Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]

Markus Krug[1] Lukas Weimer[2] Isabella Reger[2]
Luisa Macharowsky[2] Stephan Feldhaus[2] Frank Puppe[1]
Fotis Jannidis[2]

[1]Wuerzburg University, Chair of artificial intelligence and applied computer science
[2]Wuerzburg University, Chair of computer philology and modern German literary history

DARIAH-DE Working Papers Nr.27

Krug, M.; Weimer, L.; Reger, I.; Macharowsky, L.; Feldhaus, S.; Puppe, F.; Jannidis, F.: Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2018
(DARIAH-DE working papers 27)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2018-2
URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-2-9

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: In this work, we present DROC, a corpus consisting of 90 fragments of German novels, published between the 17th and 20th century. DROC consists of more than 50.000 carefully annotated character references as well as their coreferences. Additionally, we annotated direct speech instances (contained) in the fragments, along with the corresponding speaker and addressee. The corpus is released in TEI-XML and Apache UIMA .xmi. Both formats are described in this contribution.

Keywords: Figurenreferenzen, Koreferenzen, Roman, Sprecher, Direkte Rede, Korpus, Annotation, deutsch

character references, coreferences, novel, speaker, direct speech, corpus, annotation, German

Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]

Markus Krug[1] Lukas Weimer[2] Isabella Reger[2]
Luisa Macharowsky[2] Stephan Feldhaus[2] Frank Puppe[1]
Fotis Jannidis[2]

[1]Wuerzburg University, Chair of artificial intelligence and applied computer science
[2]Wuerzburg University, Chair of computer philology and modern German literary history

DARIAH-DE
Working Papers

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, Fotis Jannidis: "Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]". *DARIAH-DE Working Papers* No. 27. Göttingen: DARIAH-DE, 2017.
URN: urn:nbn:de:gbv:7-dariah-2018-2-9.

The *DARIAH-DE Working Papers* are published by Mirjam Blümm, Thomas Kollatz, Stefan Schmunk and Christof Schöch.

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

DROC was created as part of the project Kallimachos supported by the Federal Ministry of Education and Research (BMBF).

Abstract

In this work, we present DROC, a corpus consisting of 90 fragments of German novels, published between the 17th and 20th century. DROC consists of more than 50.000 carefully annotated character references as well as their coreferences. Additionally, we annotated direct speech instances (contained) in the fragments, along with the corresponding speaker and addressee. The corpus is released in TEI-XML and Apache UIMA .xmi. Both formats are described in this contribution.

Schlagwörter

Figurenreferenzen, Koreferenzen, Roman, Sprecher, Direkte Rede, Korpus, Annotation, deutsch

Keywords

character references, coreferences, novel, speaker, direct speech, corpus, annotation, German

Contents

1 Motivation

Nowadays, large collections of literary texts are available in many languages and enable new approaches to literary studies like network analysis, topic modeling or stylometry. Especially the analysis of networks of literary characters has become either a goal in itself or a building block in larger contexts (Elson, Dames, and McKeown 2010; Park et al. 2013; Trilcke 2013; Moretti 2011). In such networks, characters usually constitute the nodes while their interaction, for example the amount of conversation, is modeled as edges, often using the amount of interaction as weight. In order to create such networks, the first step is to find all references to characters in a text. However, in order to detect all character references to an entity, it is not sufficient to apply a state of the art named entity recognizer (NER) such as Stanford NER (Finkel, Grenager, and Manning 2005). In a literary text, a reference can appear in one of three broad syntactic categories: 1) as a proper noun; 2) as a nominal phrase; 3) as a pronoun. Detecting categories 2) and 3) is usually beyond the scope of named entity recognition. Furthermore, current state of the art NER were trained on newspaper articles, which means another gap to bridge with regard to the domain of novels. The mere identification of character references is not sufficient to create a complete pipeline with the goal of extracting social networks from literary text; it is required to resolve each reference to the entity it refers to in the fictional world (of a text/work), a process called coreference resolution.

Literary texts usually contain a high proportion of direct speech compared to newspaper texts. The references within direct speech, in particular pronouns of first and second person singular, correspond with the speakers or addressed entities. Therefore, speaker resolution is an essential component for a coreference resolution in literary texts.

Usually automatic tools require manually annotated data for training or at least for evaluation purposes. This work presents the corpus DROC (Deutsches ROman Corpus)[1], which consists of 90 manually annotated fragments of German novels and includes the following annotations:

1. Each character reference has been marked.
2. Each reference was assigned to one of four subcategories.
3. Each character reference has an assigned entity-identifier.
4. Each direct speech has been manually annotated.
5. The speaker and addressee of each direct speech have been manually marked.

