Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 55–64
Marseille, 11–16 May 2020

GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German

Janis Pagel, Nils Reiter
University of Stuttgart, Institute for Natural Language Processing (IMS)
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Dramatic texts are a highly structured literary text type. Their quantitative analysis so far has relied on analysing structural properties (e.g., in the form of networks). Resolving coreferences is crucial for an analysis of the content of the character speech, but developing automatic coreference resolution (CR) systems depends on the existence of annotated corpora. In this paper, we present an annotated corpus of German dramatic texts, a preliminary analysis of the corpus as well as some baseline experiments on automatic CR. The analysis shows that with respect to the reference structure, dramatic texts are very different from news texts, but more similar to other dialogical text types such as interviews. Baseline experiments show a performance of 28.8 CoNLL score achieved by the rule-based CR system CorZu. In the future, we plan to integrate the (partial) information given in the *dramatis personae* into the CR model.

Keywords: corpus, coreference, computational literary studies, literary text, dramatic text

1. Introduction

This paper introduces a new annotated corpus. A selection of (parts of) German dramatic texts (1730-1920 CE) has been annotated manually with coreference information, to be used as a gold standard for experiments on automatic coreference resolution (CR). Coreference occurs when entities are referenced multiple times in a text, to be seen in Figure 1: All occurrences of the name *Romeo* refer to the same character, and the same is true for Juliet. Additionally, the phrases *a window*, *yonder window* and *It* refer to the same window; *the sun* and *fair sun* refer to the sun.

Next to prose and poetry, drama is one of the three major literary genres, and an interesting text type with respect to both its linguistic and literary properties. Dramatic texts are highly structured: Speakers and stage directions are clearly marked (by typography or in machine-readable XML, see Fig. 1 for an example) and separated from the text characters utter. Spoken text is (most of the time) clearly associated with the character who utters it, and the entire text is hierarchically segmented into acts, scenes, and, sometimes, appearances. While dramatic texts can be considered as the 'template' for theatrical productions, not all dramatic texts have been written with the intention of being produced on stage, and not all theatre plays are based on dramatic texts. Due to the strong structure, and the fact that this structure has been encoded in TEI/XML frequently, it is straightforward to access some properties of the texts, without any language analysis per se. For instance, extracting character networks that are based on co-presence is a major research trend within the computational literary studies community (e.g., Trilcke et al. (2015)). This can be (and has been) combined with an analysis of linguistic properties of characters' voices with various goals (e.g., Bullard and Ovesdotter Alm (2014), Vishnubholta et al. (2019)).

This, however, only allows partial insight, as these analyses only access the language material uttered actively by a character. Both readers and audience perceive dramatic characters also indirectly, through what other characters say about them. Some characters are mentioned a lot more than they are actively speaking. In addition, analysing characters' speech acts does only take characters into account, and misses entities such as persons without an active role or important objects relevant to the plot.

SCENE II.
Capulet's orchard.
Enter Romeo
ROMEO. He jests at scars that never felt a wound.
Juliet appears above at a window
But, soft! what light through yonder window breaks?
It is the east, and Juliet is the sun. Arise, fair sun, and
kill the envious moon, ...

Figure 1: Excerpt from Shakespeare's Romeo and Juliet (Act II, Scene 2)

CR is therefore the key task to address for an adept analysis. Annotated coreference chains in dramatic texts would allow to extract how characters are depicted by other characters, e.g., in the form they are addressed ('the serpent') or in the propositional content that is said about them ('Juliet is the sun'). CR on dramatic texts can be expected to be easier than for prose texts: It is known who is speaking and potentially listening, which restricts the antecedents of pronouns in first and second person. The *dramatis personae* contains information about the main characters, such as familial association, gender information, etc. At the same time, dramas are literary texts, and not written to purely convey information. Some ways of mentioning entities are aesthetically motivated and the texts in general have multiple levels of meaning, which are more or less well studied and/or explicit. The texts are also heterogeneous, as many aspects are era-, epoch-, or author-dependent. This makes an application of existing coreference resolvers not straightforward.

This paper is structured as follows: We will review related work on coreference resolution for German and/or literary texts in Section 2. Section 3 describes the data set that we release, Section 4 gives insights into the annotation process, including highlighting some interesting phenomena and inter-annotator agreement. Section 5 presents a quantitative analysis of the annotated corpus and compares it with

existing German coreference corpora. Section 6 describes baseline experiments for the automatic resolution of coreferences, and Section 7 shows its potential for the analysis within computational literary studies. Section 8 concludes by highlighting our next planned research steps.

2. Related Work

CR has received a lot of attention, mostly focused on the English language and evaluated on news texts (Raghunathan et al., 2010; Björkelund and Kuhn, 2014; Clark and Manning, 2016; Martschat, 2017). The two published CR systems for German are CorZu and IMS HotCoref DE. CorZu (Tuggener, 2016) is a rule-based system that iteratively eliminates possible mention pairs by checking a number of linguistic features. It achieves 64.79 MELA F-Score (Pradhan et al., 2012) on TüBa-D/Z, an annotated corpus of German news text[1]. IMS HotCoref DE (Rösiger and Kuhn, 2016) is a machine learning system that is based on the multi-language IMS HotCoref system by Björkelund and Kuhn (2014). HotCoref searches for possible antecedents based on latent search trees and uses global features to train a perceptron for classification. HotCoref DE modifies HotCoref by adding language-specific properties such as morphological information and making use of GermaNet (Hamp and Feldweg, 1997). The system is considered to be the state of the art for German CR, evaluated on the SemEval Shared Task 2010 evaluation data.

