

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 **H**aus ‘the big **h**ouse’
das groß-e **A**ber ‘the big **b**ut’
- 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

auff=m
at=[the]DAT

Teufel-ß
devil-LE

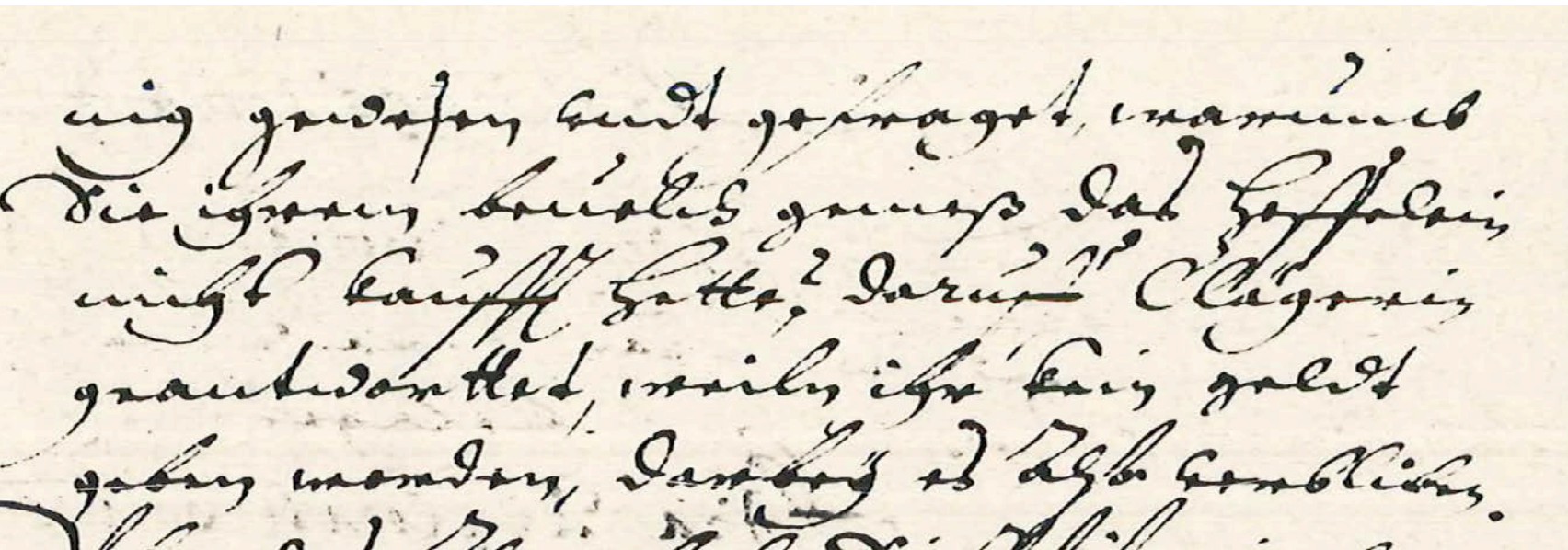
dantz
dance[DAT]

