Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena

Ina Rösiger, Sarah Schulz and Nils Reiter
Institute for Natural Language Processing
University of Stuttgart
{roesigia,schulzsh,nils.reiter}@ims.uni-stuttgart.de

Abstract

Coreference resolution is the task of grouping together references to the same discourse entity. Resolving coreference in literary texts could benefit a number of Digital Humanities (DH) tasks, such as analyzing the depiction of characters and/or their relations. Domain-dependent training data has shown to improve coreference resolution for many domains, e.g. the biomedical domain, as its properties differ significantly from news text or dialogue, on which automatic systems are typically trained. This also holds for literary texts. We therefore analyze the specific properties of coreference-related phenomena on a number of texts and give directions for the adaptation of annotation guidelines. As some of the adaptations have profound impact, we also present a new annotation tool for coreference, with a focus on enabling annotation of long texts with many discourse entities.

1 Introduction

Noun phrase (NP) coreference resolution is the task of determining which noun phrases in a text or dialogue refer to the same discourse entities (Ng, 2010). Resolving noun phrases can benefit downstream applications revolving around automatic text understanding, such as summarization and textual entailment. Furthermore, coreference resolution is an important stepping stone for analyzing narrative texts, as many such texts are built around characters – a frequently mentioned discourse entity. Coreference resolution therefore has applications within computational literary studies, which applies approaches such as network analysis, plot sentiment analysis, or distinguishing character types.

Corpora annotated with coreference information often consist of news texts or dialogues, with a few exceptions mostly in the biomedical and scientific domain. Literary texts differ from news texts and dialogues to a great extent, as their purpose is not to transfer information as it is the principle task of a newspaper, but rather to provide poetic descriptions and good storytelling. Literary texts have been shown to comprise sophisticated language, with a rich vocabulary, direct and indirect speech and a large set of syntactic constructions (van Cranenburgh and Bod, 2017). For this domain, annotated data are scarce: To the best of our knowledge, there are only a few corpora containing coreference annotations in literary texts (cf. Section 2), for which no domain-specific considerations have been described.

This paper aims at providing a theoretical yet empirically-tested basis for the annotation of coreference information in literary texts. We achieve this by a) analyzing the way coreference-related phenomena behave differently in literary texts and by b) uncovering new coreference-related phenomena that are specific to literary texts. To this end, we conduct an annotation study that uses existing annotation guidelines as a starting point, and refines them in an iterative process. Our focus here is to provide an analysis of the properties of literary texts that need to be considered when creating annotation guidelines and subsequently corpora. On a meta level, this study also gives insights into the dependency of annotation guidelines on the domain and/or text type.

The paper is structured as follows. Section 2 reviews existing corpora annotated with coreference. Section 3 presents the literary-specific considerations which are relevant for coreference annotations

Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 129–138
Santa Fe, New Mexico, USA, August 25, 2018.

and Section 4 overviews how we propose to update existing annotation guidelines. Section 4 moreover describes a new annotation tool, which was designed to handle the properties we considered crucial for coreference annotation in our domain. Finally, we conclude in Section 5.

2 Existing Corpora

Coreference resolution is a highly active NLP area, with many previous annotation efforts, which enabled the creation of well-performing automatic resolvers, such as the state-of-the-art neural coreference resolver by Clark and Manning (2016). Automatic tools are typically trained on the benchmark dataset OntoNotes, which spans multiple genres (mostly newswire, broadcast news, broadcast conversation, web text, among others) across three languages – English, Chinese and Arabic (Weischedel et al., 2011). The English portion contains 1.6M words. Before OntoNotes, the (much smaller) benchmark datasets used were the MUC (Hirschman and Chinchor, 1998) and ACE (Doddington et al., 2004) corpora. OntoNotes differs from these two corpora with respect to corpus size and the inclusion of a few more genres. Benchmark datasets have of course also been created for other languages, e.g. the Prague Dependency Treebank (Hajič et al., 2018) for Czech, the ANCORA newspaper corpora of Spanish and Catalan (Martí et al., 2007) or TübA-D/Z (Naumann and Möller, 2006) as a newspaper corpus for German.