There are only few publications that focus on **CR for literary texts**. BookNLP (Bamman et al., 2014) is a full NLP pipeline optimised for (English) long texts. To resolve coreferential links, noun phrases are clustered, following Davis et al. (2003) (w/o evaluation). The resolution of anaphora is based on linguistically motivated features and a classifier and achieves an accuracy of 82%, evaluated on under 900 mention pairs from three literary novels. Krug et al. (2015) describe a CR system that only resolves references to literary characters. The system is an adaptation of Raghunathan et al. (2010) and achieves a performance of about 56 B^3 F-Score (outperforming CorZu). No system has been published that is tailored to dramatic or dialogical texts. Only a few publications deal explicitly with domain adaptation for CR systems (Yang et al., 2012; Zhao and Ng, 2014).

TüBa-D/Z is the largest available German **data set** with coreference annotation (Naumann, 2007). It consists of newspaper texts annotated on various levels. More recently, two more news-related corpora for German have been published: DIRNDL (Björkelund et al., 2014) provides transcribed and annotated radio news and GRAIN (Schweitzer et al., 2018) transcribed and annotated radio interviews. Structurally, interview data is somewhat similar to dramatic texts, due to the existence of different speakers. Reiter et al. (2017) and Bamman et al. (2019) released corpora of German resp. English literary texts, annotated with entity types, but without coreference links. DROC (Krug et al., 2018) does contain coreference links, but for small segments of prose texts and only for literary characters, i.e., contains only a subset of coreference links. Rösiger et al. (2018) describe challenges in the annotation of literary texts, including plays.

[1] https://uni-tuebingen.de/de/134290

3. Annotated Data

The data set we release with this paper consists of 31 plays, which can be found in Table 8.[2] The plays were written between 1730 and 1920 CE. At least one play is taken from each decade in this time frame.

The plays have been digitised and enriched with structural TEI/XML markup in the TextGrid[3] project, and then enhanced as part of the GerDraCor corpus (Trilcke et al., 2015). Our annotations are created on top of GerDraCor, and comprise single acts of most plays (several plays have been annotated fully). The corpus is available in a GitHub repository[4] in three formats: (i) CoNLL, for development of automatic coreference resolution systems (we follow the GRAIN idea and added a speaker column), (ii) XMI, used in the Apache UIMA framework[5], because this is the 'original' annotation format, and (iii) TEI/XML for allowing reintegration into other drama analysis tools or corpora.

The annotation was conducted by 8 annotators in total, who used the CorefAnnotator annotation tool (Reiter, 2018). All annotators are native speakers of German, and undergraduate students of literature. Each annotator was trained on text #6 (see Appendix), and regular meetings were held to discuss difficult cases. The annotation guidelines are based on the insights from Rösiger et al. (2018), which in turn use the TüBa-D/Z guidelines as a basis.

Annotation Challenges and Phenomena. Annotating coreference chains regularly entails a number of challenges. We will focus on challenges and observations specific to this data set. One specific challenge is the **text length**. Coreference annotation is usually done on relatively short documents. An entire dramatic text may fill a book, and is thus much longer (see below for some statistics). At the same time, coreference chains may span the entire text, and characters or objects mentioned at the beginning can easily re-appear at the very end. We therefore asked the annotators to first read the entire text in print, to get an overview of the plot and the involved entities. Many plays, in particular comedies, build their plot on mix-ups of characters. In non-comedic plays, characters often wear **disguises**, and are (initially) not recognized by the others. Our annotation is done from the reader's perspective, i.e., disguised characters are linked to the 'real' ones, but this is not entirely satisfactory, as it creates coreference chains that violate agreement constraints (if a character is addressed as one of another gender, for instance). In addition, it is only a partial representation of the mental model of a reader, because they are usually aware that other characters perceive the entity differently. A related issue is the development of

[2] State at submission time. As the annotations continue, the released corpus will be larger. All numbers and analyses below refer to these 31 plays, and will also be updated for the final version.

[3] https://textgrid.de

[4] https://github.com/quadrama/gerdracor-coref – It can also be found under its DOI: 10.5281/zenodo.3559206 and its ISLRN: 259-896-856-753-3

[5] https://uima.apache.org

	Documents
TüBa-D/Z	3350
DIRNDL	55
GRAIN	23
GerDraCor-Coref	45
GerDraCor-Coref S	294

Table 1: Number of **documents** for the different corpora.

	Total	Mean	Sd
TüBa-D/Z	1 565 620	467.3	478.2
DIRNDL	38 634	702.4	212.7
GRAIN	42 324	1840.2	153.4
GerDraCor-Coref	298 352	6630.0	2601.6
GerDraCor-Coref S	252 984	860.5	1015.2

Table 2: Total number of **tokens**, as well as mean values and standard deviation over documents.

	Total	Mean	Sd	Density	AML
TüBa-D/Z	144 785	43.2	48.5	0.09	2.10
DIRNDL	2832	51.5	21.0	0.07	2.66
GRAIN	6832	297.0	40.6	0.16	2.72
GerDraCor-Coref	61 126	1358.4	478.6	0.20	1.52
GerDraCor-Coref S	49 068	166.9	182.0	0.19	1.52

Table 3: Total number of **mentions**, as well as mean values and standard deviation over documents. *Density* is the total number of mentions divided by the total number of tokens. *AML* is the *Average Mention Length* in tokens.

characters. Characters may change their name, appearance or address forms, which changes how they are referenced. Arguably due to the literary language, we also encountered a lot of **ambiguities** that were ultimately not resolvable. These include entity groupings, for which it is not clear who exactly belongs to the group and who does not. In combination with the length of the texts, in which members join and leave the group, groupings proved difficult to annotate. As can be expected, the somewhat colourful language also impacts the reference structure, as references may be paraphrased instead of repeated.

