Enhancing Character Type Detection using Coreference Information: Experiments on Dramatic Texts

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Janis Malte Pagel

aus Bielefeld

Hauptberichter	Prof. Dr. Jonas Kuhn
Mitberichter	Prof. Dr. Nils Reiter
Mitberichter	Prof. Dr. Massimo Poesio

Tag der mündlichen Prüfung: 17. November 2023

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2024

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen.

I hereby declare that this text is the result of my own work and that I have not used sources without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The thesis was not used in the same or in a similar version to achieve an academic grading.

`Stuttgart, den 19. Juni 2024`

Ort, Datum

Unterschrift

To Oma Elfriede, Oma Helga, Opa Horst and Tante Marion

Contents

List of Tables

List of Figures

List of Abbreviations

ADJ Adjective, page 68
ART Article, page 68

BT Bourgeois Tragedy, page 128

CDA Computational Drama Analysis, page 16
CL Computational Linguistics, page 1
CLS Computational Literary Studies, page 13
CR Coreference Resolution, page 3

IAA Inter-Annotator Agreement, page 40

LDA Latent Dirichlet Allocation, page 46
LLM Large Language Model, page 45
LOC Location, page 69

ML Machine Learning, page 5

NE Named Entity, page 43
NER Named Entity Recognition, page 43
NLP Natural Language Processing, page 1
NLU Natural Language Understanding, page 4
NN Normal Noun, page 68
NP Noun Phrase, page 19

ORG Organization, page 69

PC Principal Component, page 142
PCA Principal Component Analysis, page 142
PER Person, page 69
PoS Part-of-Speech, page 68
PRON Pronoun, page 68
PUNCT Punctuation, page 68

RF Random Forest, page 123

SD Standard Deviation, page 66
SuD Sturm und Drang, page 128
SVM Support Vector Machine, page 49

WC Weimar Classicism, page 128

Abstract

This thesis describes experiments on enhancing machine-learning based detection of literary character types in German-language dramatic texts by using coreference information.

Dramatic texts, or theatre plays, are one of three commonly considered main types of literary texts (the other two being prose and poetry). They have a handful of intriguing properties, such as being in dialogue form with designated speaker tags, featuring a cast list of occurring characters and usually being conceived to be performed on stage, hence also featuring stage directions. All these properties make dramatic texts an interesting type of text to explore using methods from computational linguistics and natural language processing. Naturally, characters play a major role in dramatic texts, as they drive forward the plot of a play and contribute the majority of text in the form of character speech. Literary studies has identified several distinct types of characters that frequently fulfill different roles in a play. For instance, the *tender father* (German: *zärtlicher Vater*) and the *virtuous daughter* (German: *tugendhafte Tochter*) are character types that have been identified and discussed by literary studies' works on plays of a literary genre called *bourgeois tragedy* (German: *Bürgerliches Trauerspiel*). Automatically identifying such character types has a number of advantages. On the one hand, making the potential properties of character types explicit and testing the capability of computational models to automatically detect character types using those properties can reveal new insights into the nature of theoretically motivated character types previously not in the sight of literary studies. On the other hand, having data about character types available can inform downstream applications such as narrative modelling or character relationship modelling. Previous attempts to automatically detect character types often focus on either textual features pertaining to single characters or relationships between characters based on co-occurring stage presence. However, a large part of character interactions happens on a textual level as well, in the form of characters mentioning other characters, either in their presence or absence. Knowledge about when a character mentions another character and what they say about one another might be hugely beneficial in the automatic detection of character types, as the way a character talks about other characters plays a major aspect in the role that character takes in the play. Linguistically, the phenomenon of textual mentions referring to the same person or entity is called *coreference* and the automatic resolution of such co-referring mentions is called *coreference resolution*. Having access to high quality, automatically resolved coreferences can be a major boost for automatic character type detection and can enhance or create features previously unavailable.

The thesis makes four major contributions to the research discourse of character type detection and coreference resolution: (i) a corpus of annotations of coreference on dramatic texts, called *GerDraCor-Coref*, (ii) a rule-based system to automatically resolve coreferences on dramatic texts, called *DramaCoref*, as well as experiments and analyses of results by using DramaCoref on GerDraCor-Coref, (iii) experiments on the automatic detection of three selected character types (title characters, protagonists and schemers)

using machine-learning approaches, and (iv) experiments on utilizing the coreference information of (i) and (ii) for improving the performance of character type detection of (iii).

As for (i), it turns out that dramatic texts behave differently in terms of the distribution of coreferent mentions and entities when compared to other corpora containing coreference annotations, namely corpora containing newspaper articles, radio news and radio interviews. For dramatic texts, there are many more mentions on average, however the average density of entities is lower than in other corpora. Coreference clusters containing long distances between mentions and clusters with a large amount of mentions are on average much more prevalent in dramatic texts. Lastly, methods are presented to use the coreference annotations to perform detailed character analysis and retrieve information about the relationship between specific characters.

For (ii), it could be shown that neural mention detection leads to slightly better results than resolving mentions by using the output of constituent parsers. In terms of using DramaCoref on GerDraCor-Coref, the system produced better results than other oft-the-shelf systems on the same dataset. In general, DramaCoref suffers slightly from lower recall compared to rule-based systems developed for newspaper corpora. An error analysis shows that many problems in terms of precision stem from the larger text length of dramatic texts and the resulting longer coreferential dependencies. Overall, it could be shown that DramaCoref is able to resolve coreferences on dramatic texts better than other existing systems, but still leaves plenty of room for improvements.

For (iii), it could be shown that models perform highest for the task of detecting title characters, which is intuitively the easiest task, but also structurally different from detecting protagonists and schemers. Models for detecting protagonists perform only slightly lower and the performance for detecting schemers was the lowest. An in-depth analysis of the features used in the models showed that a feature of counting the number of tokens a character utters was always important. Furthermore, the models benefit from using features of the domain of topic modelling and stage presence, as well as co-occurrence based character networks. Overall, it could be shown that a multi-dimensional approach is paramount to approaches that focus on a single type of feature.

Lastly, (iv) shows that using the coreference information gathered in (i) and (ii) has a positive impact on detecting character types and helps boosting the performance significantly. This holds true for all three character types and for all features enhanced with coreference information.

Overall, the thesis shows that dramatic texts are a challenging type of text for which future research on both natural language processing methods as well as computational literary studies methods is obligatory. In the future, it might also be interesting to utilize neural network architectures for coreference resolution and character type detection on dramatic texts without loosing the ability to interpret the results and to channel them back into a human-in-the-loop approach.

Deutsche Zusammenfassung

Die Arbeit beschreibt Experimente zur Verbesserung der auf maschinellem Lernen basierenden Erkennung von literarischen Figurentypen in deutschsprachigen Dramen mittels Koreferenzinformationen.