To the best of our knowledge, there is no comparable corpus available to the academic community in the domain of literary texts, especially for German. DROC comprises about 393.000 annotated tokens with more than 50.000 labelled character references.

The paper is structured as follows: First, a brief overview of existing corpora for named entities and coreference resolution is given, followed by a description of the textual sources of the fragments. We continue with a detailed description of our annotation guidelines and the annotation process, including the inter-annotator agreement (IAA). We then explain the two formats in which we release our data and conclude with a brief description of the statistics found in our corpus.

[1] https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release

2 Related Work

Comparing our corpus in the field of coreference resolution yields a number of similar resources – though none in the domain of (German) literary texts. In this section we restrict ourselves to German and English corpora available for academic research, starting with the latter. The best known corpora for English were released in the scope of the MUC-6 and MUC-7 conferences (Grishman and Sundheim 1996; Chinchor and Robinson 1997). Those corpora each comprise about 30.000 tokens and contain articles from the wall street journal (WSJ) and airplanes crashes. The corpus released for ACE-2005 (Walker et al. 2006) had about 400.000 annotated tokens and contains a mix of news, blog and web articles. With about 1.500.000 tokens, OntoNotes 5.0 (Weischedel et al. 2013) currently is the largest available resource for coreference resolution and consists of news articles, conversations and web articles. For German, there are currently two corpora available. The first is the Potsdam commentary corpus (Stede 2004), comprising 33.000 tokens derived from 176 newspaper commentaries. The other resource for German coreference resolution is the TüBa-D/Z corpus, released by the university of Tübingen (Telljohann et al. 2006). It is made of about 3.400 newspaper articles, with about 1.500.000 tokens. This overview shows that there is currently no resource for (German) literary texts and most articles of the aforementioned resources tend to be much shorter than an average novel – yielding new phenomena to explain with statistical methods, therefore underlining the importance of the release of DROC, a resource comprising 90 fragments of German novels, published between 1650 and 1950. There is no other resource that has manually marked direct speech passages along with the respective speaker and addressee, of which DROC has more than 2.000.

3 Description of the Textual Sources

The texts of the novels which are the basis for our corpus come from a large collection of German literary texts available as full-texts, part of the TextGrid repository[2]. The texts found in this repository are part of one of the first large-scale digitization projects in the German language. The digitization was undertaken in separate steps by a commercial company, Directmedia, over the course of ten years, which sold digital texts on CDs and DVDs. It is important to understand that the TextGrid collection comprises two different groups of texts: The first group, by far the largest, consists of canonized texts of German literature. These are usually based on scholarly editions used for decades by academics. In most editions the writing has been normalized: In our context this means mainly that "th" has been replaced by "t" (for example "Tür" instead of "Thür") and "ey" by "ei" (for example "sei" instead of "sey").

The second group has been part of a collection called *Deutsche Literatur von Frauen* (German literature by women) which tried to collect as much literature by female authors as possible. As many of these texts are not part of the literary canon, there are no scholarly editions and the creators of the collection had to base their digital texts on first prints or unchanged reprints of first prints. Therefore, the collection is not balanced or representative for the literary production of the period it covers. The collection is copyright free and has been released in TEI markup on TextGridRep with a Creative Commons-license (CC-BY).

[2] https://textgridrep.org/

The texts of the DROC have not been standardized and offer a variability of orthographic norms. The variability is especially marked in nine texts which have been published between 1650 and 1800. If an analysis of only standardized texts is required, then these can be filtered out via the metadata of the corpus.