Despite the fact that OntoNotes contains multiple genres, it is unsuited as a data basis for other domains with very different properties. One example for such a domain is the biomedical domain, for which Gasperin and Briscoe (2008) have shown that the text differs considerably from other text genres such as news or dialogue, and that the complex nature of the texts is for example reflected in the heavy use of abstract entities, such as results or variables. As a result, a lot of corpora have been annotated (Castaño et al. (2002), Cohen et al. (2010), Gasperin et al. (2007), Batista-Navarro and Ananiadou (2011), a.o.) for this domain. It has been shown that coreference resolution for the biomedical domain benefits a lot from in-domain training data (Rösiger and Teufel, 2014).

Literary texts also differ a lot from news texts or dialogue. It has been shown to comprise rather sophisticated language, with a rich vocabulary, direct and indirect speech and a larger set of syntactic constructions (van Cranenburgh and Bod, 2017). Krug et al. (2015) found that the average sentence length is larger in novels and the number of pronouns, as well as the percentage of direct speech is higher. They have also presented a German corpus in which they annotated the syntactic heads of all coreferent expressions referring to a character, and built a rule-based system to detect them automatically. In contrast to this, we decided to annotate general coreference as not to limit the use cases to DH tasks revolving around characters, but to also allow experimentation with other research questions. The German NoSta-D corpus (Dipper et al., 2013) contains a small fraction of Kafka's *Der Prozess* (about 7000 tokens). The ARRAU corpus (Poesio and Artstein, 2008; Uryupina et al., to appear) contains a number of spoken narratives as part of their PEAR domain. However, these texts consist of spontaneous speech and therefore have very different properties than the rather sophisticated literary texts that we aim to annotate in our project. The Phrase Detectives project (Poesio et al., 2013), a linguistic annotation game designed to collect an ever-growing amount of annotated data, also contains English narrative texts from Gutenberg, including a number of tales and more advanced narratives such as Sherlock Holmes and Alice in Wonderland. To the best of our knowledge, standard guidelines were applied to all the genres in this corpus, and no domain-specific considerations were taken into account for the annotation of the literary texts. We believe that the typically long and often non-standard texts pose a challenge for the annotation, which should be reflected in the annotation guidelines.

3 Literary-specific phenomena/considerations

This section gives an overview of the characteristics of literary texts which have to be taken into consideration for the annotation of coreference. We discuss phenomena that we encountered during the on-going annotation of a corpus of literary texts. It comprises thirteen literary texts which stem from different centuries and cover different literary forms. Johann Wolfgang von Goethe's *The Sorrows of Young Werther* (German title: Die Leiden des jungen Werther) is an epistolary novel from 1774. Our annotations are based on a revised version of 1787 and include the introductory words of the fictional editor as well as the first letters of Werther to his friend Wilhelm. Leo Perutz' novel *From nine to nine* (German title: Zwischen neun und neun, original title: Freiheit) was first published in 1918, the novels *Der Mond lacht*[1] and *Nur ein Druck auf den Knopf*[1] followed in 1930.

Moreover, we include the plays *Miss Sara Sampson* by Gotthold Ephraim Lessing and *The Robbers* (German title: Die Räuber) by Friedrich Schiller. As an example of shorter literary texts, we include six German fairytales by the Brothers Grimm: *Frog Prince* (German title: Der Frosckönig), *Hansel and Gretel* (German title: Hänsel und Grethel), *Cinderella* (German title: Aschenputtel), *The Town Musicians of Bremen* (German title: Die Bremer Stadtmusikanten), *Rumpelstiltskin* (German title: Rumpelstilzchen) and *Rapunzel*.