Dramen oder Theaterstücke sind eine der drei Haupttypen literarischer Texte (mit den anderen beiden Typen Prosa und Lyrik). Dramen haben eine Handvoll faszinierender Eigenschaften, wie z.B. eine Dialogform mit ausgewiesenen Sprechern, eine Dramatis Personæ der auftretenden Figuren und die Tatsache, dass sie in der Regel für die Aufführung auf der Bühne konzipiert sind und daher auch Regieanweisungen enthalten. Alle diese Eigenschaften machen dramatische Texte zu einer interessanten Textart, die mit Methoden der Computerlinguistik und Maschineller Sprachverarbeitung untersucht werden kann. Natürlich spielen Figuren in dramatischen Texten eine wichtige Rolle, da sie die Handlung eines Stücks vorantreiben und den Großteil des Textes in Form von Figurenrede beisteuern. Die Literaturwissenschaft hat mehrere verschiedene Arten von Figuren identifiziert, die häufig unterschiedliche Rollen in einem Stück erfüllen. So sind beispielsweise der zärtliche Vater und die tugendhafte Tochter Figurentypen, die in literaturwissenschaftlichen Arbeiten über Theaterstücke der literarischen Gattung "Bürgerliches Trauerspiel" identifiziert und diskutiert wurden. Die automatische Identifizierung solcher Figurentypen hat eine Reihe von Vorteilen. Einerseits kann die explizite Darstellung der potenziellen Eigenschaften von Figurentypen und die Prüfung der Fähigkeit von Computermodellen zur automatischen Erkennung von Figurentypen anhand dieser Eigenschaften neue Einblicke in die Natur theoretisch motivierter Figurentypen eröffnen, die die Literaturwissenschaft bisher nicht im Blick hatte. Andererseits kann die Verfügbarkeit von Daten über Figurentypen nachgelagerte Anwendungen wie die Modellierung von Erzählungen oder die Modellierung von Figurenbeziehungen informieren. Bisherige Versuche, Figurentypen automatisch zu erkennen, konzentrieren sich oft entweder auf Textmerkmale, die sich auf einzelne Figuren beziehen, oder auf Beziehungen zwischen Figuren, die auf einer gemeinsamen Bühnenpräsenz basieren. Ein großer Teil der Interaktionen zwischen Figuren findet jedoch auch auf Textebene statt, und zwar in Form von Erwähnungen anderer Figuren, entweder in deren Anwesenheit oder in deren Abwesenheit. Das Wissen darüber, wann eine Figur eine andere Figur erwähnt und was sie übereinander sagen, könnte bei der automatischen Erkennung von Figurentypen von großem Nutzen sein, da die Art und Weise, wie eine Figur über andere Figuren spricht, einen wichtigen Aspekt für die Rolle dieser Figur im Stück darstellt. In der Linguistik wird das Phänomen der textlichen Erwähnungen, die sich auf dieselbe Person oder Entität beziehen, als *Koreferenz* bezeichnet, und die automatische Auflösung solcher koreferierenden Erwähnungen wird als *Koreferenzauflösung* bezeichnet. Der Zugang zu qualitativ hochwertigen, automatisch aufgelösten Koreferenzen kann die automatische Erkennung von Figurentypen erheblich fördern und bisher nicht verfügbare Merkmale verbessern oder erst möglich machen.

Die Arbeit leistet vier wichtige Beiträge zum Forschungsdiskurs über die Erkennung von Figurentypen und die Auflösung von Koreferenzen: (i) ein Korpus von Koreferenzannotationen auf Dramen, genannt *GerDraCor-Coref*, (ii) ein regelbasiertes System zur automatischen Auflösung von Koreferenzen in Dramen, genannt *DramaCoref*, sowie Experimente und Analysen der Ergebnisse unter Verwendung von DramaCoref auf GerDraCor-Coref, (iii) Experimente zur automatischen Erkennung dreier ausgewählter Figurentypen (Titelfiguren, Protagonisten und Intriganten) mit Hilfe von Machine-Learning-Ansätzen und (iv) Experimente zur Nutzung der Koreferenzinformationen aus (i) und (ii) zur Verbesserung der Leistung der Figurentypenerkennung aus (iii).

Was (i) betrifft, so zeigt sich, dass sich Dramen im Hinblick auf die Verteilung von koreferenten Erwähnungen und Entitäten anders verhalten als andere Korpora mit Koreferenzannotationen, und zwar Korpora mit Zeitungsartikeln, Radionachrichten und Radiointerviews. In Dramen gibt es im Durchschnitt viel mehr Erwähnungen, die durchschnittliche Dichte der Entitäten ist jedoch geringer als in anderen Korpora. Koreferenzcluster mit großen Abständen zwischen den Erwähnungen und Cluster mit einer großen Anzahl von Erwähnungen sind im Durchschnitt in Dramen sehr viel häufiger. Schließlich werden Methoden vorgestellt, mit denen die Koreferenzannotationen für eine detaillierte Figurenanalyse verwendet werden können, um Informationen über die Beziehung zwischen bestimmten Figuren zu erhalten.

Für (ii) konnte gezeigt werden, dass die neuronale Erkennung von Erwähnungen zu etwas besseren Ergebnissen führt als die Auflösung von Erwähnungen anhand der Ausgabe von Konstituentenparsern. Was die Verwendung von DramaCoref auf GerDraCor-Coref angeht, so erzielte das System bessere Ergebnisse als andere verfügbare Systeme auf demselben Datensatz. Im Allgemeinen leidet DramaCoref unter einem etwas geringeren Recall im Vergleich zu regelbasierten Systemen, die für Zeitungskorpora entwickelt wurden. Eine Fehleranalyse zeigt, dass viele Probleme in Bezug auf die Precision auf die größere Textlänge von Dramen und die daraus resultierenden längeren koreferentiellen Abhängigkeiten zurückzuführen sind. Insgesamt konnte gezeigt werden, dass DramaCoref in der Lage ist, Koreferenzen in Dramen besser aufzulösen als andere existierende Systeme, aber noch viel Raum für Verbesserungen lässt.

Für (iii) konnte gezeigt werden, dass die Modelle für die Aufgabe der Erkennung von Titelfiguren am besten abschneiden, was intuitiv die einfachste Aufgabe ist, sich aber auch strukturell von der Erkennung von Protagonisten und Intriganten unterscheidet. Die Modelle für die Erkennung von Protagonisten sind nur geringfügig schlechter und die Leistung für die Erkennung von Intriganten ist am niedrigsten. Eine eingehende Analyse der in den Modellen verwendeten Merkmale zeigt, dass ein Merkmal, das die Anzahl der Token zählt, die eine Figur äußert, immens wichtig ist. Darüber hinaus profitieren die Modelle von der Verwendung von Merkmalen aus dem Bereich des Topic Modellings und der Bühnenpräsenz sowie von Figurennetzwerken auf der Basis von Ko-präsenz. Insgesamt konnte gezeigt werden, dass ein mehrdimensionaler Ansatz den Ansätzen, die sich auf eine einzige Art von Merkmalen konzentrieren, überlegen ist.

Schließlich zeigt (iv), dass die Verwendung der in (i) und (ii) gesammelten Koreferenz-

informationen eine positive Auswirkung auf die Erkennung von Figurentypen hat und die Leistung erheblich steigert. Dies gilt für alle drei Figurentypen und für alle mit Koreferenzinformationen angereicherten Merkmale.

Insgesamt zeigt die Arbeit, dass Dramen eine herausfordernde Textsorte sind, für die zukünftige Forschung sowohl zu Methoden der maschinellen Sprachverarbeitung als auch zu Methoden der Computational Literary Studies obligatorisch ist. In Zukunft könnte es interessant sein, neuronale Netze für die Auflösung von Koreferenzen und die Erkennung von Figurentypen in Dramen zu verwenden, ohne dabei die Fähigkeit zu verlieren, die Ergebnisse zu interpretieren und sie in einen Human-in-the-Loop-Ansatz zurückfließen zu lassen.

Acknowledgements

I would like to thank my supervisors, Prof. Dr. Jonas Kuhn and Prof. Dr. Nils Reiter, for their guidance and many immensely helpful discussions and feedback. This thesis would not have been possible without them and I am particularly thankful that they enabled me to explore different directions and were always great role models for my growth as a researcher.

I am thanking Jonas especially for agreeing to be my main examiner without hesitation, many useful meetings in the beginning of the PhD which helped me immensely to find the direction I wanted to take with my research and his support during the final stages of the thesis.

Many thanks goes to Nils for stemming the main bulk of supervision, frequent meetings with invaluable feedback and providing me the opportunity to carry out the research in this thesis.

I also wish to thank Prof. Dr. Massimo Poesio for readily agreeing to serve as an external examiner and providing insightful and detailed feedback on the submitted draft of my thesis.