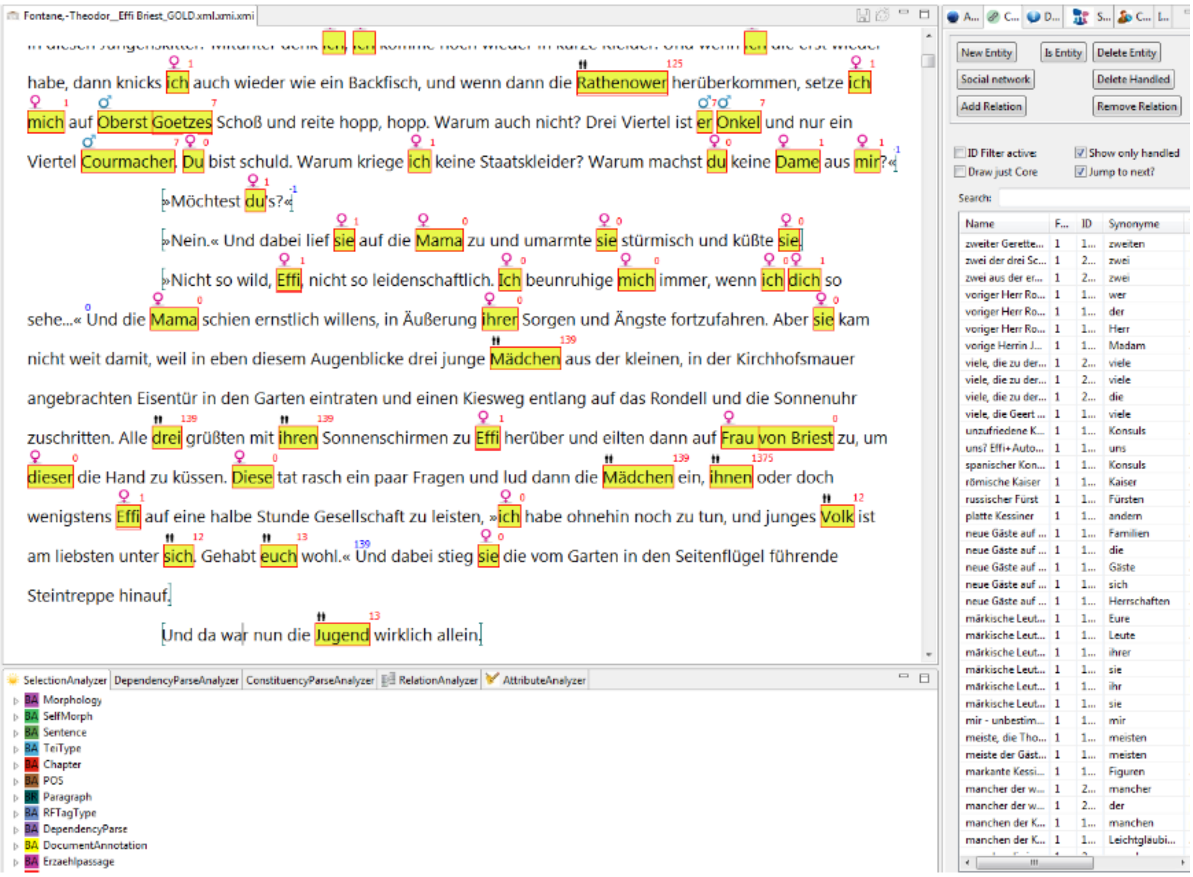
4 Creation of the Corpus

The corpus DROC comprises 90 fragments of different novels. The novels were randomly selected from 450 available novels of the TextGrid repository. We applied the Apache OpenNLP sentence detection component (Morton et al. 2005), trained on the TIGER corpus (Brants et al. 2004), to annotate sentence boundaries in the selected novels. Then, for each novel, we randomly sampled a sentence index and extended the fragment in both directions until the beginning of a chapter and the end of a chapter was reached. In some occasions, where no structural information about chapters was available, our annotators manually selected sentences that indicate the beginning of a coherent passage in the novel and therefore simulate an artificial border. The resulting fragments had an average length of 201 sentences. We implemented this procedure because we wanted to make sure that for all references either the proper nouns or the common nouns were part of the selected sentences.

The annotation process can be depicted as follows: The document were preprocessed with a rule based script, developed with UIMA RUTA (Kluegl et al. 2016), in order to generate suggestions that both of our annotators could later either accept or change. Therefore, our corpus was created semi-automatically with initial support. We annotated our novels in ATHEN[3], a selfmade desktop application based on the eclipse RCP4 framework. The perspective for character reference annotation can be seen in figure 1.

After our annotators finished their pass over the documents, resulting inconsistencies were resolved together in order to get a clean version of the annotations.

[3] https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen



habe, dann knicks ich auch wieder wie ein Backfisch, und wenn dann die Rathenower herüberkommen, setze ich mich auf Oberst Goetzes Schoß und reite hopp, hopp. Warum auch nicht? Drei Viertel ist er Onkel und nur ein Viertel Courmacher. Du bist schuld. Warum kriege ich keine Staatskleider? Warum machst du keine Dame aus mir?«

»Möchtest du's?«

»Nein.« Und dabei lief sie auf die Mama zu und umarmte sie stürmisch und küßte sie.