Inter-Annotator Agreement was determined on a subset of three acts[6] and was found to be 62.3 CoNLL score on average (MUC F-Score: 78.7). This is slightly lower than the 83 MUC F-Score reported by Versley (2006) for TüBa-D/Z, but quite reasonable given the literary nature of the texts.

4. Corpus Analysis

To get an insight into how this corpus differs from existing CR corpora, we provide several statistics about the distribution of the annotations and highlight interesting properties. We compare the annotations with three other corpora:

TüBa-D/Z which is a large corpus based on German newspaper texts, and next to coreference (Naumann, 2007) also includes other manual annotations such as parts-of-speech, named entities and syntax trees (Telljohann et al., 2004). We make use of version 10.0.

DIRNDL. A corpus of German broadcast news, annotated (among others) for information status and prosodic features (Baumann and Riester, 2012; Björkelund et al., 2014).

GRAIN. Based on 30-minutes broadcast interviews in German, covering political topics and, as DIRNDL, annotated for information status (Riester and Baumann, 2017; Schweitzer et al., 2018).

As discussed, our annotations comprise single acts of most plays. In order to compare our annotations to shorter documents as one can find in TüBa-D/Z or DIRNDL, we also provide splits into scenes, which will henceforth be called *GerDraCor-Coref S(cenes)*. Plays which only have acts and no scenes are filtered out. Each scene is treated as a single document and coreference chains spanning multiple scenes are cut.

4.1. General Statistics

Table 1 shows the number of documents for each corpus. TüBa-D/Z is clearly the largest of the corpora, followed by GerDraCor-Coref S.

Tokens. TüBa-D/Z also provides the highest number of tokens (Tab. 2), followed by the dramatic acts with the second most tokens. It should be noted that the dramatic acts supply very large documents, as seen with the mean tokens value based on the documents. Both splitting by scene and by act yields a very high variation in document length, shown by the standard deviation. GerDraCor-Coref S has fewer tokens than GerDraCor-Coref, since the acts that do not have any scenes are not included in GerDraCor-Coref S.

Mentions. We also compare the number of mentions (Tab. 3). Once again, TüBa-D/Z has the highest number in total, but on a per document basis, the plays have much more mentions, and thus potentially a higher chance to trigger mistakes in an automatic resolution. As the 'documents' vary considerably in size, we also calculate mention density (i.e., number of mentions divided by number of tokens). Interestingly, mention density in the plays is also substantially higher than for all other corpora. Since GRAIN also has a comparatively higher mention density than TüBa-D/Z or DIRNDL, we assume this is an effect of the dialogical nature of GRAIN and GerDraCor-Coref. In order to get an idea about the size of the mentions, we compute the average length of mentions (AML) measured in tokens. This reveals that for GerDraCor-Coref, the mentions are much shorter on average than for the other corpora. A plausible explanation is the higher use of pronouns in GerDraCor-Coref, which will be explored further in Section 4.2.

Entities. Table 4 shows a similar overview, but for entities, i.e., distinct coreference chains. Again, we see that the number of entities is higher in GerDraCor-Coref compared

[6]Only some acts have been annotated in parallel by multiple annotators. In particular, these are: Three annotators for the first act of text #9 and two annotators for act I and II of text #6. In order to include these acts into the gold standard, disagreements have been resolved by an additional annotator.

	Total	Mean	Sd	Density
TüBa-D/Z	39 575	11.8	11.9	0.025
DIRNDL	1171	21.3	8.9	0.030
GRAIN	1767	76.8	8.2	0.042
GerDraCor-Coref	5473	121.6	50.9	0.018
GerDraCor-Coref S	6654	22.6	19.1	0.026

Table 4: Total number of **entities**, as well as mean values and standard deviation over documents. *Density* is the total number of entities divided by the total number of tokens.

	Max	Mean	Sd	Norm
TüBa-D/Z	187	14.4	14.4	0.0013
DIRNDL	12	7.4	1.9	0.0042
GRAIN	84	44.0	14.1	0.0123
GerDraCor-Coref	543	328.8	126.8	0.0089
GerDraCor-Coref S	429	58.0	55.1	0.0087

Table 5: Number of mentions in the longest **coreference chain** (*Max*), as well as mean values and standard deviation over documents. *Norm(alized)* is the number of mentions in the longest chain divided by the total number of mentions.

to the newspaper texts. The density for entities is much lower compared to the other corpora, though. This might be due to the fact that in dramatic texts, entities are mainly characters, which are introduced once and remain present for the rest of the text.

Longest coreference chains. As a final table in this section, we show the number of mentions in a chain, and in particular in the long chains. The result can be seen in Table 5 and shows that plays contain very long chains, going up to 543 mentions in one of the documents. This is very different from the other corpora, which have much shorter chains. Interestingly, when comparing a normalized value that shows the number of mentions in a longest chain against the total number of mentions in a corpus, GRAIN shows that it has the longest chains considering its number of mentions. However, the plays follow shortly after.

Summing up these observations, dramatic texts clearly diverge from other corpora in these properties. Although they are essentially written texts, in some statistics they are closer to the GRAIN corpus than the others (e.g., mention density). This is likely a consequence of the dialogical structure of both.

4.2. Parts of Speech

Figure 2 shows the distribution of some frequently occurring parts-of-speech within mention annotations (adjectives, punctuation, articles, common nouns, proper names and pronouns) for the different corpora. The remaining parts-of-speech are subsumed into the *Other* category. This gives some insight into how entities are commonly referred to.