Thanks also goes to the members of the QuaDramA and Q:TRACK projects Nils Reiter, Marcus Willand, Melanie Andresen and Benjamin Krautter, who provided me the opportunity to carry out lot of interesting research and without whom this thesis wouldn't have been possible. Thanks also goes to the Volkswagen Foundation as the funder of the QuaDramA project and the Deutsche Forschungsgemeinschaft as the funder of the Q:TRACK project.

I would also like to thank my office colleagues Sarah Schulz, Nathalie Wiedmer and Judith Nester for feedback, encouragements and the pleasant working atmosphere they all created.

Many thanks to Prajit Dhar, Melanie Andresen, Johanna Binnewitt, Michael Göggelmann and Judith Nester for proofreading parts of this thesis and giving valuable feedback.

Furthermore, I wish to thank all members of the IMS in Stuttgart, the IDH in Cologne and the Spinfo group for many fun chats, retreats and always providing helpful feedback on my work.

Penultimately, I wish to thank my family. Danke Mama, Papa, Finn, Laurens und Louis für eure Unterstützung während dieser Zeit und das gute Gefühl immer nach Hause zu kommen wenn ich euch besuche. Danke auch an Jacqueline, Dirk, Opa Willi, Jörg, Denise, Alex und Dennis für eure Unterstützung und euren Glauben an mich und meine Arbeit.

Last but not least, I am immensely thankful to Prajit for standing by my side from the

very beginning of my dissertation journey until the very end and always being a pillar to rely on, in all the good and also the stressful moments.

Writing a dissertation proved to be a challenging endeavor and I am very grateful to have been accompanied by so many awesome and helpful people on the way!

Freilich. Aber eine Einleitung muß doch sein.
(Indeed. But there has to be an introduction.)[a]

Förster in Otto Ludwig's "Der Erbförster"

[a]The English translations of the epigraphs of all chapters were done by me.

1 Introduction

Dramatic texts are an important type of text in literary scholarship, but only in recent times have literary texts received some attention in the fields of *computational linguistics* (CL) or *natural language processing* (NLP) (e.g. Finlayson 2012; Bamman, Underwood, and Smith 2014; Krug et al. 2015; Iyyer et al. 2016; Krug et al. 2018; Bamman, Popat, and Shen 2019; Bamman, Lewke, and Mansoor 2020), while dramatic texts are still usually not being worked on. This might not be a surprise, since NLP often focuses on the development of methods applied to general and available texts, usually newspaper texts, rather than being concerned with specific domains or how different types of texts might influence how the methods have to be applied.[1] But even in domain-specific NLP research, dramatic texts, as well as literary texts in general, usually do not play a role. However, these types of text pose interesting challenges for NLP applications, as the texts are unusually long (compared to normally researched texts in NLP), contain almost exclusively direct speech in the form of dialogues, are traditionally structured into acts and scenes, and include challenging phenomena on many different linguistic levels such as orthography, syntax, semantics and discourse. Such challenges include the fact that many texts were written hundreds of years ago, or the literary nature of the texts which leads to the texts often containing highly metaphorical language and featuring complex plots and interactions of literary characters.

[1]Naturally, there are exceptions to this, see for example Ramponi and Plank (2020), which also identify bias in the kind of tasks that receive attention with regard to (unsupervised) domain adaptation.

Erste Szene
Franken. Saal im Moorischen Schloß.
Franz. Der alte Moor.

Franz.
Aber ist Euch auch wohl, Vater? Ihr seht so blaß.

Der Alte Moor.
Ganz wohl, mein Sohn — was hattest du mir zu sagen?

Franz.
Die Post ist angekommen — ein Brief von unserm Korrespondenten in Leipzig —

Der Alte Moor
begierig.
Nachrichten von meinem Sohne Karl?

Franz.
Hm! Hm! — So ist es. Aber ich fürchte — ich weiß nicht — ob ich — Eurer Gesundheit? — Ist Euch wirklich ganz wohl, mein Vater?

Der Alte Moor.
Wie dem Fisch im Wasser! Von meinem Sohne schreibt er? — Wie kommst du zu dieser Besorgnis? Du hast mich zweimal gefragt.

Franz.
Wenn Ihr krank seid — nur die leiseste Ahndung habt, es zu werden, so laßt mich — ich will zu gelegnerer Zeit zu Euch reden. *Halb vor sich.* Diese Zeitung ist nicht für einen zerbrechlichen Körper.

Figure 1.1.: TextGrid Repository (2012g): *Schiller, Friedrich. Die Räuber.* For a translation, see Figure C.1

Consider the snippet from Friedrich Schiller's *Die Räuber* in Figure 1.1. We can observe characteristics of spoken language such as contracted word forms (*unserm*, *our*, line 12; *gelegnerer*, *more suitable*, line 29), interjections (*Hm! Hm!*, line 20) or ellipses (*Aber ich*

fürchte – ich weiß nicht – ob ich – Eurer Gesundheit?, *But I fear – I do not know – if I – your Health?*, lines 20f.)[2]. There are also word meanings that have changed over time, e.g. *Zeitung* (line 30), which nowadays means *newspaper*, but is used here as to mean *message* or *news* (cf. Dudenredaktion n.d.), for which in current day German, the word *Nachricht* would be used. All of this will make working with current and state-of-the-art NLP tools and models difficult, as they are developed and trained on current day German.

Another aspect which sets literary texts apart from texts usually used in NLP is the role of entities for these texts. Literary texts are subject to literary studies research and in this research, characters and other types of entities play an important role when interpreting literary texts. On the computational side, having knowledge about entities and where which entities are placed inside the text enables a wide array of analyses and follow-up analyses that are only made possible via information about the entities. This can range from extracting co-occurrence matrices and social networks from texts based on the occurrence of entities (Blessing et al. 2017) or performing coreference resolution (CR) in order to enable plot sentiment analysis and character type detection (Rösiger, Schulz, and Reiter 2018).

This thesis aims to tackle the issue of character type detection in dramatic texts using information about the entity mentions within the texts. The main contributions can be broken down into three main parts: **(i)** the genesis of coreference annotations on a corpus of German dramatic texts (Chap. 4), **(ii)** the development of a system in order to automatically resolve coreferences on unannotated dramatic texts and enable research on a wider variety of plays (Chap. 5) and **(iii)** the introduction of methods for automatically detecting character types in dramatic texts, such as protagonists and schemers, and how to incorporate coreference information in order to improve detection rate and interpretability of the results (Chap. 6).

As mentioned, for the first and second part (i and ii), coreference resolution will be the methodological focus. CR is a high level task in NLP that involves information from many linguistic areas in order to be performed successfully and can be used as input for many down-stream applications that are of interest for DH research questions. The goal is to identify phrases in a text that refer to the same entity, which could be a person, an object or more abstract concepts such as feelings or ideas. While this is the theoretically-motivated end goal, concrete implementations of CR often focus on identifying pairs of mentions referring to the same entity, instead of finding all mentions of an entity directly. In this so called "mention-pair" model, an anaphor is linked to a

[2]See Bußmann (2008, 234f.) for an overview of linguistic features of spoken language.

preceding antecedent. Antecedents are stand-alone phrases for which the referent is given by the phrase itself, while anaphors require the antecedent to be resolved. Anaphors are typically realized as pronouns. Consider for example the following sentence:

(1) Sarah drove six miles until she reached Seattle.

In this example sentence, the phrase *Sarah* can be resolved to refer to the entity named *Sarah* without any further context, while the pronoun *she* requires the phrase *Sarah* as an anchor in order to be understood and in order to understand that *she* refers to an entity called *Sarah*. Hence, the phrase *Sarah* can be labeled as an antecedent and the phrase *she* as an anaphor. The task of CR is to automatically resolve this and other types of co-references and to return an index that indicates that *she* and *Sarah* refer to the same entity *Sarah*. Such an index could look like this:

(2) Sarah_1 drove six miles until she_1 reached Seattle.