»Nicht so wild, Effi, nicht so leidenschaftlich. Ich beunruhige mich immer, wenn ich dich so sehe...« Und die Mama schien ernstlich willens, in Äußerung ihrer Sorgen und Ängste fortzufahren. Aber sie kam nicht weit damit, weil in eben diesem Augenblicke drei junge Mädchen aus der kleinen, in der Kirchhofsmauer angebrachten Eisentür in den Garten eintraten und einen Kiesweg entlang auf das Rondell und die Sonnenuhr zuschritten. Alle drei grüßten mit ihren Sonnenschirmen zu Effi herüber und eilten dann auf Frau von Briest zu, um dieser die Hand zu küssen. Diese tat rasch ein paar Fragen und lud dann die Mädchen ein, ihnen oder doch wenigstens Effi auf eine halbe Stunde Gesellschaft zu leisten, »ich habe ohnehin noch zu tun, und junges Volk ist am liebsten unter sich. Gehabt euch wohl.« Und dabei stieg sie die vom Garten in den Seitenflügel führende Steintreppe hinauf.

Und da war nun die Jugend wirklich allein.

□ ID Filter active: ☑ Show only handled
□ Draw just Core ☑ Jump to next?

Search:

Name	F...	ID	Synonyme
zweiter Gerette...	1	1...	zweiten
zwei der drei Sc...	1	2...	zwei
zwei aus der er...	1	2...	zwei
voriger Herr Ro...	1	1...	wer
voriger Herr Ro...	1	1...	der
voriger Herr Ro...	1	1...	Herr
vorige Herrin J...	1	1...	Madam
viele, die zu der...	1	2...	viele
viele, die zu der...	1	2...	viele
viele, die zu der...	1	2...	die
viele, die Geert ...	1	1...	viele
unzufriedene K...	1	1...	Konsuls
uns? Effi+Auto...	1	1...	uns
spanischer Kon...	1	1...	Konsuls
römische Kaiser	1	1...	Kaiser
russischer Fürst	1	1...	Fürsten
platte Kessiner	1	1...	andern
neue Gäste auf ...	1	1...	Familien
neue Gäste auf ...	1	1...	die
neue Gäste auf ...	1	1...	Gäste
neue Gäste auf ...	1	1...	sich
neue Gäste auf ...	1	1...	Herrschaften
märkische Leut...	1	1...	Eure
märkische Leut...	1	1...	Leute
märkische Leut...	1	1...	ihrer
märkische Leut...	1	1...	sie
märkische Leut...	1	1...	ihr
märkische Leut...	1	1...	sie
mir - unbestim...	1	1...	mir
meiste, die Tho...	1	1...	meisten
meiste der Gäst...	1	1...	meisten
markante Kessi...	1	1...	Figuren
mancher der w...	1	2...	mancher
mancher der w...	1	2...	der
manchen der K...	1	1...	manchen
manchen der K...	1	1...	Leichtgläubi...

Figure 1: The main user interface for the coreference annotation in ATHEN. The left depicts the main editor which shows the currently opened document (in this case "Effi Briest" by Theodor Fontane (Fontane 1973)) and on the right, there is the view to accept or change a selected annotation.

5 Annotation Guidelines

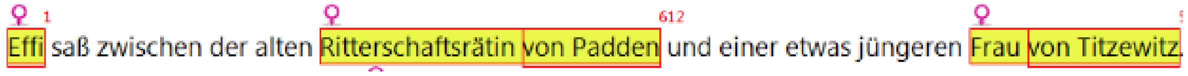
We present our annotation guidelines in a three step process. At first, we describe which references were annotated, followed by the description of the resulting phenomena we had to deal with in terms of coreference resolution and some borderline cases in DROC. We conclude the guidelines section with our guidelines for the annotation of direct speech utterances along with their speakers and addressees.

5.1 Annotated Character References

The annotation of character references follows a single rule:

Mark every text snippet in the novel that references a (literary) character.

Furthermore, we decided not to mark the complete nominal phrase surrounding the reference and only marked the heads instead.



Effi saß zwischen der alten Ritterschaftsrätin von Padden und einer etwas jüngeren Frau von Titzewitz.

