




**Universität Stuttgart**  
Institut für Maschinelle Sprachverarbeitung  
Center for Reflected Text Analytics



# Automatisierungs- perspektiven für die Erforschung von Literatur

Nils Reiter



Connection to Literary Studies



The human in the loop  
Automatic != fully automatic



Cost Benefit Ratio

# Connection to Literary Studies

## Ensuring the Satisfaktionsfähigkeit



- Top-Down

- Operationalize concept of interest
- Ensure intersubjective application
  - Annotation guidelines (e.g., in a shared task)
- Corpus, reference data, machine learning, ...
- Large-scale analysis of texts ✓

Humans

- Bottom-Up

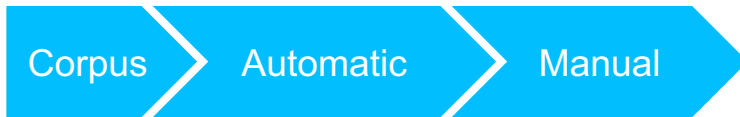
- Select appropriate corpus
- Operationalize concept of interest
  - E.g., instrumental variables / approximate operationalization
- Find patterns (e.g., unsupervised), interpret them
- Connect pattern interpretations to a literary interpretation ✓

Computers

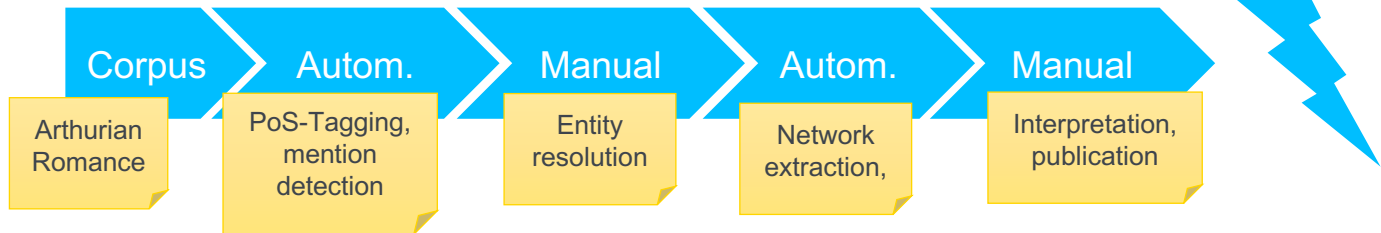
# Automatic != fully automatic



- Full automatization requires (lots of) reference data
- Not available for interesting phenomena
- Interactive tools for exploration



- Manual steps can be integrated earlier



# Cost Benefit Ratio

## Method development

- NLP
    - Large data sets, *and constantly new texts*
    - (Almost) no upper bound on invested time
  - CLS
    - In many cases: No new texts (unless contemporary literature)
    - Paying someone to annotate Goethes Werther might be faster and cheaper than optimizing an NLP model for it
    - Unless: The work on the optimization itself leads to something interesting
- ➡ Focus on generic phenomena





Use cases



Context



Black box

# Automatization goals / use cases



- Finding interesting cases (“Finde-Heuristik”)
  - Precision matters
  - Interactivity allows users to express/fine-tune their preferences
- Annotation support (candidate generation)
  - Recall matters
  - Practical questions How much annotation do you need before it makes sense? It depends.
- Hypothesis testing
  - “Female characters have a stronger association with illness than male characters, in 19th century novels of genre X.”
  - Claims in literary studies do not come in hypothesis form
  - Corpus linguistics toolbox
  - Operationalization crucial
  - Representativity

# Context, Ambiguity, Polyvalence

“Peter saw Judy with the binoculars.”

- Linguistic ambiguity: Ambiguous sentence, but can be resolved with context
- “Resolved”: Readings become implausible
- Claim: Only if assumptions are made
  - that Peter and Judy didn’t switch binoculars in between the sentences
  - that binoculars can be used for seeing
  - that ...
- Narratological categories behave similarly
- Ambiguities may remain
- Many ambiguities are unnoticed by humans





# Context, Ambiguity, Polyvalence

- Result of text analysis: Propositional content of the text
- Interpretation = Interpretation Theory + Text + Context
- Theory suggests
  - which parts of the text are paid attention to
  - which context to use
- Deterministic, following rules?
- If not, is ruling out interpretations possible?



# Black box

- Are black boxes a problem?
  - Depends on use case
    - DHd panel on deep learning: black box OCR ok
  - Empirically validated black boxes are often not a problem, unpredictable performance on new texts is
- Performance on a new text (type) is unknown
  - But we can do empirical validation
  - “Annotate these (representative/difficult) sentences, and we tell you how reliable the pos-tagger was”
- Domain adaptation
  - Users willing to do some annotation can get better results



# Conclusions

- Support for semi-automatic processing (integration of manual and automatic processing steps)
- Optimize tools for the right kind of error
  - In general, precision errors are easier to fix than recall errors
- No one knows the performance of a tool on a new text (type)
  - Support empirical validation
  - Users willing to do some annotation can get better results





**Universität Stuttgart**  
Institut für Maschinelle Sprachverarbeitung  
Center for Reflected Text Analytics

**Vielen Dank!**



**Nils Reiter**

E-Mail [nils.reiter@ims.uni-stuttgart.de](mailto:nils.reiter@ims.uni-stuttgart.de)

<https://nilsreiter.de>

Universität Stuttgart  
Institut für Maschinelle Sprachverarbeitung  
Pfaffenwaldring 5b  
70569 Stuttgart