While this shows a concrete distinction between an only theoretically framed task (resolving all mentions in a text) and the concrete implementation (linking pronouns to antecedents)[3], there is a third level (iii): The concrete application of the task. Downstream applications for CR include relationship modelling (Iyyer et al. 2016), network modelling (Pagel 2022a), character detection (Jahan et al. 2020) and analysis of characters (Andresen and Vauth 2018a). Since CR is a natural language phenomenon and humans generally have no problems resolving coreference on dramatic texts, working on CR also serves the more high level goal of *natural language understanding* (NLU). But also as a task in its own right, optimising CR on dramatic texts can be a rewarding endeavour, because of the aforementioned challenges and what can be learned about CR on dramatic texts as well as about the dramatic texts themselves. Figure 1.2 shows the beginning of Gotthold Ephraim Lessing's play *Miss Sara Sampson*.

1 **Erster Aufzug**
2
3 **Erster Auftritt**
4
5 *Der Schauplatz ist ein Saal [im Gasthofe]$_1$.*
6
7 *[Sir William Sampson]$_2$ und Waitwell treten in Reisekleidern herein.*

[3]This difference and the bridge between the two levels will later be called and discussed as *operationalization*.

8
9 [Sir William]$_2$.
10 Hier [meine]$_2$ Tochter? Hier in [diesem elenden Wirtshause]$_1$?

Figure 1.2.: TextGrid Repository (2012e): *Lessing, Gotthold Ephraim. Miß Sara Sampson*, extended with markup showing coreference relations for nominal phrases. For a translation, see Figure C.2

In this snippet, the phrases *im Gasthofe* (*at the inn*) and *diesem elenden Wirtshause* (*this miserable inn*), as well as *Sir William Sampson*, *SIR WILLIAM* and *meine* (*my*), are coreferent, i.e. refer to the same entity, indicated by indices. In contrast to for example CR on newspaper articles, we can use the inherent structure of the text to resolve first person pronouns. In this example, *meine* is marked as being uttered by Sir William Sampson via the speaker tag in line 8. This means that all instances of first person pronouns occurring under the speaker tag of Sir William can easily be resolved as belonging to the same entity. Such information can then be used to perform analyses on the use of pronouns for literary characters. The fact that the structure of dramatic texts allows to resolve first person pronouns more easily shows that dramatic texts do not only hold challenges for NLP but also provide the potential to use their unique nature to improve NLP methods in a way that is not applicable to texts from other domains. In this vein, the thesis aims to explore how structural information in dramatic texts can be used to improve and benefit CR.

This leads to the third part (iii), concerned with the question of how coreference resolution can be used to enhance character type detection. To shed light on this question, we will not only use CR, but also investigate classification tasks such as protagonist detection and classification of literary character types such as the *schemer*. In order to determine which character(s) is/are the potential protagonist(s) of a dramatic text, machine learning (ML) using linguistic and structural features will be applied and the results analysed in order to make the output of the ML algorithms more interpretable. For example, we can investigate which features were most useful in classifying protagonists and hence learn more about the characters and how they are characterized. This knowledge can in turn be used to serve as an input for literary scholarship, e.g. interpreting the characters with the quantitative evidence from the ML experiments. However, in this work, we will focus on the NLP and ML aspects necessary to provide literary scholarship with this kind of information, not on the results of literary interpretation of texts using this knowledge. Protagonist detection can also be aided by using coreference information, as features

can be improved or new features become possible once information about the references and coreferences are available. The insights from classifying protagonists will be used to classify a much more complex character type: the *schemer*.
In the end, all previous results, insights and learnings are combined to present experiments on the automatic classification of character types using deeply annotated coreference information. This way, the thesis not only provides methodologies for automatically extracting coreference information and character types from dramatic texts and hence enables further research on the information gathered, but also shows one possible way of applying NLP methods in order to render high-level interpretations in the realm of computational literary studies.

1.1. Outline of the Thesis

The thesis is structured in the following way:

Chapter 2 will give a high-level overview of the theoretical and methodological background that will be relevant throughout the thesis. In particular, it will give overviews about the fields of digital humanities, computational literary studies, drama in scholarly and quantitative works and coreference resolution from a linguistic and a NLP perspective.

Chapter 3 will provide a literature review of related work regarding literary coreference annotation and resolution and the detection of protagonists and character types.

Chapter 4 introduces GerDraCor-Coref, a corpus of coreference annotations on German-language dramatic texts and offers a variety of statistical and single-play analyses on the data.

Chapter 5 introduces DramaCoref, a rule-based system for resolving coreferences on dramatic texts. The chapter furthermore describes several experiments on GerDraCor-Coref and other corpora with the aim to resolve coreference, both using DramaCoref and other previously introduced systems.

Chapter 6 presents series of experiments on character type detection, namely on title character, protagonist and schemer detection.

Chapter 7 combines the character type detection of Chapter 6 and the information about coreference relationships gathered in the Chapters 4 and 5 in order to enhance the detection and interpretation of automatic character type detection.

Chapter 8 concludes the thesis by giving a summary of the insights and results gained in this thesis and gives an outlook on future research possibilities.

1.2. Research Questions

The main research questions of chapters 4 to 7 can be summarized as follows:

Chapter 4

- Do dramatic texts differ with regards to their coreferential structure in comparison with other types of (literary) texts?
- What properties do long coreference chains and chains with spread-out mentions in dramatic texts have?
- What can coreference information tell us about the characters of a play and their relationship to each other?

Chapter 5

- What is the performance of a sieve-based system for CR on dramatic texts?
- Which passes of the sieve-based system perform best for CR on dramatic texts?
- How does the performance differ with regards to other types of (literary) texts?

Chapter 6

- How can literary character types be operationalized and automatically detected?
- What is the difference in performance of different types of literary characters and how can these differences be explained?

Chapter 7

- Can coreference information be used to improve the automatic detection of literary character types?

1.3. Publications & Contributions

The following section lists papers I have published or co-authored during the preparation of this thesis and highlights the research contributions I have made. In the beginning of each applicable chapter, I will indicate which papers form the basis of the results and insights presented in the respective chapter and in what respect the information shown in the thesis are summaries of the information in the paper or new contributions.

- Nils Reiter, Benjamin Krautter, Janis Pagel, and Marcus Willand (Dec. 2018). "Detecting Protagonists in German Plays around 1800 as a Classification Task". In: *Book of Abstracts of the European Association for Digital Humanities (EADH)*. Galway, Ireland. DOI: 10.18419/opus-10162. URL: https://elib.uni-stuttgart.de/bitstream/11682/10179/1/article.pdf
  - This papers presents results on automatically detecting protagonists of German theatre plays using the support vector machine algorithm. I was mainly involved in co-writing the text and retrieving new results necessary to gain a better understanding of the main findings, such as results for the majority baseline.
- Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand (Nov. 2018). "Titelhelden und Protagonisten — Interpretierbare Figurenklassifikation in deutschsprachigen Dramen". In: *LitLab Pamphlets* 7. URL: https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf
  - A study on automatic protagonist and title character detection for German theatre plays. I was responsible for carrying out the experiments and retrieving all experimental results. I also co-wrote the text with a focus on the experiment and result sections.
- Benjamin Krautter and Janis Pagel (Mar. 2019). "Klassifikation von Titelfiguren in deutschsprachigen Dramen und Evaluation am Beispiel von Lessings "Emilia Galotti"". In: *Book of Abstracts of DHd*. Frankfurt am Main, Germany, pp. 160–164. DOI: 10.5281/zenodo.4622195. URL: https://elib.uni-stuttgart.de/bitstream/11682/10382/1/KRAUTTER_Benjamin_Klassifikation_von_Titelfiguren_in_deutsch.pdf
  - A paper focusing on title character detection with extended annotations and features compared to Krautter et al. (2018). I carried out the experiments and obtained all experimental results. Benjamin Krautter and I co-wrote the text.

- Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand (2020). ""[E]in Vater, dächte ich, ist doch immer ein Vater". Figurentypen und ihre Operationalisierung". In: *ZfdG* 5.7. DOI: 10.17175/2020_007. URL: http://www.zfdg.de/2020_007
  - This article presents results on automatically classifying complex (literary) character types such as tender fathers and virtuous daughters. I was responsible for carrying out all experiments and obtaining the experimental results. I co-wrote the text with a focus on the experiment and result description.
- Janis Pagel and Nils Reiter (May 2020). "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German". In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 55–64. URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf
  - This paper presents a new corpus of plays annotated for coreference. I was responsible for supervising the students annotating the data and conducted the statistical analyses presented in the paper. Nils Reiter had an advisory role and we wrote the text of the paper collaboratively.
- Janis Pagel and Nils Reiter (Nov. 2021). "DramaCoref: A Hybrid Coreference Resolution System for German Theater Plays". In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021)*. Punta Cana, Dominican Republic, pp. 36–46. URL: https://aclanthology.org/2021.crac-1.4
  - This paper presents a rule-based system for resolving coreferences and uses transformer models to extract mentions from the data. The system is catered towards drama and the paper provides results and error analyses of experiments on applying the system on the data shown in Pagel and Reiter (2020). I implemented the system and conducted the experiments presented in the paper. Nils Reiter had an advisory role. We wrote the text of the paper collaboratively.
- Janis Pagel, Nidhi Sihag, and Nils Reiter (Nov. 2021). "Predicting Structural Elements in German Drama". In: *Proceedings of the Second Conference on Computational Humanities Research (CHR2021)*. Amsterdam, The Netherlands (Online), pp. 217–227. URL: http://ceur-ws.org/Vol-2989/short_paper34.pdf
  - The paper investigates the application of transformer models in order to predict structural elements in German plays such as speaker tags, stage directions or character speech. The paper is the result of a student project by Nidhi

Sihag whom I supervised. I carried out the experiments with the CRF model and gathered the statistical information on the corpus. Nils Reiter had an advisory role. We wrote the text of the paper collaboratively.

- Janis Pagel (July 2022a). "Co-reference networks for dramatic texts: Network analysis of German dramas based on co-referential information". In: *Book of Abstracts of DH2022*. Tokyo, Japan (Online), pp. 326–329. URL: https://dh2022.dhii.asia/dh2022bookofabsts.pdf
  - This abstract presents findings on constructing social networks using coreference information and compares these networks to traditional co-presence-based networks.
- Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand (Dec. 2022). "Properties of Dramatic Characters: Automatic Detection of Gender, Age, and Social Status". In: *Computational Stylistics in Poetry, Prose, and Drama*. Ed. by Anne-Sophie Bories, Petr Plecháč, and Pablo Ruiz Fabo. Berlin/Boston: De Gruyter, pp. 179–202. DOI: 10.1515/9783110781502-010
  - In this publication, broad character properties, in particular age, gender and social status, are assigned to a number of German play characters and automatically classified. I was responsible for carrying out the classification experiments and retrieving the experimental results. I co-wrote the text with a focus on the experiment and result sections.
- Benjamin Krautter and Janis Pagel (2024, to appear). "The Schemer in German Drama. Identification and Quantitative Characterization". In: *Computational Drama Analysis. Reflecting Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin/Boston: De Gruyter
  - This paper presents results on automatically classifying a certain literary character type, the schemer. I was responsible for carrying out the classification experiments and gathering the experimental results presented. Benjamin Krautter and I collaboratively wrote the text.

During the preparation of this thesis, I was also involved in papers whose content and findings were not included in this thesis and which will not be listed here.

Theorie? Ich denk doch, wenn eine Sache praktisch wird, geht's an die Anwendung von Theorien.
(Theory? I think that when something becomes practical, it's time to apply theories.)

Schenk in Erich Mühsam's "Judas"

2 Background

This chapter gives an overview of recurring concepts and definitions that appear throughout the thesis. It mainly focuses on possible definitions of the fields of digital humanities, computational linguistics and computational literary studies, in which this thesis places itself. Furthermore, necessary concepts from literary studies, linguistics and computational linguistics are explained, summarized and presented, in particular dramatic texts, coreference, coreference resolution, evaluation in machine learning and social network analysis.

2.1. Digital Humanities and Computational Literary Studies

2.1.1. Digital Humanities

Digital humanities is an umbrella term encompassing several different trends and subfields. In general, it can be used for any research or practical work that deals with using and applying digital media, methods and/or computational resources on humanities research questions or objects of interest. In principle, there is no limitation on the kind of humanities research that can use methods of the DH and hence vastly different humanities fields can be found in the DH, like literary studies, history, philosophy, political and social sciences, religious studies, musicology or archaeology. Since this thesis deals exclusively

with literary texts, the following descriptions of DH will focus on this type of research. The larger amount of work in the DH which is concerned with literary texts can be grouped into the following fields:

1. Digital editorial work
2. Representation of humanities-related knowledge and research in digital forms, e.g. on websites, searchable online archives, etc.
3. Processing of literary texts with computational methods

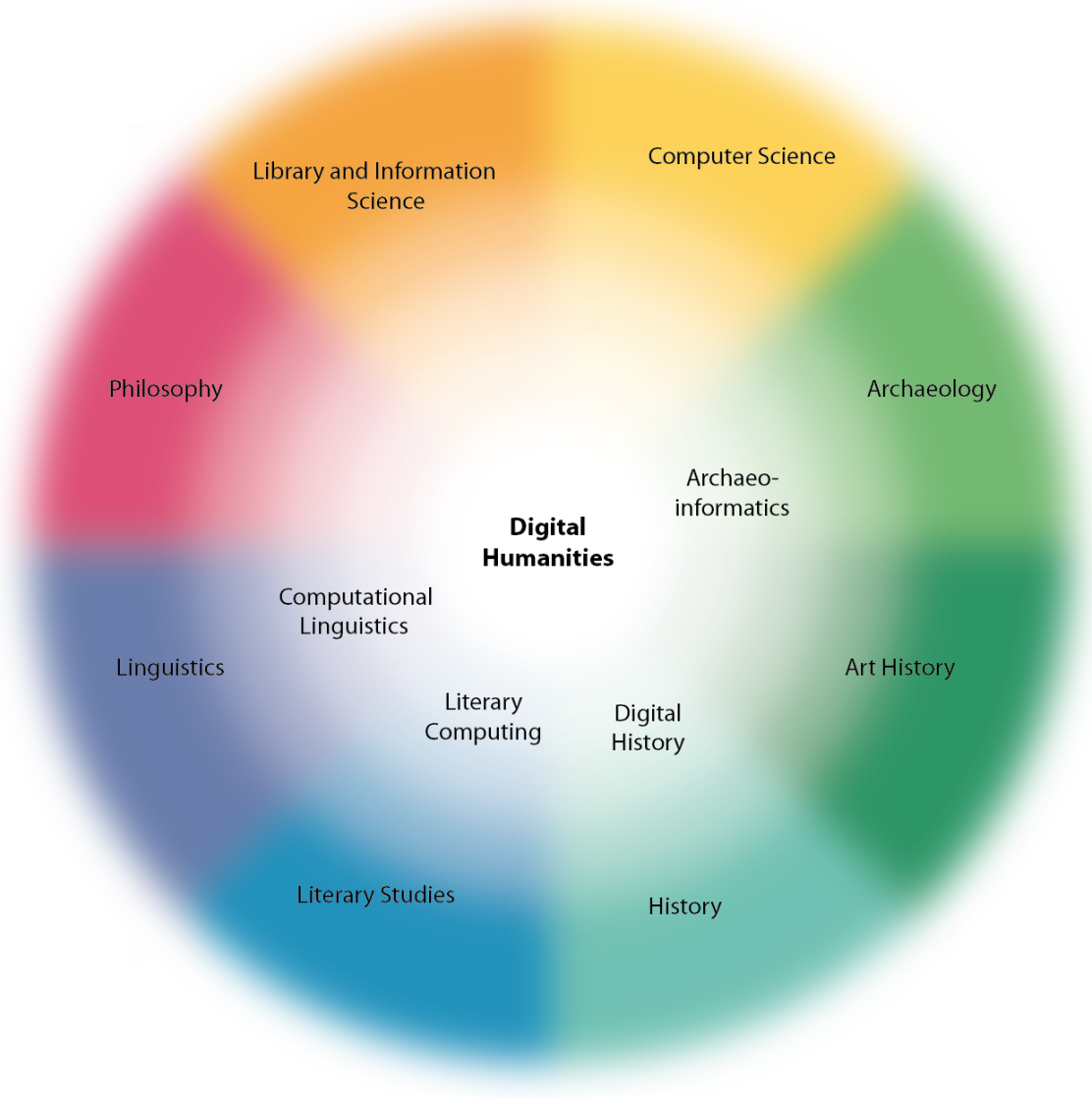


Figure 2.1.: "3-spheres model" from Sahle (2015).

