Computational Narratology: Extracting Tense Clusters from Narrative Texts

Thomas Bögel, Jannik Strötgen, Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Context of this Work

Computational Narratology

- context: Digital Humanities
- facilitate annotations from literary scientists
- support hypotheses [1]
- methods: Natural Language Processing

Temporal phenomena

- field of study in narratology
- temporal structure of literary texts
- examples: time shifts, order phenomena (e.g., prolepsis)

The heureCLÉA Project [4]

Cooperation

- BMBF-funded eHumanities project
- narratologists (Hamburg)
- computer scientists (Heidelberg)
- temporal phenomena in literary text

Goals

- collaborative annotation framework that automatically suggests annotations
- reduce effort of manual annotations
- analysis of temporal aspects in narrative texts

Temporal Phenomena in Literary Texts

Temporal expressions

- explicit temporal information
- can be extracted automatically (HeidelTime [2])
- less frequent in literary text

Tense information

- tenses: information about temporal structures
- shifts in tenses indicate order phenomena (e.g., prolepsis)

Prior work

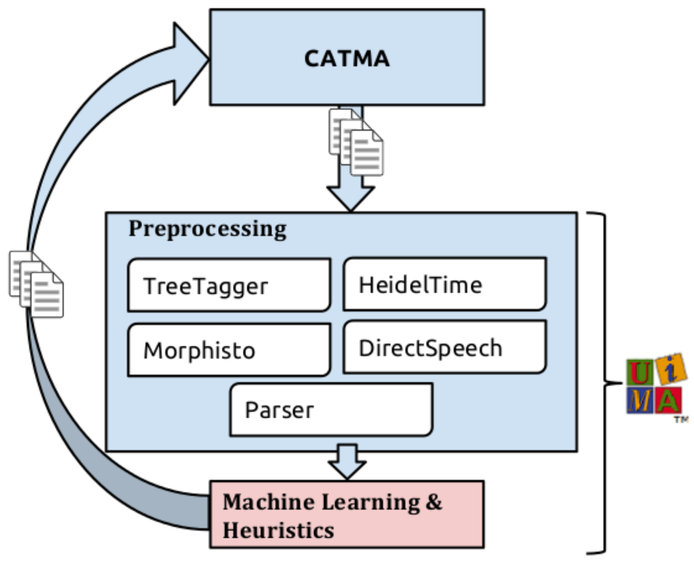
- laborious manual annotations
- automatic systems focus on English
- no existing system for German tense annotation

Data set

- German narrative texts (20^{th} century)
- annotations by literary scientists
- tagset: narratological aspects

Task: robust annotation of tenses in narrative texts

NLP Pipeline Architecture [3]



Component Description and Tools

NLP components

- POS tagging: TreeTagger
- morphology: Morphisto
- time expressions: HeidelTime
- syntactic parsing: Parzu & Stanford parser

CATMA interface

- CATMA: collaborative annotation platform
- flexible CATMA $\leftrightarrow$ UIMA interface
- tailored to narratologists

Machine learning interface

- feature extraction and machine learning
- interchangeable algorithms
- goal: predict annotations automatically

Feedback loop on predicted annotations

- manual corrections
- improvement of future predictions (ML)

Use Case: Tense Annotations

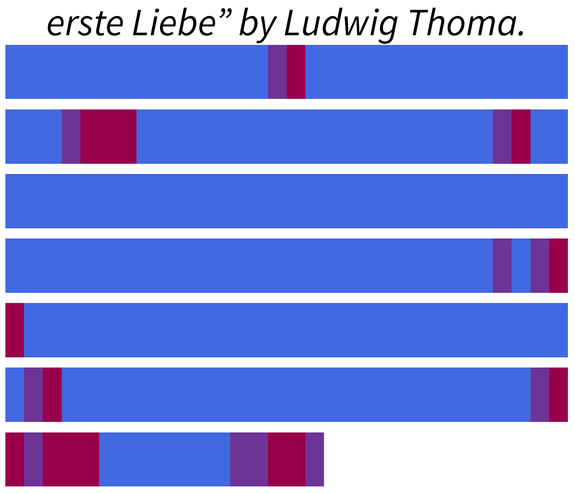
Extracting temporal clusters

- temporal cluster: all tokens governed by the same verb
- exploitation of tense markers (e.g., auxiliaries)
- morphological features & heuristics
- heuristic for sentences with unknown tense
- evaluation: comparison to manual annotations
- high inter-annotator agreement ($\kappa > 0.8$)

Ongoing work

- machine learning based system for additional annotations, e.g., narrative levels
- hybrid, self-improving system: heuristics + machine learning

Visualization of temporal clusters in "Meine erste Liebe" by Ludwig Thoma.



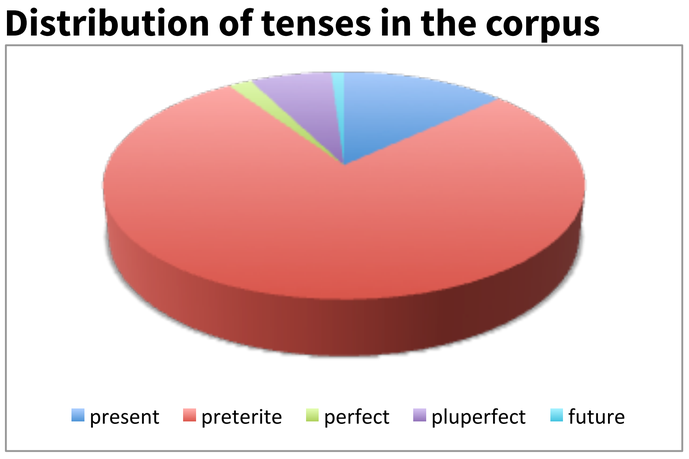
- each block: temporal cluster
- colors correspond to tenses
  - *blue*: preterite
  - *red*: future
  - *purple*: present
- text mostly written in preterite, interrupted by forecasts.

Evaluation Results

tense	correctly tagged verbs
present	93.10
preterite	95.73
perfect	96.43
pluperfect	84.71
future	90.00

$\Rightarrow$ reliable and robust prediction of tense clusters

Distribution of tenses in the corpus



References

[1] I. Mani: Computational Narratology. **The living handbook of narratology**. http://www.lhn.uni-hamburg.de/article/computational-narratology, 2013.

[2] J. Strötgen and M. Gertz: **Multilingual and Cross-domain Temporal Tagging.** *Language Resources and Evaluation*, 47(2), 269–298, 2014.

[3] T. Bögel, J. Strötgen, C. Mayer and M. Gertz: **A Flexible NLP Pipeline for Computational Narratology.** *1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd)*, 2014.

[4] The heureCLÉA Project: http://www.heureclea.de/.

Contact Information:

Thomas Bögel

boegel@uni-hd.de

http://dbs.ifi.uni-heidelberg.de/