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



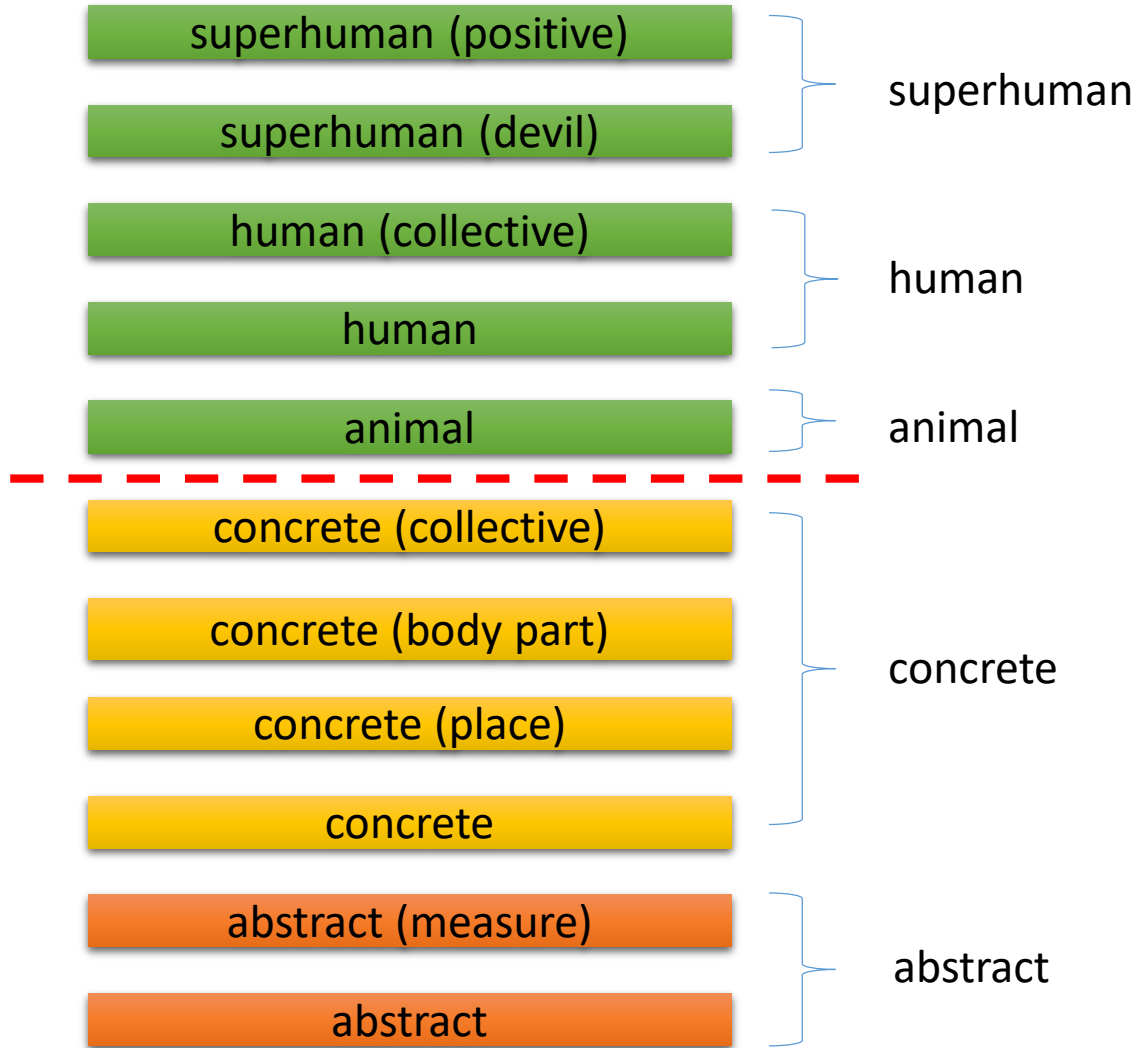
daruff Clägerin
geantwortet, weiln
ihr kein geldt geben
worden, darbey es
Also verblieben.

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



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

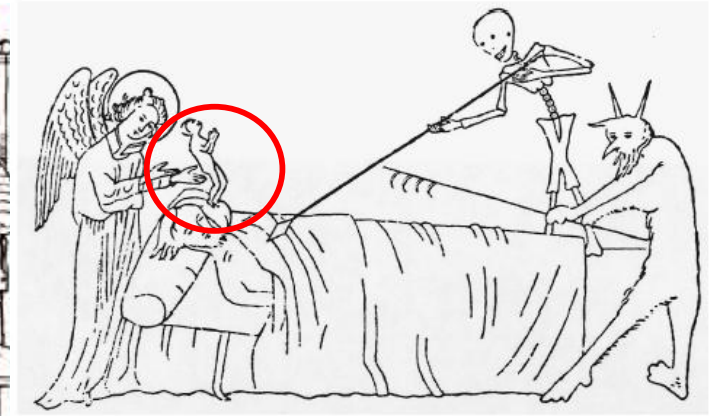
Animacy

- ... 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

$$\text{uppercase} \sim \text{Animacy} * \log(\text{Frequency}) + (\text{Animacy} | \text{Protocol}) + (1 | \text{Lemma})$$