Figure 2.1 shows an attempt by Sahle (2015) to position the DH inside a spectrum of other disciplines. According to Sahle (2015), the image can be read as containing three spheres, where the center represents DH as its own discipline, the middle sphere the disciplines which were digitally transformed, i.e. which separated from their original humanities origins and the outer sphere the humanities disciplines which use digital methods. The borders are blurred, since DH overlaps with all the spheres and disciplines in ways that are not always obvious.

2.1.2. Computational Literary Studies

In the past couple of years, Computational Literary Studies (CLS) emerged as a subfield of literary-text-focused DH with a stronger focus on computational processing and experimentation of and on literary text, in contrast to research concerned with for example the creation of digital editions and resources. CLS is still a developing field, but a general goal of CLS that is already apparent is the better understanding of computational methods applied to literary texts. Hereby, the research interest can be two-fold: (i) a better understanding of the methods used, for example within the larger concept of *explainable AI*, (ii) a better understanding of the specific literary texts, genres or epochs under investigation.

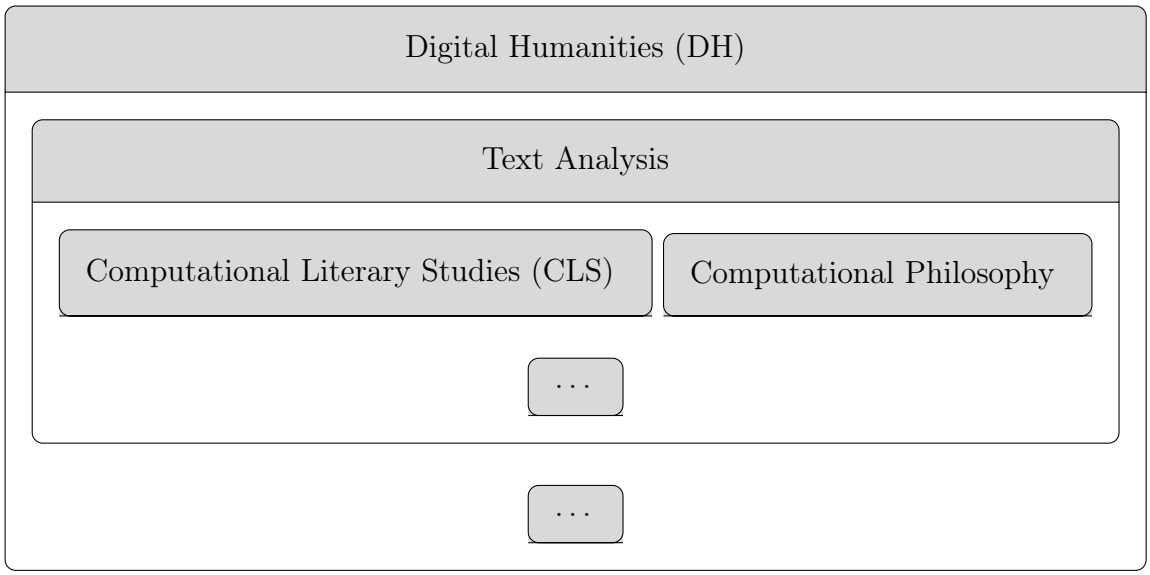


Figure 2.2.: Venn diagram of CLS as a sub-discipline in DH.

Figure 2.2 is an attempt to visualise a general possible understanding of how CLS relates to the digital humanities: In this Venn diagram, CLS is a proper subset of DH in a broader

subfield of DH, which could be labeled as text analysis (in comparison to other possible subfields of DH such as digitalization, image analysis, etc. and as opposed to subfields in text analysis which do not cover literary texts, e.g. political debates etc.) Another example of a subfield of DH concerned with text analysis is given as Computational Philosophy, however, many more fields are possible. Note that it would also be possible to understand CLS as its own field or as a subfield of computational linguistics, also dependent on the focus of the research done under the label of CLS. Another aspect that the diagram does not cover is that CLS (as well as computational philosophy) are not only concerned with text analysis but can also analyze other form of data, such as text performance, text reception, biographical data, etc. Hence, other subfields next to text analysis not depicted here might also contain CLS as a (sub-)subfield.
In Germany, there has been a recent attempt to anchor CLS more firmly into the research landscape via a DFG (Deutsche Forschungsgemeinschaft, German Research Foundation) funded priority programme[1]. The programme hosts several research projects related to CLS and promotes work of the projects in the community via general meetings and several forms of dissemination like workshops and blogs.
Also note that CLS has been under attack for allegedly not delivering results which are generally usable by or interesting to literary scholars (Da 2019, pp. 604, 638–639). Such claims would also imply that some literary scholars see the main goals of CLS in the development of computational methods (in their view blindly applied to literary texts, see point (i) above), instead of the furthering of an understanding of literary texts (see point (ii) above). The discussion and the defense of CLS by CLS researchers that followed from this has also sparked further discussions on how to define CLS as a field and what its goals and methods might exactly be (for instance Piper 2020; Jannidis 2020).

2.2. Drama

2.2.1. Drama as a Concept in Literary Studies

Drama is one of three major traditional literary genres, the other two being prose and poetry (Aristoteles 1982). Drama itself is often divided into two sub-genres: Comedy and tragedy (Aristoteles 1982). Pfister (1988, chap. 1) presents several characteristics that, through history, have been proposed to distinguish dramatic texts from prose:

[1] `https://dfg-spp-cls.github.io/`, `https://gepris.dfg.de/gepris/projekt/402743989/`

1. Dramatic texts do not contain a narrator through which the author (might) speak to their audience, but only characters that are speaking (Pfister 1988, pp. 2–3)
2. Dramatic texts are multi-modal in that they are performed and during their performances contain visual and acoustic information (Pfister 1988, pp. 6–7)
3. In contrast to other similar performance activities, performances of dramatic texts are ritualized and non-spontaneous. Pfister (1988) however also notes that more unstructured performances might be seen as predecessors to drama and the lines are often blurred (Pfister 1988, pp. 11–12)

Generally, it should be noted that these are not clear-cut properties with which every text can be identified as drama, but rather indicators on a spectrum. For instance, some texts are considered to be dramatic texts that weren't (originally) conceived to be performed,[2] such as Friedrich Schiller's *Die Räuber* (see Schiller's foreword in Schiller 1781/2017, pp. 3–7) or texts that are not only focused on dialogue and also contain epic modes of narration, such as Bertolt Brecht's *epic theatre*[3] (see also Pfister 1988, chap. 3.6 for a more in-depth discussion).