In addition to these German texts, our corpus contains one English text, *A Narrative of the Captivity and Restoration of Mrs. Mary Rowlandson*, from the 17th century which describes encounters between European explorers and settlers and the native peoples (US captivity narratives)[2].

Two narrative spheres. Literary narrative texts differ from other text types in the fact that they exist in "two basic spheres: that of the narrator (and optionally, his narratee) and that of the narrated (individual narrative agents, events, and states of affairs)" (Margolin, 1991, p. 518). This division leads to two levels of texts: the textual actual world which is the intersubjective reality of the text and the subjective subdomain of the narrator and the narrative agents which include subjective views, beliefs and which leads to layers of alternative narrated worlds (Margolin, 1991). As a result, narrative texts pose a complex web of different levels of knowledge of the reader, the narrator and the characters of the fictional world. This difference in knowledge can give ground to a purposefully deployed play with given and new information which influences the information structure and thus the operating principle of coreference in a text.

Genericity. While genericity is not specific to literary texts, the rate of switching between the generic and non-generic use of noun phrases is. The fact that readers readily and often unconsciously connect instances and classes makes the annotation rather challenging.

(1) DER WIRT. [...] Was liegt mir daran, ob ich es weiß, oder nicht, was Sie für eine Ursache hierher führt, und warum Sie bei mir im Verborgnen sein wollen? Ein Wirt nimmt sein Geld, und läßt seine Gäste machen, was ihnen gut dünkt. Waitwell hat mir zwar gesagt, [...]

EN translation: LANDLORD. [...] What is it to me, whether I know or not, what cause has brought you hither, and why you wish to live in seclusion in my house? A landlord takes his money and lets his guests do as they think best. Waitwell, it is true, has told me that [...].

In 1, "A landlord" is a generic expression that refers to the class of landlords. Given the context it is clear, however, that the landlord (who is uttering this sentence), *also* talks about himself.

(2) [Sie]$_1$ ließ [[ihren]$_2$ Regenschirm]$_3$ fallen. [Jeder junge Mann]$_4$ wird in einem solchen Fall blitzschnell nach [dem Schirm]$_5$ greifen und [ihn]$_6$ [der Dame]$_7$ überreichen. Und [die Dame]$_8$ bedankt sich vielmals. Aber diesmal geschah etwas Unerhörtes. [Stanislaus Demba]$_9$ ließ [den Schirm]$_{10}$ liegen.

[She dropped [[her]$_2$ umbrella]$_3$. [Every young man in such a situation]$_4$ would catch [the umbrella]$_5$ blazingly fast and hand [it]$_6$ on to [the lady]$_7$. And [the lady]$_8$ would thank [him]$_8$ a great many times. But this time, something outrageous happened. [Stanislaus Demba]$_9$ left [the umbrella]$_{10}$ on the ground.

A similar case is shown in 2, yet a bit more complex. Mentions 1 and 2 refer to a concrete individual, whose umbrella (mention 3) falls down. The following sentence switches to a generic reading (signified by mention 4). The definite noun phrase 5 also needs to be understood as generic, and thus starts a new coreference chain, similarly to the entity lady (7 and 8). The last sentence again switches to the individual level, and describes that a concrete individual (9) does not pick up the umbrella introduced in 3.

[1] Published in: Herr, erbarme dich meiner. Not translated into English.

[2] Made available under Creative Common licence by EEBO-TCP which is a partnership between the Universities of Michigan and Oxford and the publisher ProQuest to create accurately transcribed and encoded texts based on the image sets published by ProQuest via their Early English Books Online (EEBO) database (`http://eebo.chadwyck.com`).

Entity development. One of the key characteristics of narratives is that change takes place: The (fictional) world changes, and so do the characters that reside in it. This does not only include the name of marrying (mostly female) characters, but also promotions (e.g., in Goethe's Elective Affinities, Captain Otto is promoted to major after half the book, and henceforth only referred to as 'Major'). In such cases where the name of a character or his profession changes, one can argue that the referent probably stays the same. However, characters can also change more drastically, e.g. turning evil or good, where it is then questionable whether the referent remains the same. Another interesting issue is the appearance of dead or unreal characters: Are dreamed and fictional but real characters the same entity? Is the ghost of Hamlet's father the same entity as Hamlet's father?