- **Random effects:**

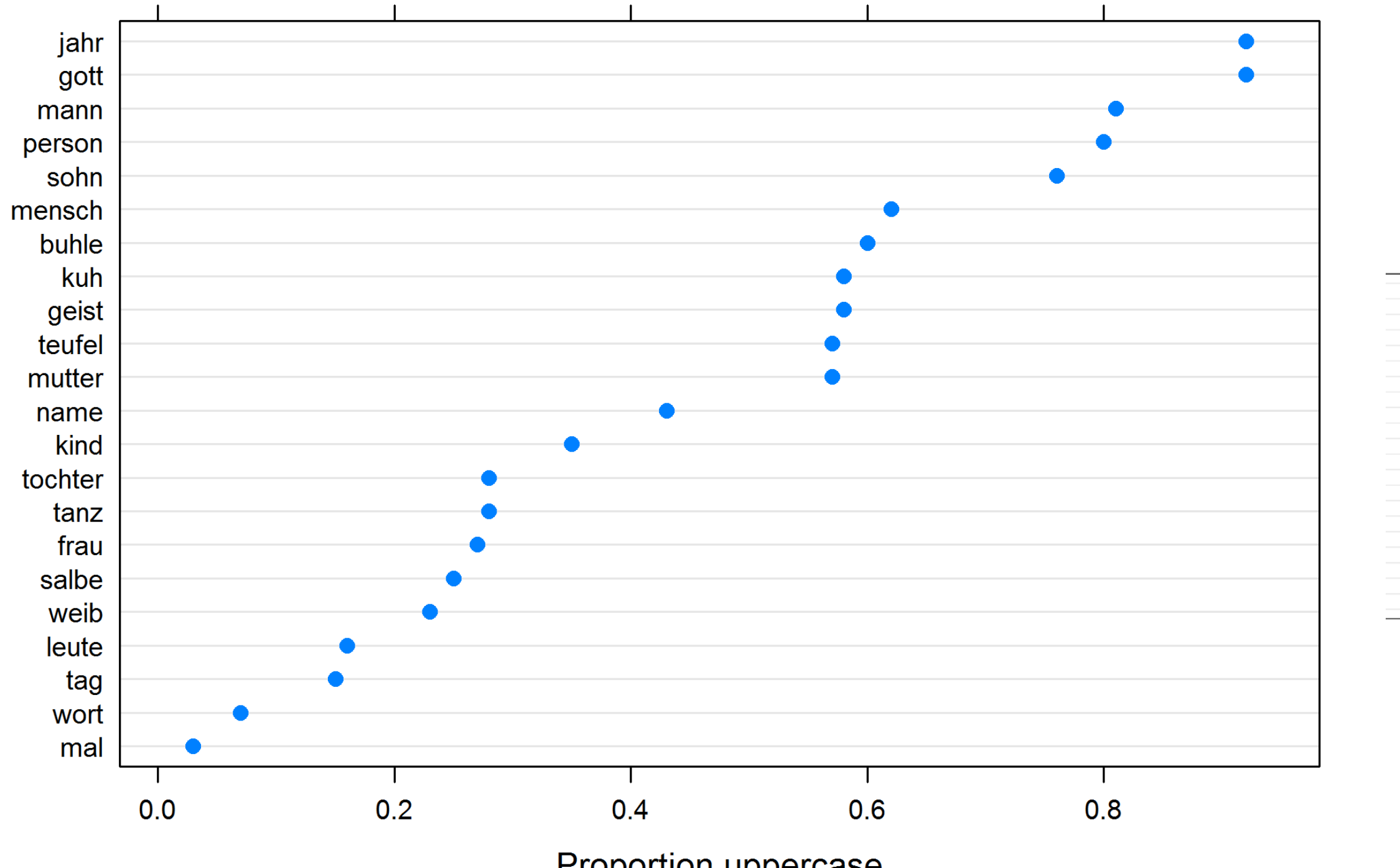
- Protocol (\approx Scribe)
- Lemma

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*** |
| log₁₀(Freq) | 0.39 | 0.2 | 1.9 | 0.06 . |
| Animacy-conc×log₁₀(Freq) | -0.85 | 0.26 | -3.32 | <0.001*** |
| Animacy-anim×log₁₀(Freq) | -0.17 | 0.68 | -0.25 | 0.8 |
| Animacy-hum×log₁₀(Freq) | -0.35 | 0.32 | -1.11 | 0.27 |
| Animacy-sup×log₁₀(Freq) | -1.79 | 0.76 | -2.36 | 0.02 * |

Model diagnostics:

$C = 0.94$, $C_{xy} = 0.87$, all VIFs < 5



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)



experiencer

stimulus

Sie fürchtet die Dunkelheit

She fears the darkness

stative

„daß sie vnschuldigh wehre“ (Minden 1614)

that she was innocent

Meta roles

- *ambiguous:*

solches hab sie ein bettell frau Ir gelehret (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)*