In this thesis, the focus will be put on drama as text and hence the performative aspect of drama will not play a role for the following chapters, or only a marginal one.[4]

Pfister (1988) identifies several structural properties of drama:

- Primary text vs. secondary text, where the former is the spoken dialogue exchanged between characters and the latter is all non-spoken text, such as the title, forewords, dramatis personæ, stage directions and speaker indications. The secondary text is typically typographically distinct from the primary text (Pfister 1988, chap. 2.1.2)
- The dramatis personæ is a list of all characters occurring on stage during a play (also non-speaking characters). Pfister (1988) uses the term to refer to the abstract list of characters as well as the actually printed list at the beginning of a play. Usually the characters which are only spoken about are not considered to be part of the dramatis personæ, however Pfister (1988, pp. 164–165) remarks that these characters also have the ability to influence the plot or be characterized in a certain way. Pfister (1988) also identifies several structuring properties of general dramatis personæ like social status, gender and age, however, as he notices, the structure of dramatis personæ varies greatly between different texts, epochs and genres (Pfister

[2] So called *closet drama* or *Lesedrama* in German

[3] In German *episches Theater*

[4] It should be noted that the performative aspects of drama play a larger role in literary science, e.g. Pfister (1988, pp. 13–38) dedicates a full chapter to it and the interest of scholars in the relationship between text and performance has only increased since then (compare e.g. Worthen 1998; McIntyre 2008; Kallenbach and Kuhlmann 2018)

1988, chap. 5.3.1).

- Stage directions may refer to actors, e.g. entrances and exits, mime and gestures etc. or to the surroundings like the way the set is supposed to look like (Pfister 1988, chap. 2.1.3). Pfister (1988, chap. 2.1.4) also describes the possibility of implicit stage directions, e.g. when characters describe the actions they are performing within the primary text.
- The division of the play into scenes and acts. While these divisions are usually marked inside the text, there is no consistent definition on when a scene or act boundary takes place. Pfister (1988) points out several historical and national differences in how scene and act divisions have been handled. However, there are certain properties that authors have often used to base their separation into scene and acts on, e.g. a change in configuration[5] (either partial or complete), a change in setting, time, etc. or a change in mood or to indicate important changes in the plot (Pfister 1988, chap. 6.4.1–6.4.2).

While this list is in no way complete, it comprises the most important aspects that will become relevant later on in this thesis.

2.2.2. Computational Drama Analysis

Early research in Computational Drama Analysis (CDA) has been focused on *social network analysis* (Moretti 2011; Trilcke, Fischer, and Kampkaspar 2015; Fischer et al. 2017; Lee and Lee 2017; Krautter 2023). Social network analysis makes use of graph theoretical assumptions and measures to quantify relationships of actors in social settings. In the context of drama, a node in such a network usually represents a character appearing on stage and edges between nodes represent if a character is co-present with another character, i.e. if both characters appear on stage at the same time. The edges might also be weighted, with higher weights indicating that two figures co-appeared more often on different occasions. If a network is constructed in such a way as described above, it can also be called a *co-presence or co-occurrence network*. Common measures to quantify and compare the relationships of nodes in such networks include degree, weighted degree, betweenness, closeness and eigenvector centrality (Newman 2010, chap. 7).
Apart from co-presence, other aspects of drama have been studied quantitatively, such as stage directions (Maximova and Fischer 2018; Trilcke et al. 2020), protagonists (Fischer et al. 2018; Krautter et al. 2018; Reiter et al. 2018), character speech (Krautter 2018),

[5] *Configuration* is a term that Pfister (1988) uses to describe the current set of characters on stage.

sentiment (Nalisnick and Baird 2013; Schmidt and Burghardt 2018; Schmidt et al. 2019) and emotion (Saif 2011; Yavuz 2020; Schmidt, Dennerlein, and Wolff 2021a; Schmidt, Dennerlein, and Wolff 2021b; Dennerlein, Schmidt, and Wolff 2023).

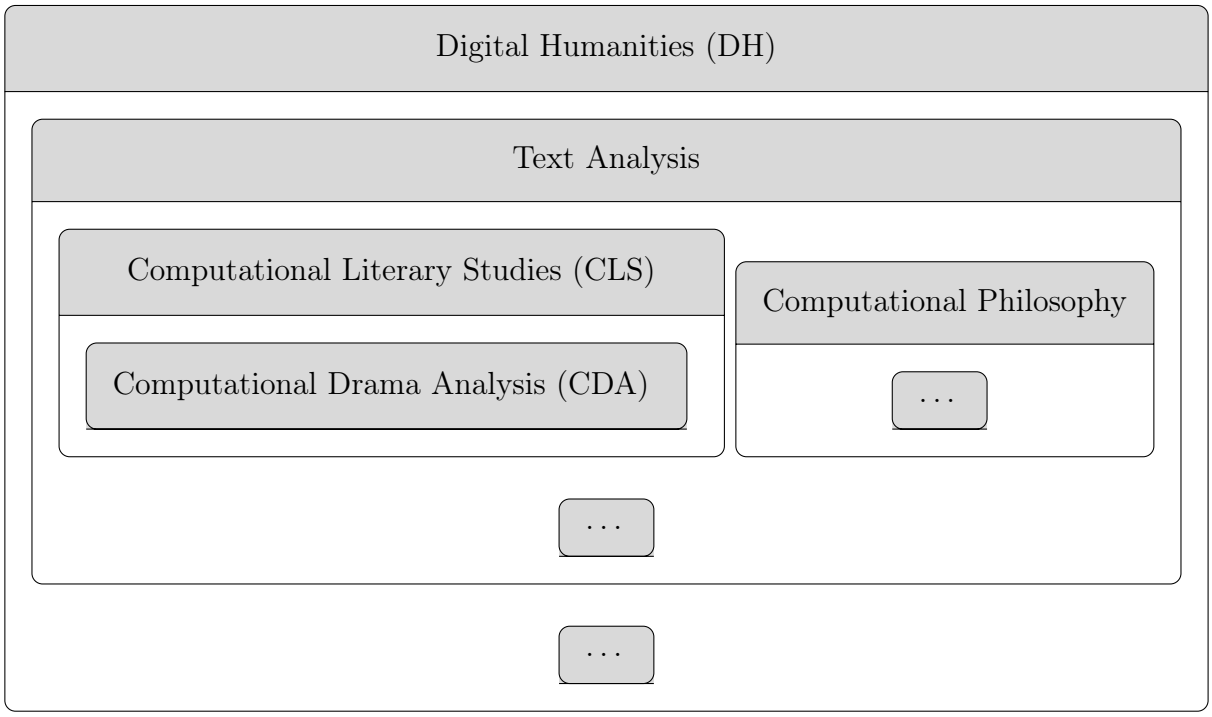


Figure 2.3.: Venn diagram of CDA as a sub-discipline in DH.

Figure 2.3 extends the Venn diagram of Figure 2.2 and locates CDA as a subfield of CLS.

2.3. Coreference

2.3.1. Coreference from a Linguistic Perspective

Coreference is a phenomenon that occurs when two or more mentions in a discourse refer to the same entity of this discourse (Reinhart 1983; Lasnik 1989; Crystal 2008, pp. 116–117; Halliday and Hasan 2013, p. 3). In this case, we say that the mentions *co-refer*. While coreference generally describes mentions as equal participants in a coreferential relationship, it is often useful to look at the relationship two mentions can have to one another. In general, a mention that occurs later in a text and therefore refers back to a previous mention can be called an *anaphor* as it links back to a previous mention *anaphorically* and the mention referred back to can be called *antecedent* (Lyons 1977,

p. 659; Crystal 2008, pp. 25–26). An illustrative example of coreference is given in Figure 2.4. Shown is the prologue of William Shakespeare's *Romeo and Juliet*.

THE PROLOGUE

Enter $[Chorus]_0$.