Figure 2: A text snippet, taken from "Effi Briest" by Theodor Fontane. (Effi was sitting between the old knighthood councilor von Padden and a somewhat younger Mrs. von Titzewitz.)[4] *The picture shows the marked head "Ritterschaftsrätin von Padden" which is embedded in the nominal phrase "der alten Ritterschaftsrätin von Padden". Analogously, the snippet "Frau von Titzewitz" is marked as the head of the phrase "einer etwas jüngeren Frau von Titzewitz"*

Following this rule, the resulting phrases can be classified into the following subcategories:

1. **Proper noun**

Proper nouns, for example forenames, surnames or family names. These names can also refer to entities that are not part of the fictional world (e.g. another author, historic persons, etc.) In our schema, the text snippets representing proper nouns are marked as "Core". Sometimes a "Core" snippet is only a part of a reference (As shown in figure 2, where "von Padden" is the Core snippet of "Ritterschaftsrätin von Padden").

2. **Heads of common noun phrases**

A head of a common noun phrase can be an arbitrary composite consisting of:

- Occupational titles (e.g. "Bäcker" – "baker")
- Relational expressions (e.g. "Mutter" – "mother")
- Gender terms (e.g. "Mann" – "man")
- Different titles (e.g."Graf" – "earl")
- Action terms (e.g. "Spaziergänger" – "stroller")
- Defamations (e.g. "Idiot" – "idiot")
- Substantival verbs (e.g. "Rufende" – "shouter")
- Substantival adjectives (e.g. "Schöne" – "beauty")

This listing is not complete, showing the complexity of this class. Annotations of this kind were marked as "AppTdfW" (Appellativ, Teil der fiktionalen Welt) if they are part of the fictional world or as "AppA" (Appellativ, Abstraktum) if they refer to generic or abstract entities that are not part of the fictional world.

[4] For better clarity, all German citations were translated into English.

3. **Pronouns**

This category, marked as "Pron", comprises all sorts of pronouns, the most prominent examples being:

- Personal pronouns (e.g. "er", "sie," – "he", "she")
- Possessive pronouns (e.g. "seine", "ihre" – "his", "her")
- Reflexive pronouns (e.g. "sich" – "himself", "herself", "themselves")
- Relative pronouns (e.g. "der", "die" – "who")

For each resulting character reference, we marked the following features:

- Type: one of "Core","Pron","AppTdfW" or "AppA", as described above
- Range: (used only for Cores) span of character offsets for the identification of the core text snippet
- Number: singular or plural
- ID: a unique identifier for each entity appearing in the text, used to represent coreference.
- Pseudo: This means that the person is mentioned in the text, but does not take part in the action or does not exist. An example for this case is "War nicht auch Cromwell erst in hohem Alter nach vergeudeter Jugend erweckt worden zum Dienste Gottes?" ("Only in his old age and after wasting his youth Cromwell was called to serve God, wasn't he?" (Bleibtreu 1888)). Both, *Cromwell and Gott*, are identified as pseudos, because both are not taking part in this novel's action.
- Uncertain: A boolean flag that can be set by the annotator if the decision is unclear.

5.2 Annotated Coreferences

With the definition of the character references our annotators had the task to assign a unique identifier to each entity in the text, and to reuse this Id for each mention of an entity. To enable an easier comparison of DROC to existing corpora with annotated coreference we discuss a selected list of coreferential linguistic phenomena and elaborate whether we marked them as coreferent or not.

Coordination and plural references

Plural references are included if the phrase that is required to mark them does not consist of multiple smaller references. Therefore our annotations are not hierarchical.

Split Antecedents

Split antecedents, that is plural references which can only be mapped to more than one reference, are not marked. (e.g. "sie" in "Effi und Innstetten planten eine Reise, sie...")[5]

[5] Effi and Innstetten planned a trip, they...

Expletives

Expletives, such as "It" in "It is raining" are not included in DROC.

Appositions and Predicatives

Appositional references (e.g. "Otto, ihr ältester Sohn,..."[6]) as well as references in predicative position (e.g. "Er ist Bäcker"[7] are (usually) marked as coreferent.)

Bridging Anaphora

Bridging anaphora, such as the relation between tyre and bicycle in "I bought a bicycle. A tyre was already flat", are not marked within DROC.

Discourse

The information whether an entity is discourse new has to be parsed from the ID feature of the references.

We conclude this section with a prototypical example taken from DROC:

"Bekannte (ID=1, AppTdfW, Plural) traten zu ihnen (ID=2, Pron, plural) heran und das Gespräch war unterbrochen. Michael (ID=3, Core) fuhr mit Käthe (ID=4, Core) in einer offenen Droschke, in der milden Märznacht, nach Hause. Ihre (ID=4, Pron) Blicke hingen am gestirnten Himmel, die seinen (ID=3, Pron) an ihrem (ID=4, Pron) Antlitz. In Beiden (ID=2, Pron, plural) klang die Stimmung von Tristan (ID=5, Core, pseudo) und Isolde (ID=6, Core, pseudo) nach."[8]

5.3 Borderline Cases

During the annotation process, some borderline cases were discovered and will be discussed in the following.

