Can We Describe a Literary Character by Its Explicit Attributions Based on Syntactic Annotation?

Introduction
A great deal of distant reading research focuses on the investigation of character constellations, for instance character networks (Moretti 2011). Far fewer attempts have been made to automatically describe characters. The studies that exist are usually word-based analyses. For instance, Braun et al. (2018) present a study on character complexity using the context of character mentions. John et al. (2017) visualize the context of a character’s mentions in the text as a word cloud. The potential of syntactic annotation for this purpose has only been explored in a few works (e.g. Sudhahar et al. 2015). Our study will contribute to this line of research and is part of the research project hermA (Gaidys et al. 2017).

Characterization in a narrative text can take two forms: It is either implicit and relies on the reader’s knowledge of the extratextual world, or it is explicit and happens by ‘textually explicit ascription of properties to a character’ (Jannidis 2009: 22). The latter can potentially be identified at the text surface in terms of certain lexical or syntactic patterns. In this paper, we explore the potential of describing a character with the help of explicit attributions. In order to do so, we exploit syntactic dependency annotation and coreference annotation. One of the simplest syntactic types of explicit attribution is non-verbal predication like the following:

(1) Rosentreter ist ein netter Junge.
‘Rosentreter is a nice boy.’

Adelmann et al. (2018) have shown that syntactic dependency parsers achieve reasonable results on literary texts, but their performance is low when analyzing non-verbal predicates. Our goal is to determine whether this type of analysis is nevertheless feasible.

Data and Method
Our analysis is based on two German novels: the contemporary Corpus Delicti by Juli Zeh from 2009 (46,499 tokens) and Aus guter Familie by Gabriele Reuter from 1895 (72,534 tokens). This choice allows us to test our approach on two texts of different time and style. For both texts we manually resolve coreference of character mentions and annotate the texts accordingly in CorefAnnotator (Rösiger et al. 2018). In addition, we annotate syntactic dependencies (including part of speech and lemma) using two different parsers: Mate (Bohnet 2010), which achieved the best overall performance in Adelmann et al. (2018), and Malt (Nivre 2003) with the model ‘covnonproj’, which performed best on the specific task of detecting non-verbal predicates (see Adelmann et al. 2018 for training details). Based on these annotations, we pick all sentences with a character mention in subject function and a non-verbal predicate (label: PRED). We extract triples of subject, verb and non-verbal predicate (e.g. Rosentreter – be – a nice boy) for further investigation.

Results
In terms of quantitative evaluation, we want to know how many non-verbal predicates the parsers detect and how many of them are correctly attributed to a character. Tables 1 and 2 give an overview of the number of non-verbal predicates that were identified by the parsers. For Corpus Delicti, these are 308 non-verbal predicates in total. Note that in only 52 (~17%) of these the character is mentioned by its proper name. For the rest, we had to rely on our
coreference resolution. In *Aus guter Familie* we find 424 non-verbal predicates, 77 (~18%) of them for proper names. We can see that the overlap between the results of the two parsers is surprisingly small: Only 85 (28%, *Corpus Delicti*) and 48 (11%, *Aus guter Familie*) of the results are shared by Mate and Malt. Generally speaking, Malt finds many more instances than Mate (58% and 137% more).

<table>
<thead>
<tr>
<th>Found by Mate</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found by Malt</td>
<td>yes</td>
<td>85</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>67</td>
<td>-</td>
</tr>
<tr>
<td>total</td>
<td>152</td>
<td>156</td>
<td>308</td>
</tr>
</tbody>
</table>

Table 1: Number of non-verbal predicates found by the two parsers in *Corpus Delicti*

<table>
<thead>
<tr>
<th>Found by Mate</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found by Malt</td>
<td>yes</td>
<td>48</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>92</td>
<td>-</td>
</tr>
<tr>
<td>total</td>
<td>140</td>
<td>284</td>
<td>424</td>
</tr>
</tbody>
</table>

Table 2: Number of non-verbal predicates found by the two parsers in *Aus guter Familie*

A high number of instances is of no value if the analysis is false or—in this particular case—the predicate is attributed to the wrong entity. Two annotators checked for all instances whether the relation between subject and verb as well as verb and non-verbal predicate were correct. In table 3 we present the precision for the two parsers, their combined output and the instances found by one of the parsers only.\(^1\) When both parsers agree, we can be very confident that the attribution is correct (0.94). Each parser on its own achieves much lower scores. This is especially true for Mate. Our current setting does not allow us to determine the recall of the parsers.\(^2\) As we do not have a gold-standard annotation of our texts, there is no way to determine how many non-verbal predicates were missed by both parsers. In the study presented in Adelmann et al. (2018), the average recall for the label PRED was 0.74 for Malt and 0.26 for Mate. Hence, we have to assume that quite a number of instances were missed in the present study, too.