A more frequent issue is the creation of groups, and plural references to such groups. In Act II, Scene 2 of Schiller's *The Robbers*, a group of six different characters argue. They decide on whether to become robbers but are divided. Various coalitions are formed, and their exact composition is unclear in many cases.

Text knowledge. Characters in narratives simulate individual entities and thus have varying states of knowledge. For plays, mix-ups are an important plot element for an entire genre with a tradition of over 2000 years (e.g., Electra unknowingly meets her brother in Sophocles' play, written about 400 BCE). A lot of the tension in these cases comes from the fact that the reader or audience indeed realizes the truth. Some crime novels also reveal the identity of the perpetrator early on. It is an important question whether the annotations reflect characters' or readers' knowledge, and how much text knowledge one can assume in readers. In our guidelines, we have decided to annotate from the reader's point of view. To deal with the fact that some of the novels have been read by the annotators, the text knowledge is fixed to the knowledge of the annotator after one read-through. Another stylistic device which is sometimes used in literary text is gender confusion. Gender is typically an important marker in pronouns which aids the resolution to the correct noun phrase. Hirst (1981, p. 10) cites an example from *Even Cowgirls get the blues* (Robbins, 1976), where a character called *The Countess* turns out to be male after a couple of pages by implicitly using the pronoun *he* for referencing him.

World knowledge. In historic literary texts, we have the problem that contemporary annotators do not have the same knowledge as the author or typical reader at the time of the works' publication. To overcome this, we try to approximate the world knowledge of a typical reader: wherever the annotator gets the feeling that something is assumed to be common knowledge, he/she is allowed to look up the missing facts to derive the right references.

(3) HERMANN. [...]Da [Karl] auf der Welt nichts mehr zu hoffen hatte, zog ihn der Hall von [Friederichs]$_1$ siegreicher Trommel nach Böhmen. Erlaubt mir, sagte er [zum großen Schwerin]$_2$, da ich den Tod sterbe auf dem Bette der Helden [...].

EN translation: For Karl had nothing to hope for in the world, he was drawn to Bohemia to the sound of [Friederich's]$_1$ triumphant drums. Allow me, he said to [the Great Schwerin]$_2$ that I might die on the beds of heroes ...

For example in 3, "zum großen Schwerin" refers to 'Kurt Christoph Graf von Schwerin', an important and popular general under Friedrich II. This is obvious for contemporaries, but not for today's readers. From discourse context alone, an annotator might annotate 1 and 2 as co-referent, which is not in line with the author's (presumed) intentions.

Lexical variation. One of the main properties of literary texts is the high amount of lexical variation and paraphrasing as a stylistic means. As a result, it is sometimes difficult to decide where coreference ends and bridging begins, such as in 4, where *Verbindung* refers to both the wedding (event) and marriage (state):

(4) MELLEFONT. Mit Unrecht tadelt sie die Verzgerung [einer Zeremonie] [...].
SARA. Neue Freunde sollen die Zeugen [unserer Verbindung] sein? Grausamer [...]

MELLEFONT. Aber überlegen Sie denn nicht, Miss, dass [unserer Verbindung] hier diejenige Feier fehlen würde, die wir ihr zu geben schuldig sind?

EN translation: MELLEFONT. Unjustly, she condemns the delay of [a ceremony].
SARA. New friends shall be the witnesses of [our union]?
MELLEFONT. But bear in mind, Miss, that [our bond] would be lacking the festivity, which we are responsible to give.