The character references are usually human beings, albeit in some cases animals or even other things can play an important role as an agent for the plot of a story. In such instances these protagonists will also be annotated, as shown in the following example: "Eine Woche später und der Alraun war in seiner Art völlig ausgewachsen, etwa dreieinenhalben Fuß hoch;"[9] Mandrake is labeled as an entity, because in the course of the story the plant comes to life, is named *Cornelius Nepos* and is able to move and talk. Therefore it becomes an important agent.

Sometimes a novel is partly interrupted by a stichomythia, which then resembles a drama. In this incident the names, which introduce the speech, will be regarded as entities e.g. "Einsiedel: Wie heißest du? Simpl.: Ich heiße Bub."[10] Both, Einsiedel and Simpl. (short for Simplicissimus), are marked.

[6] Otto, her oldest son

[7] He is a baker

[8] "Friends came up to them and made their conversation stop. Michael went home with Käthe in an open hansom through the mild March night. Her eyes focussed on the starry sky, his eyes focussed on her countenance. Both of them reveled in the mood of Tristan and Isolde." (Dohm 1894)

[9] "One week later the mandrake was fully grown, approximately three and a half feet tall;" (Arnim 1962)

[10] "Einsiedel: What is your name?; Simpl.: My name is boy." (Grimmelshausen 1956)

In rare cases a definitive decision is not possible – due to a lack of knowledge or imprecise references. In the following example, it is not clearly determinable what "man" ("someone") refers to. It could refer to a concrete person or group, to human species generally or to nothing at all. "Sollte ich etwa mit gebundenen Händen immer weiter zusehen, wie man mir mein Leben zertritt, bis die Jugend vorbei ist und alles zu spät?"[11]

5.4 Annotated Direct Speech

Direct speech passages and every text section enclosed by single or double quotation marks are annotated. Such annotations range from opening quotation marks till closing ones – both included. In most cases these are french quotation marks (»«), infrequently dashes. To every annotation one (or in rare cases more) speaker and addressed character references is assigned. If it was not possible to determine who speaks or who is addressed, they are marked as "unknown".

The annotation process obeys strict rules. If speaker and addressed reference are connected to the relevant direct speech with a communication verb, then these entities are labelled. If not, we looked for direct addressees within the direct speech which are not pronouns (e.g. "..., my dear friend"). If something like that does not exist either, the last mention of speaker and/or addressed person which lies outside of direct speeches was annotated, independent of being a noun or pronoun.

Apart from real direct speeches every text section within quotation marks is annotated. This might be names of places, quotations or thoughts. (...unten in der »Gelben Straße«...[12])(Dauthendey 1923) In this case, category, which is set to "directspeech" by default, was changed. The following categories are defined: "thought", "citation" (e.g. quotations of absent characters or of other fictional works), "fictionalspeech (speeches of a text entity that is not labeled as a reference, e.g. "my heart says...", "roses say..."), "name" (e.g. place names) and "other" (if further classification is not possible, e.g. a word highlighted with quotation marks by the author for accentuation).

Sometimes direct speeches are not marked up by quotation marks. In rare cases, direct speeches are labeled by dashes or even without any marker. These cases have been annotated either way.

6 Inter-Annotator Agreement

There are multiple ways to measure an inter-annotator agreement (IAA). We used 12 documents that were labelled by both our annotators and measured the IAA for character reference annotation and for coreference resolution. For evaluating the quality of the character reference annotation we only took the annotated span into account and calculated Cohens Kappa (Cohen 1960).

We did this on a per token basis and converted the output of each annotator into a sequence of B-I-O labels. A measurement on a per token basis awards our annotator for not marking a token as a character reference on top of the rewards for marking the same span. This yields 31.185 instances and resulted in a Kappa κ of 94.3%

[11] "Should I really keep on watching with tied hands, how someone destroys my life until the youth faded and everything is too late?" (Reventlow 1980)

[12] ...down in »Yellow Street«...

On the same documents, we measured the IAA of the assigned coreference clustering with MUC-6 and B-Cube scores (Luo 2005). Our evaluation resulted in a MUC-6 F1 of 88.5% and a B-Cube F1 of 69%. Since both evaluation metrics require the amount of references to be equal we added references if necessary and treated them as singletons. The B-Cube metric punishes singleton clusters which explains the much lower score compared to the MUC evaluation. Removing unmatchable annotations yields a MUC-6 F1 of 92.4% and a B-Cube F1 of 76%. For the final version of DROC, the documents were annotated by one annotator and afterwards both annotators revised the documents together to guarantee a corpus of high quality.