For GerDraCor-Coref, the part-of-speech annotations were generated automatically, using the implementation of the

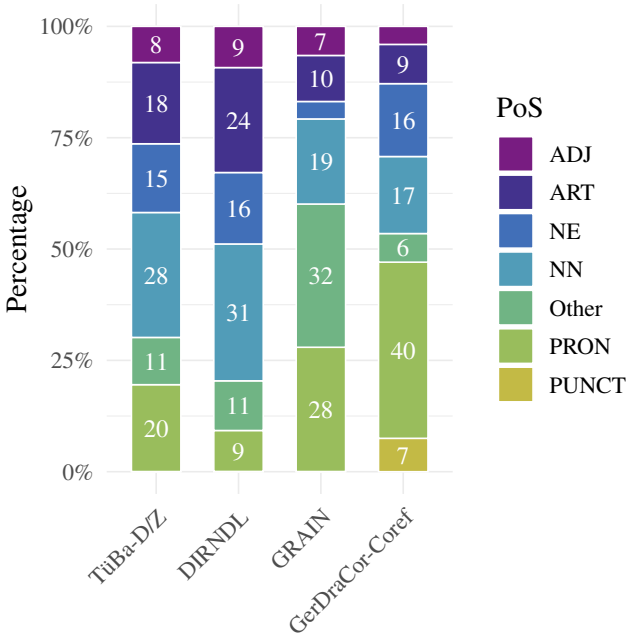

Figure 2: Barplot showing the counts of certain parts-of-speech in all the mentions. *Other* is the sum of the remaining parts-of-speech.

Stanford POS tagger (Toutanova et al., 2003) by the DKPro NLP project.[7]

In GerDraCor-Coref, pronouns are by far the most common way of referring to entities (40%), followed by noun and name mentions, which are on par (17% and 16%). These findings are in line with Krug et al. (2018), who also found pronoun references to highly outnumber other references for their annotation of characters in German novels. Punctuation is occurring relatively commonly (7%) since speaker designations often contain a final full stop (compare 'ROMEO.' in Figure 1). In GerDraCor, the speaker tags including dots are already marked as mentions, which we adopted.

In comparison to the other corpora, some interesting observations can be made. Firstly, both 'spoken' language corpora, GRAIN and GerDraCor-Coref, make more use of pronouns compared to the other corpora, and less use of common nouns. Apart from GRAIN, all corpora make approximately the same use of names. For the news domain, this is not surprising, since in a news setting reporting about people is quite common, whereas in the interview setting in GRAIN, the participants are given by context. For the plays, names are also a common way of referring to people. Slightly surprising is the small percentage of adjectives in GerDraCor-Coref (4%), while the other corpora make equal use of adjectives. One could assume that the literary language of the plays also induces a higher rate of adjectives to more eloquently describe entities; this is apparently not the case.

[7] https://dkpro.github.io

	Total	Mean	Sd
Generic Entities	113	3.65	4.83
Ambiguous Mentions	32	1.03	2.68
Abstract Antecedents	115	3.71	6.91
Abstract Anaphora	309	30.90	41.64

Table 6: Overview of the distribution of certain entity and mention properties: Generic entities, abstract antecedents and anaphora, ambiguous mentions. Total counts apply for the whole corpus. Mean value and standard deviation are calculated over the complete dramas (not for single acts or scenes).

4.3. Abstract Anaphora, Generic Entities & Ambiguous Mentions

We also annotated generic entities and ambiguous mentions. Abstract anaphora and antecedents (see e.g. Kolhatkar et al. (2018)) have been annotated for a part of the corpus. We did not annotate predicates.

For the generic entities, some general observations can be made: Generics in GerDraCor-Coref mostly consist of pronouns like *man* (one) or *wir* (we). Other common themes are gender: *der Mann* (the man), *das Weib* (the woman, archaic); nationality: *die Schweden* (the Swedes), *die Sarazenen* (the Saracens); or humans in general: *die Menschen* (humans). Sometimes, characters make use of generics to talk about other characters in an indirect way. In *Der sterbende Cato* by Gottsched (Table 8, #1), the character Cato is talking about *einen Tyrann* (a tyrant) in a general way and how such a person will always punish the ones who helped in ascending the throne. However, from context it is very clear that he is actually referring to the character Cäsar (cf. Rösiger et al. (2018, p. 131)), which we also annotated accordingly.

An overview of the total amount of occurrences in the corpus can be found in Table 6. The high standard deviation suggests that the findings are very diverse and dependent on the text.

5. Baseline Experiments

In addition to the statistics of the annotated data, we carried out baseline experiments with an automatic CR system. The results are setting an orientation for further experiments on the corpus and showcase several challenges when dealing with coreference resolution in literary data.

We opted to make use of CorZu (Tuggener, 2016) for all experiments, as it is a readily usable rule-based system, and decisions about training setup and the like do not need to be made. Krug et al. (2015)'s system would have been another plausible alternative, as it was developed on the literary domain; however, it only resolves mentions of characters, and is thus not completely suited for our needs.

Table 7 shows the results of applying CorZu on the same data sets as used in Section 4. We report MUC, B^3 and $CEAF_e$ scores (Pradhan et al., 2014), as well as the average of these three scores, often referred to as CoNLL score. The performance is highest for TüBa-D/Z, which is not surprising, as CorZu has been developed mainly on this resource (Tuggener, 2016, p. 74). The performance for GerDraCor-Coref is on par with the performance on DIRNDL (CoNLL score); the performance is considerably lower for B^3 and $CEAF_e$ scores, though. The same can be said about GRAIN, on which the overall performance is also the lowest.