Text length. Whereas a discourse in a newspaper is typically rather short, a discourse in literary texts can span hundreds of pages, which poses a challenge for the annotators. For pronominal reference this is not problematic, as pronominal coreference is limited with respect to the attention span of the reader in every text, i.e. the author chooses a pronoun when he/she can be sure that the reader remembers the referent/antecedent. Cases of nominal coreference, however, might now span hundreds of pages. One example for a text which is broken up into different planes and surrounded by a frame story is the appearance of Scheherazade in *One Thousand and One Nights*, where we observe co-references across very long distances.

Obviously, annotators cannot remember every discourse entity that appeared in the first chapter while reading the last chapter. Depending on the importance of the entity for the narrative, co-reference might still be established with additional means by the authors (e.g., the gun that Werther uses for shooting himself is introduced early on). To allow annotation of long coreference chains, the annotators have access to all previously annotated entities and can for example additionally search the text for a certain headword (c.f. Section 4.2 for the annotation tool we developed).

Idiomatic expressions. Idiomatic expressions are arguably more frequent in literary texts than in newspaper text. While they are generally considered non-referring expressions, it is sometimes difficult to decide which of the expressions are idiomatic expressions and which are referring:

(5) 'mit verzerrtem Gesicht' (with a twisted face)

(6) 'Gott sei dank' (thank God)

In the context of (5), the face is mentioned regularly and an established discourse referent. The 'twisted face' is used an idiomatic expression, but can also be understood literally in this context. In (6) this distinction probably depends on the religious beliefs of the person.

Sub-token annotation. While it is typically assumed that pre-modifiers in compounds can only be picked up again in case they are proper nouns (cf. for example in the OntoNotes guidelines), this is not always true in literary texts, as in 7. Additionally, for German, we are faced with the problem that compounds are not multi-words, i.e. they are not separated by blanks. As a result, the word level is sometimes unsuited as an annotation level. Our new tool thus also allows the annotation of mentions within tokens.

(7) Eine schlechte Vorbereitung, eine [trost]suchende Betrübte zu empfangen. Warum sucht sie [ihn] auch bei mir?
A bad preparation for receiving one who seeks [comfort]. But why does she seek [it] from me?

True ambiguity A basic assumption made in many NLP tasks is that there is a ground truth, thus one label that is the correct label when annotating certain phenomena in data. In our corpus, however, we encounter a kind of disagreement between annotators that is not a mistake resulting from the incorrect application of the annotation guidelines but from diverging interpretations of the text. These ambiguities can be used intentionally as a stylistic device in literary texts. Such ambiguities should be preserved in the annotation since they are interesting for literature analysis. An example for how different readings can lead to such a disagreement in the annotation of coreference comes from *The Sorrows of Young Werther*. In (8), the chosen antecedent of the abstract anaphor *Das* (mention 3) was different in the two annotations: one annotator chose a non-nominal-antecedent (mention 1), whereas the other chose a shorter, nominal antecedent (mention 2). Another disagreement occurred in the resolution of *seiner selbst* (mention 6). One annotator assigned it to the chain with mention 5 (*the feeling heart*). This reflects a rather poetic interpretation where the feeling heart enjoys the beauty of the garden. Annotator 2 chose *the Count of M.* (mention 4) as the antecedent which enjoys the beauty of the garden. Note that in the English translation this ambiguity is not preserved.

(8) [[Die Stadt selbst ist unangenehm, dagegen rings umher [eine unaussprechliche Schönheit der Natur]$_1$]$_2$. [Das]$_3$ bewog [den verstorbenen Grafen von M.]$_4$ seinen Garten auf einem der Hügel anzulegen, die mit der schönsten Mannichfaltigkeit sich kreuzen, und die lieblichsten Thäler bilden. Der Garten ist einfach, und man fühlt gleich bei dem Eintritte, daß nicht ein wissenschaftlicher Gärtner, sondern [ein fühlendes Herz]$_5$ den Plan gezeichnet, das [seiner selbst]$_6$ hier genießen wollte.