The code for the evaluation, as well as the documents that were used for the measurements can be downloaded from DROCs git repository.[13]

7 Release Formats

DROC is available in two formats. The first format is XMI. These files are standard for Apache UIMA and come with a typesystem definition, required to open DROC. The second format is TEI-XML (TEI-Consortium 2017). This section gives a brief overview over the representation used within these formats. A more thorough definition can be found on the git repository of the release.

7.1 XMI-format

In the UIMA format, each annotation is stored with at least two features, a begin indicating the character offset where the annotation starts and an end feature indicating where an annotation ends. Additionally, each annotation has its own type, defined in a separate descriptor xml-file. For DROC, we defined two types:

Type NamedEntity

This type represents a character reference. One annotation is created for every character reference. Table 1 gives an overview of the features used.

Table 1: Overview over the UIMA type: NamedEntity used in the .xmi encoding for DROC

Featurename	Range	Featurevalues	Description
ID	String	any	A unique id, referring to the entity this reference belongs to.
Pseudo	String	true	false
Numerus	String	pl (plural) or si (singular)	A string referring to the number of the reference.
NEType	String	AppTdfW|AppA|Core|Pron	The type of the reference.
Uncertain	String	true|false	A flag, indicating whether our annotators were uncertain about their decision.
CoreRange	String	[from:to] or null	A string, used for core references to show the text that is a proper noun.

[13] https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release

Type DirectSpeech

This type represents an instance of a direct speech. It stores information about speaker and addressee as well as its type. Table 2 shows the structure of the type.

Table 2: Overview over the UIMA type: DirectSpeech used in the .xmi encoding for DROC

Featurename	Range	Featurevalues	Description
Speaker	Annotation		An annotation of type NamedEntity, depicting the speaker.
SpokenTo	Annotation		An annotation of type NamedEntity, depicting the addressee.
category	String	directspeech|thought|citation|ficitionalspeech|name|other	The type of the direct speech annotation.

7.2 TEI-XML

The second format DROC is available in is TEI-XML. Within the `<body>` element of each document, a sequence of elements is added for each token. Character references have been encoded using the `<persName>` element and direct speech utterances using the `<quote>` element with embedded speech elements `<sp>` that direct to the speaker of the utterance. Sentence and paragraph borders have been added as virtual elements at the end of each document. We used the `prev` attribute of the element `<persName>` to refer to the first appearance of the corresponding entity of a character reference. Speakers have been encoded using the `who` attribute that refers to the `xml:id` of the speaking character reference.

8 Corpus Statistics

DROC contains 90 fragments of different novels. The corpus comprises about 393.000 tokens, determined by the tokenizer script of the TreeTagger (Schmid 2013). On average each fragment consists of 4.368±2.334 tokens and 202±131 sentences. We manually annotated 52.079 character references with the majority of 65% being pronouns (34.060). About 23% (12.005) of the references belong to the type "appellative" and the remaining 12% (6.013) are "Core" references. These 52.081 references are clustered into 5.288 entities, therefore on average 10 references per entity and 59±31 entities per document. Compared to the statistics from the study in (Kabadjov 2007), pronouns in DROC appear more frequently, with a proportion of 65% compared to 44% evaluated by Kabadjov, with the amount of proper nouns being almost constant with a small increase from 10% to 12% in DROC. 35 of those fragments were written by female authors and the remaining 55 were written by male authors resulting in a slightly imbalanced 40%-60% gender ratio. As shown by table 3, most novels of DROC were published between 1801 and 1900.

Table 3: Overview of the amount of novels published during an epoch of 50 years. It can be seen that most novels were published during the 19th century.

Epoch	1651 - 1700	1701 - 1750	1751 - 1800	1801 - 1850	1851 - 1900	1901 - 1950	1951- 2000
Amount novels	2	3	4	31	35	14	1

9 Conclusion

In this work we presented DROC, a corpus with manual annotations of character references and direct speeches alongside their speakers and addressees. DROC is the only resource of this kind for German-language literary texts and the only resource that includes the assignment of speakers and addressees to the corresponding direct speech.

10 License

The corpus is licensed under the Creative Commons license CC-BY[14].