	MUC	B^3	$CEAF_e$	CoNLL
TüBa-D/Z	54.3	49.7	53.3	52.4
DIRNDL	32.3	33.4	37.8	34.5
GRAIN	32.3	19.6	20.1	24.0
GerDraCor-Coref	47.5	17.2	21.7	28.8
GerDraCor-Coref S	44.9	24.9	25.6	31.8

Table 7: Mean values for MUC, B^3, $CEAF_e$ and CoNLL scores on the different corpora, using CorZu.

CR generally works better for GerDraCor-Coref S than for GerDraCor-Coref, which can be explained by the shorter average document length of GerDraCor-Coref S.

When looking at the distribution of CoNLL scores across documents, as can be seen in Figure 3, it becomes clear that coreference resolution is very dependent on the document in question. Especially for the scenes, there is a lot of variation and outliers. Please note that the number of 'documents' in GerDraCor-Coref S is much higher, which explains the higher number of potential outliers shown in the plot. The same is also true for TüBa-D/Z when compared to the other corpora.

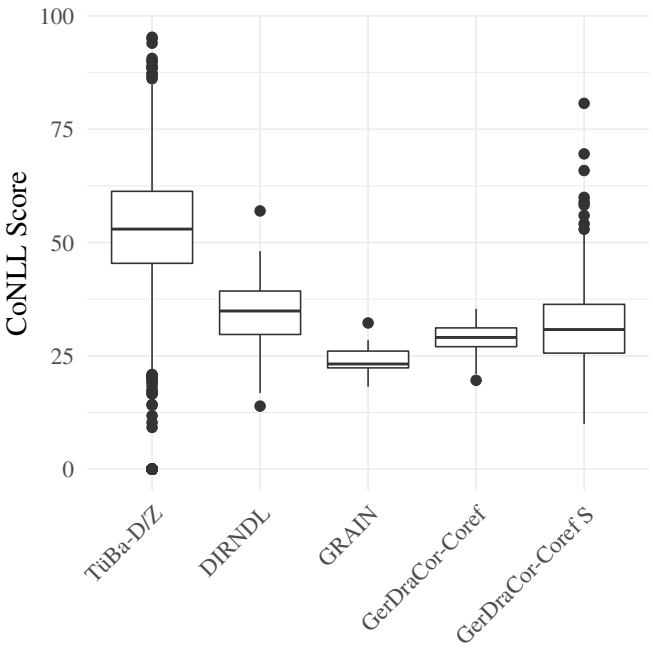

Figure 3: Boxplots showing the CoNLL scores for single documents by applying CorZu on the different corpora.

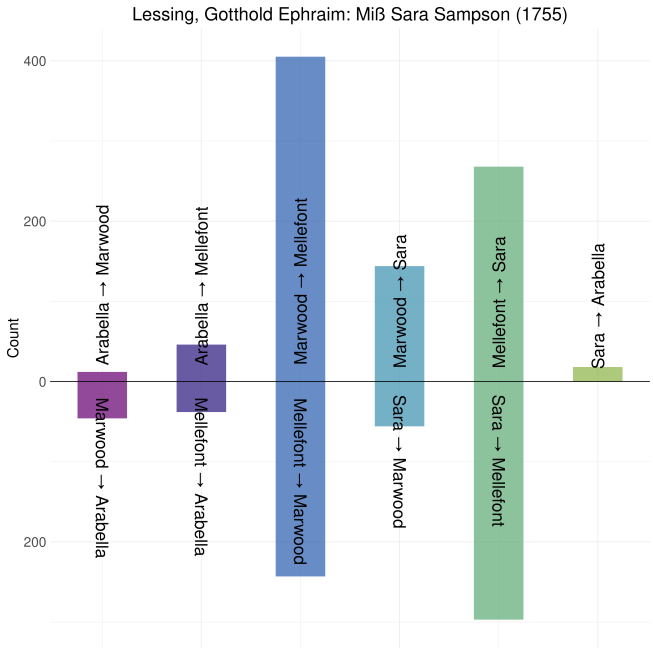

Figure 4: Who mentions whom how often?

6. Applications

Combining coreference annotations with the strong structure of plays allows for a number of interesting analyses. In this section, we showcase a few of these analyses to demonstrate possible applications of coreference on dramatic texts. All showcases are based on manual annotations. Since the speaker of an utterance is directly accessible, it is straightforward to analyse which character mentions which other character how often. This is shown in Figure 4, for the play *Miß Sara Sampson* by Lessing. In the play, the male protagonist Mellefont is torn between his long-time mistress Marwood, with whom he has an illegitimate child called Arabella, and the young Sara. Both Sara and Marwood are mentioned by him roughly 300 times in the play. Of interest in this plot is the difference in the fourth column: Marwood mentions Sara much more often than vice versa, which gives a hint as to the characterisation of both. Arabella only makes a brief appearance on stage. Since Sara learns about Arabella's existence only during the play, she mentions her much less often than Marwood or Mellefont do. Arabella does not know about Sara, so she never mentions her. Finding out which characters are never mentioned by other characters can give insights into the character constellation of a play in general.

Figure 5 shows all individual utterances and mentions of characters over the course of the play. The plot shows that Sara is mentioned continuously in the entire play, even when she is not on stage at all. This 'passive presence' of a character (being mentioned without being on stage) is also relevant for Marwood in the last act, and Sir William (father of Sara) in the third and fourth act. In the last act, Marwood has poisoned Sara and left the country, but is still very present in the characters' mentions. In the third act, Sara receives a letter by her father Sir William, who writes that he is forgiving her for eloping with Mellefont. With some right it can be said that Sir William is the main topic of the act, without being present for the most part of it, which can be visualised by using the coreference annotations.