[[The city itself is unpleasant, whereas round and round there is [an inexpressible beauty of nature]$_1$]$_2$. [This]$_3$ made [the late Count of M.]$_4$ to build his garden on one of the hills, which have cross-bred with the most beautiful diversity, and which make up lovely valleys. The garden is simple, and one can feel it instantly that the plans were not made by a scientific gardener, but [a feeling heart]$_5$, which wanted to enjoy [itself]$_6$.

4 Annotation

4.1 Guidelines

Our annotation guidelines are based on the NoSta-D guidelines.[3] In the following, we outline the differences and extensions between this template and our annotation guidelines. These differences are mainly motivated by factors explained in Section 3. The guidelines have been tested and clarified in an iterative process through parallel annotations. Cases in which differences in the double annotations appeared were discussed. When necessary, the guidelines were modified. In the following, we describe in which points we deviate from the NoSta-D guidelines.

Annotation of entity clusters instead of binary links. We do not annotate links between mentions, but rather assign mentions to entities. This entity-centric view will be explained in more detail in Section 4.2.

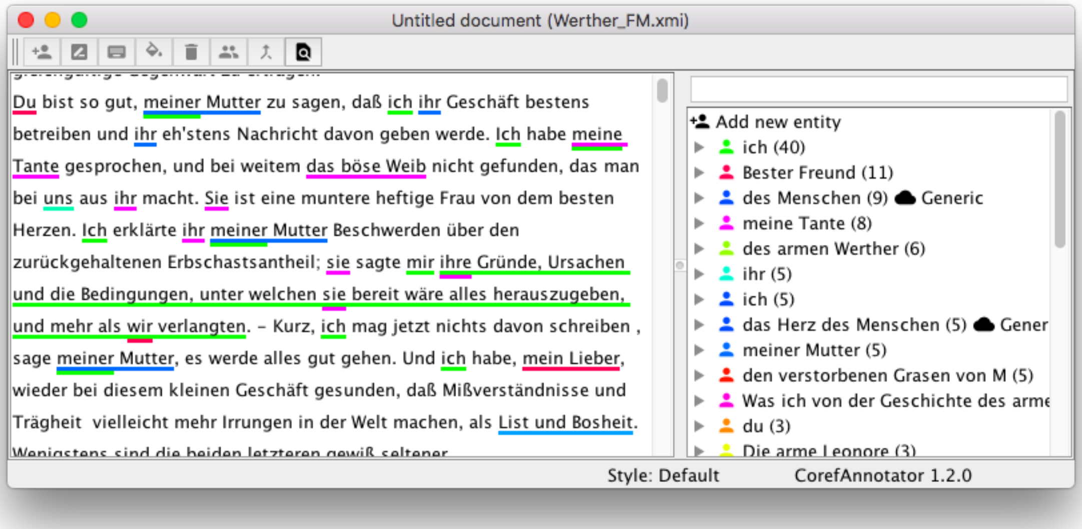
World and text knowledge. As explained above, the world and text knowledge poses a challenge for the annotation. To standardize the annotations, we set the text knowledge to the knowledge of the annotators after one thorough read-through. Temporally developing changes throughout the text are fixed to the knowledge at the end of a text. As for world knowledge, we allow annotators to look up relevant knowledge, which is presented in a way that we can assume that the expression refers to given information (see the example about *zum großen Schwerin*, above).

Genericity. In contrast to the NoSta-D guidelines, we do not annotate "bound" relations but introduce generic entity types instead. Generic chains can only contain generic entities and should not be mixed with non-generic entities.

No annotation of link type. As we assign mentions to entity clusters, we do not annotate the link type (e.g. anaphoric, coreferent, cataphoric).

No singletons. In NoSta-D, all mentions are marked, and at the end of the annotation process, singletons are filtered out. We do not initially mark all NPs as potential chain members but manually determine the candidates during the annotation process (and after the text has been read in its full length before the annotation).