<table>
<thead>
<tr>
<th>Precision for instances detected by</th>
<th><em>Corpus Delicti</em></th>
<th><em>Aus guter Familie</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mate</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>Malt</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>both parsers</td>
<td><strong>0.94</strong></td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>Mate only</td>
<td>0.61</td>
<td>0.53</td>
</tr>
<tr>
<td>Malt only</td>
<td>0.72</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 3: Precision of attributions detected by different parser constellations in the two novels

From a qualitative perspective, we want to know whether these data are informative for character description. It is an open question how this type of evaluation can be formalized in an objective way. For now, we rely on our judgment as informed readers of the texts under investigation.

In the novel *Corpus Delicti*, the main character is called Mia and rebels against a dictatorship based on the absolutization of health. Among the nominal predicates for Mia

---

1. Precision = # correct instances found by the parser / # all instances found by the parser.
2. Recall = # correct instances found by the parser / # all instances in the corpus.
are ‘not a school girl’, ‘scientist’, ‘nihilist’, ‘a witness’, ‘supporter of the METHOD’3, ‘a saint’, ‘a hero’, and ‘a good child’. Her antagonist Kramer on the other hand is ‘a patient man’, ‘a machine’, ‘a fanatic’, ‘a media figure’, ‘man of conviction’, and ‘a brilliant demagogue’. These attributions give a good first impression of the roles these characters fulfill in the novel. The second text, Aus guter Familie, is about the girl Agathe who suffers from expectations society places on women of her time (19th century) and is considered mentally ill by the end of the novel. Attributions to her include ‘so happy’, ‘very sensitive’, ‘very excited’, ‘insecure’, ‘very sad’, ‘very tired’, ‘pale and ill’, and ‘no longer young’.

In both texts we get different types of information about the characters, i. e.:
- biographical information (‘not a school girl’; ‘scientist’),
- psychological information (‘so happy’; ‘very sensitive’; ‘very excited’),
- information about the characters’ narrative function (‘a hero’; ‘a witness’) and
- their political and philosophical attitude (‘a brilliant demagogue’; ‘a fanatic’; ‘supporter of the METHOD’, ‘nihilist’).

A comparison of both texts’ attributions shows, that the information type can additionally be used for assumptions about character types, genre and literary epoch: The attributions about Agathe mainly correspond to psychological information while the attributions about Mia and Kramer tend to be one of the three other information types.

The attributions cannot all be taken at face value, though, as they can have different degrees of factuality. For instance, some of them just describe possibilities. This can sometimes, but not always, be detected by modal verbs, a verb in subjunctive mood or adverbs such as vielleicht (‘maybe’):

(2) Wäre ich ein Hund [...] ‘If I were a dog [...]’

Also, the attribution might be in the scope of a negation missed by our analysis. In other cases, the attributions represent the opinion of one specific character which does not necessarily do justice to the character as a whole:

(3) Interessanterweise [...] sind Sie ein noch größeres Arschloch, als wir dachten. ‘Interestingly, [...] you are an even bigger asshole than we thought.’

Still, these attributions give us access to specific facets of a character.

Conclusion

To conclude, we have seen that despite the mediocre performance of dependency parsers in the task of detecting non-verbal predicates, we can still use these annotations to get a general impression of characters. It proved to be highly beneficial to combine the two parsers Mate and Malt as their results differ substantially. The overlap between their results is small, but for those instances we can be very confident that the attribution is correct. In the future, we intend to use this information to identify genre-specific character types, e. g. members of the resistance in dystopias. Another future field of application are interview data collected in a palliative setting (Gaidys & Begerow 2018). Here, we want to determine the mental state of patients.

One of the drawbacks of our approach is the necessity of coreference annotation, which as yet cannot be automated to a satisfactory extent. When only relying on proper names, the number of identifiable non-verbal predicates drops considerably and so does the number of informative triples that we can extract for analysis.

---

3 Name of the health-based state system which she used to trust before her brother was falsely convicted.
References