As, in contrast to other corpora, our annotations aim at full coreference, they include reference relations for objects and abstract entities. Figure 6 shows their distribution over the text. The plot shows that some entities are used throughout the entire text ('Tugend', engl. virtue), while others only have 'local' relevance ('Brief Sir William', engl. letter by Sir William). About one half of the frequently referred things are abstract concepts like virtue or the love someone experiences.

7. Conclusions

In this paper, we have presented an annotated corpus of full coreference chains for German dramatic texts. We also presented an analysis of the coreference phenomenon in dramatic texts and how it behaves differently than in news texts. The availability of an annotated corpus is a starting point for developments of automatic CR systems, we therefore have also made first baseline experiments using the rule-based CR system CorZu. Finally, we have described some of the applications that will be made possible with coreference annotation on dramatic texts.

Based on our analysis of the data and the baseline performance, we see three avenues for improving automatic CR systems: i) Domain adaptation using similar documents: The most gain in combining data sets can be expected when the documents in the data sets are similar. We will therefore explore the use of GRAIN data to support a machine learning CR system. ii) In the baseline experiment, no additional information is taken into account, despite the fact that more information is available: Plays feature a *dramatis personae*, from which names, professions, personal relations of the literary characters can be extracted. To exploit this knowledge for CR, the model needs to handle knowledge that is only present for a subset of the entities in the text. iii) Lastly, the influence of different writing styles on the referential systems still needs to be investigated more thoroughly.

Even though CR remains a challenging task for the time being, it is clear that coreference chains are an important component for the analysis of character-driven texts such as theatre plays. In particular, they allow insight into indirect presentations of characters, e.g., through the speech of other characters. This creates new options for analysing the way character (proto)types are introduced and represented.

8. Acknowledgements

The work described here has been conducted in the QuaDramA project[8], funded by Volkswagen foundation. We thank the foundation for making this possible. We also thank our annotators for their work, and our colleagues and collaboration partners Benjamin Krautter and Marcus Willand for theirs.

Appendix

On overview of all texts used in the corpus with detailed information of author, time of creation and size can be found in Table 8.

[8] https://quadrama.github.io

	Author	Title	Year	Acts	Scenes	Characters	Ch. Speech	Mentions	Ann. Tokens	Ann. Act(s)
1	Gottsched, Johann Christoph	Der sterbende Cato	1732	5	31	11	539 896	652	5193	II
2	Quistorp, Theodor Johann	Der Hypochondrist	1745	5	32	7	719 936	1174	5315	III
3	Krüger, Johann Christian	Die Candidaten oder Die Mittel zu einem Amte zu gelangen	1748	5	50	10	1 369 850	1510	6380	V
4	Mylius, Christlob	Die Schäferinsel	1749	3	30	8	412 110	1198	5219	III
5	Lessing, Gotthold Ephraim	Der Freigeist	1755	5	35	11	894 215	1160	4943	III
6	Lessing, Gotthold Ephraim	Miß Sara Sampson	1755	5	44	11	1 372 668	7542	33 147	I-V
7	Cronegk, Johann Friedrich von	Der Mißtrauische	1760	5	35	7	1 023 890	1356	5731	II
8	Gessner, Salomon	Evander und Alcimna	1762	3	19	14	191 615	621	2803	I
9	Lessing, Gotthold Ephraim	Emilia Galotti	1772	5	43	13	1 090 695	6261	27 896	I-V
10	Wagner, Heinrich Leopold	Die Reue nach der That	1775	6	6	16	146 172	517	2215	V
11	Klinger, Friedrich Maximilian	Sturm und Drang	1777	5	37	13	762 126	1236	5221	V
12	Schink, Johann Friedrich	Hanswurst von Salzburg mit dem hölzernen Gat	1778	4	30	17	486 060	1079	6215	II
13	Schiller, Friedrich	Die Räuber	1781	5	15	25	687 810	8522	49 528	I-V
14	Goethe, Johann Wolfgang	Egmont	1788	5	5	23	146 320	1189	6631	II
15	Kotzebue, August von	Die Indianer in England	1790	3	43	14	1 076 419	2274	9906	III
16	Schiller, Friedrich	Die Piccolomini	1798	5	31	23	698 027	1152	6597	II
17	Hensler, Karl Friedrich	Die Teufelsmühle am Wienerberg	1799	4	58	33	960 016	957	4805	IV
18	Schlegel, August Wilhelm	Jon	1803	5	23	6	451 950	498	3308	III
19	Weißenthurn, Johanna von	Das Manuscript	1817	5	34	15	758 880	775	3384	V
20	Grillparzer, Franz	Ein treuer Diener seines Herrn	1830	5	5	33	104 420	1077	4700	IV
21	Braun von Braunthal, Karl Johann	Faust	1835	5	5	35	101 610	695	3399	IV
22	Freytag, Gustav	Graf Waldemar	1850	5	7	19	182 301	1761	7602	IV
23	Hebbel, Friedrich	Gyges und sein Ring	1856	5	5	8	92 130	777	3465	IV
24	Moser, Gustav von	Das Stiftungsfest	1862	3	52	11	1 126 112	1744	8212	I
25	Anzengruber, Ludwig	Der Meineidbauer	1871	3	26	23	751 972	2245	13 171	I
26	Wildenbruch, Ernst von	Die Quitzows	1888	4	72	56	2 157 984	1416	6432	IV
27	Hauptmann, Carl	Ephraims Breite	1900	5	35	28	842 905	1059	5779	V
28	Wedekind, Frank	König Nicolo oder So ist das Leben	1902	4	4	33	77 284	2148	10 022	I
29	Hofmannsthal, Hugo von	Der Rosenkavalier	1911	3	3	34	44 328	5065	25 501	I-III
30	Rubiner, Ludwig	Die Gewaltlosen	1919	4	69	38	1 940 004	1675	6938	IV
31	Mühsam, Erich	Judas. Ein Arbeiterdrama	1921	5	5	33	171 815	1791	8694	I

Table 8: Overview of the texts used for the corpus. *Character Speech* is measured in tokens. *Year* is the year of creation or the year of publication or first performance, if not available. *Annotated Tokens* denotes the totality of tokens in the annotated act or acts, not only the tokens that are part of a mention. If a drama has no scenes and only acts, the values for *Acts* and *Scenes* are identical.