[3] `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-cor-1.1.`



Du bist so gut, meiner Mutter zu sagen, daß ich ihr Geschäft bestens betreiben und ihr eh'stens Nachricht davon geben werde. Ich habe meine Tante gesprochen, und bei weitem das böse Weib nicht gefunden, das man bei uns aus ihr macht. Sie ist eine muntere heftige Frau von dem besten Herzen. Ich erklärte ihr meiner Mutter Beschwerden über den zurückgehaltenen Erbschastsantheil; sie sagte mir ihre Gründe, Ursachen und die Bedingungen, unter welchen sie bereit wäre alles herauszugeben, und mehr als wir verlangten. – Kurz, ich mag jetzt nichts davon schreiben , sage meiner Mutter, es werde alles gut gehen. Und ich habe, mein Lieber, wieder bei diesem kleinen Geschäft gesunden, daß Mißverständnisse und Trägheit vielleicht mehr Irrungen in der Welt machen, als List und Bosheit. Wenigstens sind die beiden letzteren gewiß seltener

- ich (40)
- Bester Freund (11)
- des Menschen (9) Generic
- meine Tante (8)
- des armen Werther (6)
- ihr (5)
- ich (5)
- das Herz des Menschen (5) Gener
- meiner Mutter (5)
- den verstorbenen Grasen von M (5)
- Was ich von der Geschichte des arme
- du (3)
- Die arme Leonore (3)

Figure 1: Screenshot of the annotation tool, showing the text left and the annotated entities on the right

Non-nominal antecedents. We allow non-nominal antecedents in the case of event references. This means that in addition to NPs, VPs, clauses as well as several sentences can be members of a chain in some cases.

(9) I love [baking cakes]. [It] is an activity I used to do with my mom.

Text/document length. We annotate chains that span the entirety of a book, instead of considering chapters or other sub-divisions as documents. This way, we can cover long dependencies that are e.g. spanning frame stories.

Groups/aggregated coreference. We annotate group relations if two members of a group appear apart from each other in a text and if they are subsequently referred to by a mass pronoun subsuming them as in the example below, where mention 3 refers to mention 1 and mention 2 (*John and Paul*).

(10) [John]$_1$ cut the tomatoes and [Paul]$_2$ the carrots. [They]$_3$ liked cooking together.

Idiomatic expressions. We do not annotate NPs in an idiomatic expression:

(11) [the ball] is in your court (not-annotated) vs. [the ball] bounced off the ground (annotated if coreferent with another mention)

Lexical variation. We distinguish cases of bridging from coreferent anaphora. Cases of bridging are not annotated. As mentioned above, the distinction can be tricky in literary texts, and the resolution of individual cases is left to the annotators.

4.2 Annotation tool

The annotations have been produced with a new annotation tool that we make available with an open source license (Apache 2.0).[4] The tool is optimized for fast annotations with a keyboard and departs from existing coreference annotation tools in a number of ways:

Coreference annotations are conceptualized as equivalence sets. All mentions that belong to one coreference chain form a set, and are treated equally. The tool does not support the annotation of relations between mentions (i.e., we cannot annotate a binary relation as cataphoric, for instance). Annotating a

[4] https://github.com/nilsreiter/CorefAnnotator.

mention into a chain adds it directly to the set. Each entity is represented by a color, and can optionally be named. All mentions that belong to the same entity are underlined with the same color in the text view, multiple annotations on the same span result in multiple underlines on different levels.

We also make no assumptions on related tasks. Arbitrary text spans can be annotated as mentions, including sub-token annotations (by default, however, partial token annotations are expanded to the full token). Internally, all annotations are represented as stand-off annotation, using the UIMA framework[5] for text and annotation representation. This allows flexible import and export in a variety of formats (e.g., CoNLL 2012 [given token and sentence boundaries], TEI/XML [as long as the result is valid XML]).

