Motivation

- Coreference resolution could benefit many DH tasks
- Character analysis
- Relation analysis
- We analyze the specific properties of coreference-related phenomena in literary texts
- We give directions for the adaptation of guidelines
- We present a new annotation tool

Analyzed texts

<table>
<thead>
<tr>
<th>Title</th>
<th>Literary form</th>
<th>Language</th>
<th>Published in</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Sorrows of Young Werther (Goethe)</td>
<td>Epistolary novel</td>
<td>German</td>
<td>1778</td>
</tr>
<tr>
<td>from Nine to Nine (Perutz)</td>
<td>Novel</td>
<td>German</td>
<td>1918</td>
</tr>
<tr>
<td>Der Mond lacht (Perutz)</td>
<td>Novel</td>
<td>German</td>
<td>1930</td>
</tr>
<tr>
<td>Nur ein Druck auf den Knopf (Perutz)</td>
<td>Novel</td>
<td>German</td>
<td>1930</td>
</tr>
<tr>
<td>Miss Sara Sampson (Lessing)</td>
<td>Play</td>
<td>German</td>
<td>1755</td>
</tr>
<tr>
<td>The Robbers (Schiller)</td>
<td>Play</td>
<td>German</td>
<td>1782</td>
</tr>
<tr>
<td>six fairytales by the Brothers Grimm</td>
<td>Fairytale</td>
<td>German</td>
<td>1812</td>
</tr>
<tr>
<td>N Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson</td>
<td>US captivity</td>
<td>English</td>
<td>1682</td>
</tr>
</tbody>
</table>

Literary-specific phenomena/considerations

Text knowledge

- Characters (and reader) have varying states of knowledge
- Often purposefully deployed play with mix-ups
- Example: In crime novels, when should the description of the murderer be linked to a certain character?
  - We annotate from the reader’s point of view
  - Fixed knowledge after one read-through (for standardization)

Text length

- Long texts pose challenges to the annotator’s attention span
- Pronominal coreference is limited to a certain distance
- Nominal coreferences might span hundreds of pages, e.g., the appearance of Schehezerade in One Thousand and One Night
  - We present a new tool

Genericity

- High rate of switching between the generic and non-generic use
  - Annotations should distinguish them

Entity development

- The concept of identity is often played with
- Are the prince and the frog in Frog Prince the same entity?
- Also difficult: creation of groups, plural references to groups
  - In exact composition often unclear, e.g., in The Robbers
  - In unclear cases, we prefer separate entities

World knowledge

- Contemporary annotators do not have the same knowledge as a typical reader at the time of the work’s publication
  - Where something is assumed to be common knowledge, the annotator may look up missing facts

Idiomatic expressions

- Very frequent in literary text
- Sometimes difficult to decide which expressions are idiomatic
  - with a twisted face

Sub-token annotation

- High rate of switching between the generic and non-generic use
- For example: Are they the same entity?
  - Leave true ambiguities in the form of divergent annotations

Annotation tool

- Focus on enabling annotation of long texts with many discourse entities
- Coreference annotations are conceptualized as equivalence sets represented by colors
- Sub-token annotation: Arbitrary text spans can be annotated
- To address issue of text length:
  - entities can be searched for
  - found spans can be annotated as a new or existing entity
  - already annotated entities are shown
- Tool can also compare two annotations and highlight differences
- Flexible import and export in a variety of formats
- Tool is openly available:

https://github.com/nilsreiter/CorefAnnotator

Restoration of Mrs. Mary Rowlandson
A Narrative of the Captivity and...