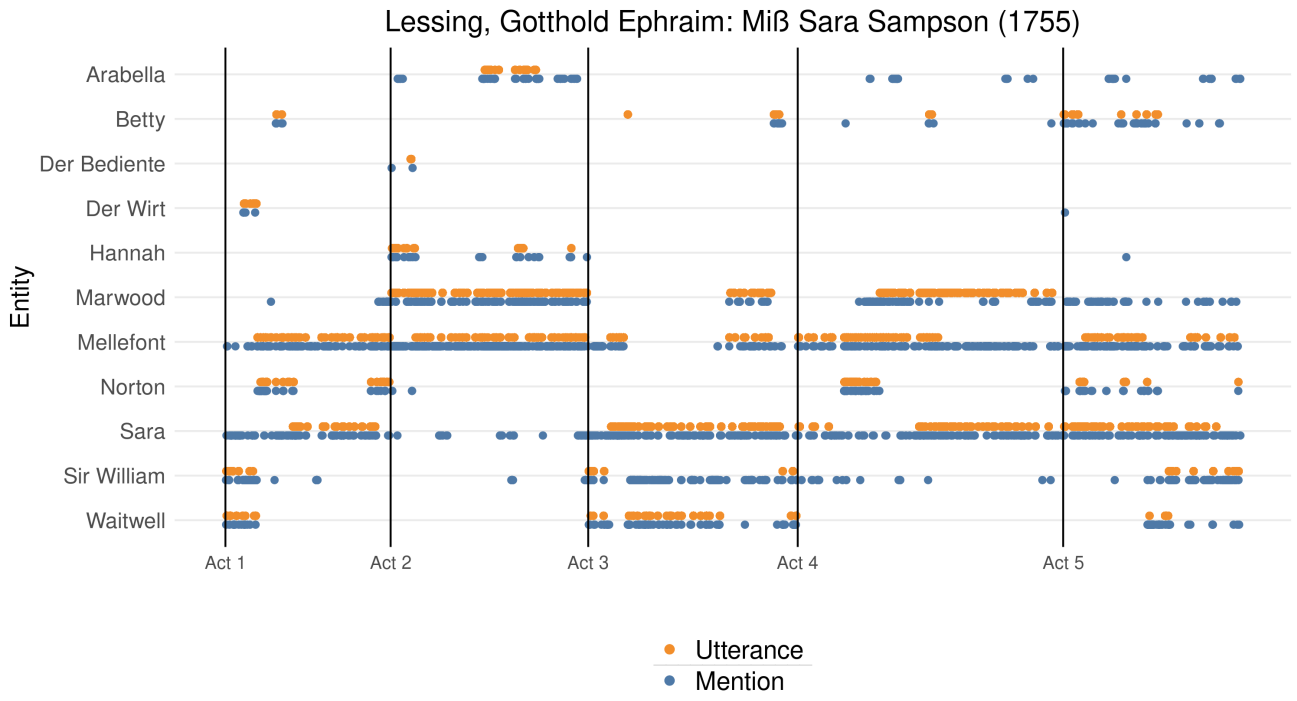

Figure 5: Utterances by and mentions of characters over the course of the play. Self-references (like first person pronouns, etc.) were removed.