The general usage philosophy is to support keyboard based annotation. Text spans can be selected with the keyboard, and the appropriate entity can be searched for. In addition, selected text spans can be dragged onto the entity. Fast, large scale annotation can be performed via the search function. It supports regular expressions, and all or some found spans can be annotated as a new or existing entity with a single click or press. We are currently exploring ways to represent conflicting/diverging annotations.

If the texts contain appropriate annotation (e.g., stage directions or headings), they can be used to control the formatting (bigger headings and italic stage directions, for instance). This makes reading and annotating more accessible, in particular for long texts. The annotation tool is fully localized in English and German, and can be localized to more languages.

5 Conclusion and Future Work

We have presented an analysis of coreference-related phenomena that we encountered during an annotation study conducted on literary texts. Based on these observations, we propose a number of adaptations to the annotation guidelines. We show that a domain can have a considerable influence on linguistic phenomena and that this has to be reflected in the guidelines to adequately capture them in the annotation. To fully incorporate the peculiarities of this domain and the potential needs of scholars using coreference as an analysis step, deeper changes to the annotation workflow are required: Literary texts may contain references to discourse entities that are intentionally ambiguous. Making a majority-based decision is certainly possible, but does not do justice to the complexities of language use. This has already been observed for annotation in the area of narratology (Gius and Jacke, 2017), but we can showcase it also for 'classic' NLP annotation layers.

Therefore, it will be required to cope with annotated corpora that contain multiple, conflicting annotations, which has severe implications in several areas: a) A number of conceptual questions arise regarding representation in file formats. E.g., do we take annotators' decisions as a whole or do we break them apart into smaller units? b) How do we evaluate annotations properly? Measuring inter-annotator agreement quantitatively assumes that a single ground truth is achievable, which just might not be the case here. c) What does this entail for the way method development works in computational linguistics? How can we use such data sets for training/testing purposes?

As a next step, we will employ the annotation guidelines sketched above to create a corpus of a selection of German literary texts. This corpus can be made openly available and will contain texts written by Goethe, Perutz, as well as several folktales.

References

R.T. Batista-Navarro and S. Ananiadou. 2011. Building a coreference-annotated corpus from the domain of biochemistry. *ACL HLT 2011*, page 83.

José Castaño, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature. In *Proceedings of the International Symposium on Reference Resolution for NLP*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

[5] `http://uima.apache.org`

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogrena, Martha Palmer, and Lawrence E. Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of the Second Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2010), LREC 2010.*

Stefanie Dipper, Anke Lüdeling, and Marc Reznicek. 2013. NoSta-D: A corpus of German non-standard varieties. *Non-Standard Data Sources in Corpus-Based Research*, (5):69–76.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *LREC*. European Language Resources Association.

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 257–264. Association for Computational Linguistics.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium.*

Evelyn Gius and Janina Jacke. 2017. The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis. *International Journal of Humanities and Arts Computing*, 11(2):233–254, October.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2018. Prague dependency treebank 3.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lynette Hirschman and Nancy Chinchor. 1998. Appendix F: MUC-7 Coreference Task Definition (version 3.0). In *MUC*.

G. Hirst. 1981. *Anaphora in Natural Language Understanding: A Survey*. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Uri Margolin. 1991. Reference, coreference, referring, and the dual structure of literary narrative. *Poetics Today*, 12(3):517–542.

M.A. Martí, M. Taulé, M. Bertran, and L. Márquez. 2007. AnCora: Multilingual and Multilevel Annotated Corpora.

Karin Naumann and V Möller. 2006. Manual for the annotation of in-document referential relations. *University of Tübingen.*

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):3.

T. Robbins. 1976. *Even cowgirls get the blues*. Bantam Books: Novel. Bantam Books.

Ina Rösiger and Simone Teufel. 2014. Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of EACL.*

Olga Uryupina, Ron Artstein, Antonella Bristot, Ferederica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. to appear. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Journal of Natural Language Engineering*.

Andreas van Cranenburgh and Rens Bod. 2017. A data-oriented model of literary language. In *Proceedings of EACL*.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. pages 54–63.