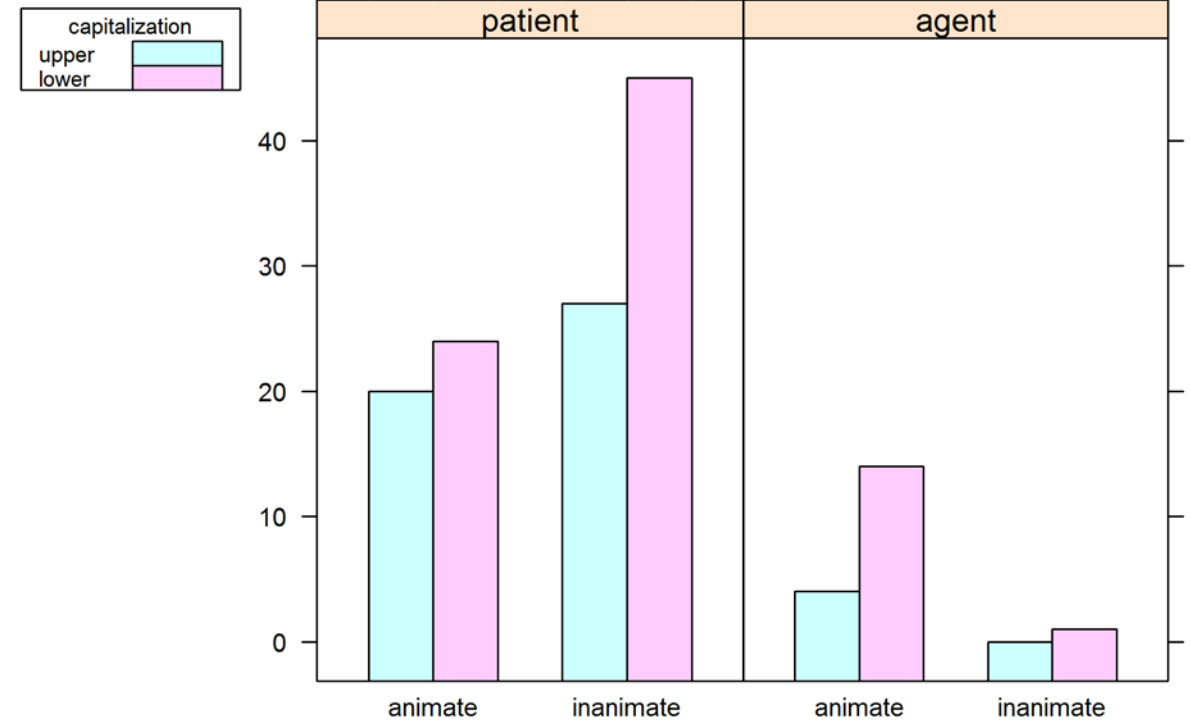
Capelle was startled and [he] shivered

- *no role:*

„Derselbe habe Ihr ein stücke 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

- Bergmann, Rolf/Nerius, Dieter (1998): Die Entwicklung der Großschreibung im Deutschen von 1500 bis 1700. 2 Bände. Heidelberg: Winter.
- Cunningham, Hamish/Maynard, Diana/Bontcheva, Kalina et al. (2014): GATE Devel-oper. General Architecture for Text Engineering. Version 8.1. Sheffield.
- Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. *Language* 67(3). 547–619.
- Macha, Jürgen, Elvira Topalović, Iris Hille, Uta Nolting & Anja Wilke (eds.). 2005. *Deutsche Kanzleisprache in Hexenverhörprotokollen der Frühen Neuzeit. Bd. 1: Auswahl-edition*. Berlin, New York: De Gruyter.
- Risse, Ursula (1980): Untersuchungen zum Gebrauch der Majuskel in deutschsprachigen Bibeln des 16. Jahrhunderts: Ein historischer Beitrag zur Diskussion um die Substantivgroßschreibung. Heidelberg: Winter.
- Rössler, Paul (1998): Die Großschreibung in Wiener Drucken des 17. und frühen 18. Jahrhunderts. In: Bauer, Werner/Scheuringer, Hermann (Hgg.): Beharrsamkeit und Wandel. Festschrift für Herbert Tatzreiter zum 60. Geburtstag. Wien: Edition Praesens, 205–238.
- Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor & Tom Wasow. 2004. Animacy Encoding in English: Why and How. In Bonnie Webber & Donna Byron (eds.), *DiscAnnotation '04*, 118–125. Stroudsburg, PA: Association for Computational Linguistics.
- Weber, Walter Rudolf (1958): Das Aufkommen der Substantivgroßschreibung im Deutschen: Ein historisch-kritischer Versuch. zugl. Diss. Univ. Bern, 1952. Uni-Druck: München.

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!