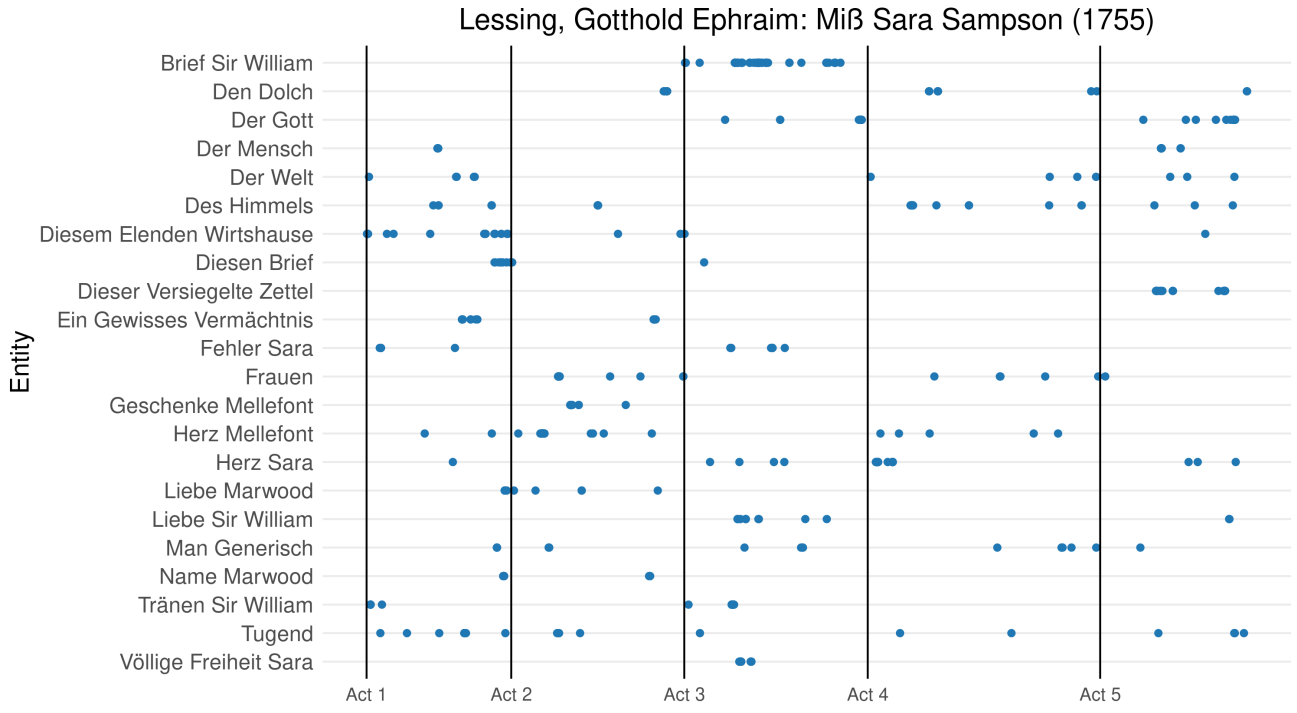

Figure 6: Most mentioned non-characters. Entity designations were chosen by the annotators.

9. Bibliographical References

Bamman, D., Underwood, T., and Smith, N. A. (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 370–379.

Bamman, D., Popat, S., and Shen, S. (2019). An annotated dataset of literary entities. In *Proceedings of NAACL*.

Baumann, S. and Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In Gorka Elordieta et al., editors, *Prosody and Meaning*, volume 25 of *Interface of Explorations*, pages 119–162. Mouton de Gruyter, Berlin.

Björkelund, A. and Kuhn, J. (2014). Learning structured

perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 47–57.

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 3222–3228.

Bullard, J. and Ovesdotter Alm, C. (2014). Computational analysis to explore authors' depiction of characters. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 11–16, Gothenburg, Sweden. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 643–653.

Davis, P. T., Elson, D. K., and Klavans, J. L. (2003). Methods for precise named entity matching in digital collections. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '03, pages 125–127, Washington, DC, USA. IEEE Computer Society.

Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Anaphora with non-nominal antecedents in computational linguistics: A survey. *Computational Linguistics*, 44(3):547–612.

Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., and Weimer, L. (2015). Rule-based coreference resolution in german historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Krug, M., Puppe, F., Reger, I., Weimer, L., Macharowsky, L., Feldhaus, S., and Jannidis, F. (2018). Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers*, 27.

Martschat, S. (2017). *Structured Representations for Coreference Resolution*. Ph.D. thesis, Heidelberg University.

Naumann, K., (2007). *Manual for the Annotation of in-document Referential Relations*. Abt. Computerlinguistik Universität Tübingen.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL 2012, pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 30–35.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 492–501.

Reiter, N., Blessing, A., Echelmeyer, N., Kremer, G., Koch, S., Murr, S., Overbeck, M., and Pichler, A. (2017). CUTE: CRETA Unshared Task zu Entitätenreferenzen. In *Proceedings of DHd*.

Reiter, N. (2018). CorefAnnotator - A New Annotation Tool for Entity References. In *Abstracts of EADH: Data in the Digital Humanities*.

Riester, A. and Baumann, S. (2017). The RefLex Scheme – Annotation guidelines. SinSpeC. Working papers of the SFB 732 Vol. 14, University of Stuttgart.

Rösiger, I. and Kuhn, J. (2016). IMS HotCoref DE: A data-driven co-reference resolver for German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 155–160.

Rösiger, I., Schulz, S., and Reiter, N. (2018). Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138.

Schweitzer, K., Eckart, K., Gärtner, M., Faleńska, A., Riester, A., Rösiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2232.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proccedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 173–180.

Trilcke, P., Fischer, F., and Kampkaspar, D. (2015). Digital network analysis of dramatic texts. In *DH2015 Conference Abstracts*.

Tuggener, D. (2016). *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zürich.

Versley, Y. (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-)reference. In *ESSLLI 2006 Workshop on Ambiguity in Anaphora*, pages 83–89.

Vishnubholta, K., Hammond, A., and Hirst, G. (2019). Are fictional voices distinguishable? classifying character voices in modern drama. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics*

for Cultural Heritage, Social Sciences, Humanities and Literature, pages 29–34.

Yang, J. B., Mao, Q., Xiang, Q. L., Tsang, I. W., Chai, K. M. A., and Chieu, H. L. (2012). Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Comutational Natural Language Learning (EMNLP-CoNLL)*, pages 744–753.

Zhao, S. and Ng, H. T. (2014). Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Mining and Information Analysis (Louhi)*, pages 21–29.

10. Language Resource References

Bamman, D., Popat, S., and Shen, S. (2019). An annotated dataset of literary entities. In *Proceedings of NAACL*.

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 3222–3228.

Krug, M., Puppe, F., Reger, I., Weimer, L., Macharowsky, L., Feldhaus, S., and Jannidis, F. (2018). Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers*, 27.

Reiter, N., Blessing, A., Echelmeyer, N., Kremer, G., Koch, S., Murr, S., Overbeck, M., and Pichler, A. (2017). CUTE: CRETA Unshared Task zu Entitätenreferenzen. In *Proceedings of DHd*.

Schweitzer, K., Eckart, K., Gärtner, M., Faleńska, A., Riester, A., Rösiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 2229–2232.