• ... sie wolte Gott eine reine Seele vberantworten (Crivitz 1642) 'She wanted to deliver a pure soul to God'

abstract?

animate?

concrete (body part?)

• Ir buel hannß Federle (Messkirch 1644) 'her lover Hans Federle'

human

superhuman (devil)
Results: Animacy and frequency

Binomial mixed-effects model:

- **Response variable:**
  - Uppercase

- **Fixed effects:**
  - Animacy
  - Frequency

- **Random effects:**
  - Protocol (≈ Scribe)
  - Lemma

\[
\text{uppercase} \sim \text{Animacy} \times \log(\text{Frequency}) + (\text{Animacy} | \text{Protocol}) + (1 | \text{Lemma})
\]
### Results: Animacy and frequency

|                      | Estimate | Std. Error | z value | Pr(>|z|)   |
|----------------------|----------|------------|---------|------------|
| (Intercept)          | -2.01    | 0.25       | -8.05   | <0.001***  |
| Animacy-concrete     | 1.27     | 0.27       | 4.62    | <0.001***  |
| Animacy-animal       | 1.7      | 0.59       | 2.87    | <0.001***  |
| Animacy-human        | 1.82     | 0.38       | 4.81    | <0.001***  |
| Animacy-superhuman   | 3.79     | 1.28       | 2.97    | <0.001***  |
| log\(_{10}\)(Freq)  | 0.39     | 0.2        | 1.9     | 0.06 .     |
| Animacy-conc×log\(_{10}\)(Freq) | -0.85    | 0.26       | -3.32   | <0.001***  |
| Animacy-anim×log\(_{10}\)(Freq) | -0.17    | 0.68       | -0.25   | 0.8        |
| Animacy-hum×log\(_{10}\)(Freq) | -0.35    | 0.32       | -1.11   | 0.27       |
| Animacy-sup×log\(_{10}\)(Freq) | -1.79    | 0.76       | -2.36   | 0.02 *     |

**Model diagnostics:**

\(C = 0.94, C_{xy} = 0.87, \) all VIFs < 5
High frequency seems to have an effect on capitalization, but this effect is not straightforward. There is much lexeme-specific variation between highly-frequent lexemes. This points to the importance of additional factors: socio-pragmatics, semantic roles, and syntactic functions. For example, in the core corpus, terms denoting women are significantly less often capitalized than terms denoting men (Fisher exact test: p<0.001, odds ratio=3.01).
Semantic role annotation

• Proto-roles for NPs (Dowty 1991)
• corpus-specific additions

• coded annotated by two trained annotators
• high inter-annotator agreement (F=0.875)
Semantic roles in the SiGS corpus

+ volitional involvement
+ sentience
+ causing an event or change of state
+ movement (relative to other participants)
+ undergoes change of state
+ incremental theme
+ affected by another participant
+ stationary (relative to other participants)

proto-agent

experiencer
stimulus
proto-patient

Sie fürchtet die Dunkelheit
She fears the darkness

„daß sie vnschuldigh wehre“ (Minden 1614)
that she was innocent

09.03.2017 Dücker/Hartmann/Szczepaniak 14
Meta roles

• ambiguous:
  
  solches hab sie ein bettell frau Ir gelehrt (Gaugrehweiler 1610)
  a begger woman taught her that
  she, a beggar woman, taught her that

• multiple roles:
  
  Cappelle Aber sehr erschrocken (experiencer) vnd gezittert (proto-agent) (Helmstedt 1580)
  Capelle was startled and [he] shivered

• no role:
  
  „Derselbe habe Ihr ein stucke goldes geben“ (Jever 1592)
  He gave her a piece of gold
Semantic roles

• only four protocols annotated so far
• very tentative first results:
  • patient nouns are capitalized slightly more often than expected
  • no significant differences in distribution across animate/inanimate
  • however, only very few agentive nouns
Summary and outlook

• Animacy encoding has already yielded interesting results
• additional factors seem to play a role as well
• annotation for semantic roles and syntactic functions as a promising approach
Summary and outlook

• Corpus will be released via ANNIS
• Principles of annotation are made transparent in detailed annotation guidelines
• As such, the corpus can also be used to address a variety of follow-up questions and other research questions related to ENHG
• additional research questions could address (or have addressed)
  • the role of gender,
  • capitalization of other parts of speech (e.g. pronouns and adjectives)
  • morphological and syntactic questions (e.g. compounding in ENHG)
References


Thank you for your attention!

And a very special thank you goes out to our former and current annotators: Annemarie Bischoff, Aleksa Krieg, Sophie Muehlenberg, Merle Pfau, Nicolai Pudimat, Eleonore Schmitt, Tanja Stevanovic, and Annika Vieregge!