If you use the corpus, please use the following quote:

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, Fotis Jannidis: "Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]". DARIAH-DE Working Papers Nr. 27. Göttingen: DARIAH-DE, 2018. URN: urn:nbn:de:gbv:7-dariah-2018-2-9

[14] https://creativecommons.org/licenses/by/4.0/deed.de

Bibliography

Arnim, Achim von. 1962. *Sämtliche Romane Und Erzählungen. Bde. 1–3*. Edited by Walther Migge. München: Carl Hanser Verlag.

Bleibtreu, Karl. 1888. *Größenwahn. Pathologischer Roman*. Leipzig: Wilhelm Friedrich.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. "TIGER: Linguistic Interpretation of a German Corpus." *Research on Language and Computation* 2 (4). Springer: 597–620.

Chinchor, Nancy, and Patricia Robinson. 1997. "MUC-7 Named Entity Task Definition." In *Proceedings of the 7th Conference on Message Understanding*. Vol. 29.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1). Sage Publications: 37–46.

Dauthendey, Max. 1923. *Lingam. Zwölf Asiatische Novellen.* München: Albert Langen.

Dohm, Hedwig. 1894. *Wie Frauen Werden*. Breslau: Schlesische Buchdruckerei, Kunst- und Verlags-Anstalt v. S. Schottlaender.

Elson, David K, Nicholas Dames, and Kathleen R McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138–47. Association for Computational Linguistics.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 363–70. Association for Computational Linguistics.

Fontane, Theodor. 1973. *Romane Und Erzählungen in Acht Bänden. Bd. 7*. Edited by Peter Goldammer, Gotthard Erler, Anna Golz, and Jürgen Jahn. Berlin; Weimar: Aufbau.

Grimmelshausen, Hans Jakob Christoffel von. 1956. *Der Abenteuerliche Simplicissimus*. Edited by Alfred Kelletat. München: Winkler-Verlag.

Grishman, Ralph, and Beth Sundheim. 1996. "Message Understanding Conference-6: A Brief History." In *Coling*, 96:466–71.

Kabadjov, MA. 2007. "Task-Oriented Evaluation of Anaphora Resolution." PhD thesis, Ph. D. thesis, University of Essex, Colchester, UK.

Kluegl, Peter, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. 2016. "UIMA Ruta: Rapid Development of Rule-Based Information Extraction Applications." *Natural Language Engineering* 22 (01). Cambridge Univ Press: 1–40.

Luo, Xiaoqiang. 2005. "On Coreference Resolution Performance Metrics." In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 25–32. Association for Computational Linguistics.

Moretti, Franco. 2011. "Network Theory, Plot Analysis." *New Left Review*.

Morton, Thomas, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. "OpenNLP: A Java-Based Nlp Toolkit."

Park, Gyeong-Mi, Sung-Hwan Kim, Hye-Ryeon Hwang, and Hwan-Gue Cho. 2013. "Complex System Analysis of Social Networks Extracted from Literary Fictions." *International Journal of Machine Learning and Computing* 3 (1). IACSIT Press: 107.

Reventlow, Franziska Gräfin zu. 1980. *Autobiographisches. Ellen Olestjerne. Novellen, Schriften, Selbstzeugnisse.* Edited by Else Reventlow. München: Langen Müller.

Schmid, Helmut. 2013. "Probabilistic Part-Ofispeech Tagging Using Decision Trees." In *New Methods in Language Processing*, 154. Routledge.

Stede, Manfred. 2004. "The Potsdam Commentary Corpus." In *Proceedings of the 2004 Acl Workshop on Discourse Annotation*, 96–102. Association for Computational Linguistics.

TEI-Consortium. 2017. "TEI P5: Guidelines for Electronic Text Encoding and Interchange." March. http://www.tei-c.org/Guidelines/P5/.

Telljohann, Heike, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2006. "Stylebook for the Tübingen Treebank of Written German (TüBa-d/Z)." In *Seminar Fur Sprachwissenschaft, Universitat Tubingen, Tubingen, Germany*.

Trilcke, Peer. 2013. "Social Network Analysis (Sna) als Methode Einer Textempirischen Literaturwissenschaft." *Philip Ajouri/Katja Mellmann/Christoph Rauen (Hg.), Empirie in Der Literaturwissenschaft, Münster*, 201–47.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. "ACE 2005 Multilingual Training Corpus." *Linguistic Data Consortium, Philadelphia* 57.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, et al. 2013. "Ontonotes Release 5.0 Ldc2013t19." *Linguistic Data Consortium, Philadelphia, PA*.