$[\text{Two households}]_1$, $[\text{both}]_1$ alike in dignity
(In fair $[\text{Verona}]_2$, $[\text{where}]_2$ $[\text{we}]_3$ lay $[\text{our}]_3$ scene),
From ancient grudge break to new mutiny,
Where civil blood makes civil hands unclean.
From forth the fatal loins of $[\text{these two foes}]_1$
$[\text{A pair of star—crossed lovers}]_4$ take $[\text{their}]_4$ life;
$[\text{Whose}]_4$ misadventured piteous overthrows
Doth with $[\text{their}]_4$ death bury $[[\text{their}]_4 \text{ parents'}]_5$ strife.
The fearful passage of $[\text{their}]_4$ death—marked love
And the continuance of $[[[\text{their}]_4 \text{ parents'}]_5 \text{ rage}]_6$,
$[\text{Which}]_6$, but $[[\text{their}]_5 \text{ children's}]_4$ end, naught could remove,
Is now the two hours' traffic of $[\text{our}]_3$ stage;
The which, if you with patient ears attend,
What here shall miss, $[\text{our}]_3$ toil shall strive to mend.

$[Chorus]_0$ *exits.*

Figure 2.4.: DraCor (2020): *Shakespeare, William. Romeo and Juliet.* extended with markup showing coreference relations.

Mentions are indicated by squared brackets, marking the words the respective mentions span, and mentions that refer to the same entity are additionally marked with identical subscripts. For example, in the line *A pair of star-crossed lovers take their life*, *A pair of star-crossed lovers* and *their* refer to the same supra-textual entity, namely the lovers described. This immediately leads to three observations: (i) entities do not need to be real world entities, as the two lovers Romeo and Juliet described here are purely fictional, (ii) mentions may refer to entities which themselves describe entities, e.g. here the term *lovers* refers to Romeo and Juliet, which themselves are entities to which may be referred, (iii) mentions are full noun phrases as in the case of *A pair of star-crossed lovers*, i.e. the mention includes all parts the noun phrase is composed of, such as articles (*a*), adjectives (*star-crossed*) and embedded prepositional phrases (*of star-crossed lovers*). Note that

in the given example, only mentions are marked which have at least one coreferential connection, i.e. at least two mentions refer to the same entity. However, basically almost all noun phrases in the text can potentially co-refer. Noun phrases that do not refer, so called *expletives*, are usually pronouns that are stand-ins for the syntactic subject of a sentence; for example in *It rains.*, the pronoun *it* is an expletive that does not refer to any entity, but has only the functional purpose of giving the sentence a subject. As there are many different perspectives on coreference and the correct annotation of coreference, the following will give an outline of the most important concepts related to coreference, explain the view on coreference taken in this work and, where ever, possible mention alternative possibilities.

Mentions Generally, all noun phrases (NPs) are considered mentions if they are referring. While there are different understandings of what *referring* means in different schools of linguistics, we will understand it as described in Bußmann (2008, pp. 573–74), i.e. as denoting objects, places, properties or events which can be part of the real world or a projected, imagined world. The phrases denoting referents are typically nominal phrases when concerning objects and places. There are also non-referring nominal phrases, which we pragmatically set to be idiomatic and expletive expressions, following Riester and Baumann (2017, p. 12).

Entities Entities are viewed as sets of mentions, the entirety of mentions in a set constitute the entity. We can also speak of the set of mentions as a *coreference cluster*, as the mentions form a cluster representing the entity and all mentions in a cluster are coreferent.
Another view on coreference would be the metaphor of a *coreference chain* that is spanning the text and connecting the mentions. While seemingly having the same end results, there are subtle conceptual differences. While the clustering view puts the focus stronger on the entity, and the mentions merely happen to appear at certain points in the text, the chain view puts a stronger focus on the mentions in the text which constitute the entity by appearing in certain positions in the text (see also Lyons 1977, p. 660). Adopting the chain view, it is also possible to view coreference as a phenomenon, in which coreferences only exist between anaphors and preceding antecedents. This purely anaphorical view has been used in early works on ML based coreference resolution and is known as the *mention-pair model* (cf. Soon, Ng, and Lim 2001).

Expletives Expletives are non-referring noun phrases, usually pronouns in certain contexts such as *it* in the English sentence *It rains* (Crystal 2008, p. 179). These pronouns are also called *dummy pronouns*, since they only fulfil a syntactic purpose (filling the subject position) without contributing to the semantic content of a sentence. Expletives can neither refer nor co-refer and are hence not considered as mentions in this work.

Non-nominal antecedents and anaphors Coreference clusters involving non-nominal antecedents, in the literature also referred to as *abstract anaphora* (cf. Kolhatkar et al. 2018) or discourse deixis (Webber 1988), are coreference clusters that contain at least one mention that is not nominal. Typical non-nominal antecedents are verbal phrases, sentences or sequences of sentences. Often, pronouns such as *that* or *it* do not refer to other NPs, but rather to previous statements. Consider the following example from Kolhatkar et al. (2018, p. 551):

> Anna finally **made her butternut squash recipe** this morning. **It** took her twenty minutes.

Here, the pronoun *it* refers to the process of Anna finishing the recipe, i.e. to a verbal phrase. *Shell noun phrases*, so called because they contain *shell nouns* (cf. Schmid 2000), such as *this fact*, are also possible contenders for referring to non-nominal antecedents (Kolhatkar et al. 2018, pp. 559ff.).

Predicative constructions Predicative constructions are combinations of a subject, a copula and a predicative (Bußmann 2008, pp. 542–543; Crystal 2008, pp. 381–382). The copula and the predicative form the predicate of the clause and describe a property of the subject. An example is *She is a teacher*, where *she* is the subject, *is* functions as a copula verb and *a teacher* is the predicative. While it may first look like *she* and *a teacher* are coreferent, *a teacher* functions as an attribution; a possible paraphrase would be "She has the property of being a teacher". Borthen (2004) presents two major arguments for why predicative noun phrases and their subject do not co-refer: (i) the reference of the predicative can not always be resolved, even if the reference of the antecedent (the subject) is known, (ii) predicatives are often non-nominal and if nominal, they are often indefinite. Van Deemter and Rodger (2000) advice against the annotation of predicative constructions as coreferent and criticize its annotation as coreferent as it was done for the MUC-6 and 7 conferences.

Bridging Bridging is a phenomenon that is related, but not identical, to coreference and denotes the occurrence of anaphors that are not coreferent to a preceding antecedent. Nonetheless, the bridging antecedent is necessary in order to make sense of the reference of the anaphor. For example, in the sentences *Lucy stood in front of a big house. The door was open.*, the reference of the term *The door* can only be properly resolved if it can be inferred that this door is part of the aforementioned house (see for example Clark 1975; Hou 2016; Riester and Baumann 2017; Rösiger 2019). Since bridging and coreference are related, but not identical, we will not consider cases of bridging in all following analyses.

Binding Since Chomsky (1993), several rules have been formalized that describe the interplay of certain syntactic constructions with the choice of co-referring pronouns. Chomsky noticed that the choice of reflexive and non-reflexive pronouns when referring back to an antecedent is not arbitrary, but depends on the syntactic structure of a sentence. In simplified terms, the binding theory of Chomsky states that an expression β binds an expression α if and only if β c-commands α and α and β are coindexed (Chomsky 1993, p. 184) and that different types of expressions must or must not be bound; in this theory, anaphors are always bound and pronominals and R-expressions are never bound (Chomsky 1993, p. 188) Notice that the term "coindexed" in Chomsky (1993) is called "coreferent" in this work. Furthermore, Chomsky refers to all reflexive and reciprocal pronouns as *anaphors* (*himself*, *herself*, *each other*, etc.), to all other types of pronouns as *pronominals* (*he*, *she*, *him*, *her*, etc.) and to referring NPs as *R-expressions* (e.g. proper NPs such as *wood* or *book* and proper names such as *John* or *Mary*, etc.) (Chomsky 1993, pp. 101–102). "C-command" is a type of syntactic relationship, in which a node β c-commands another node α in a syntactic phrase-structure tree, if one can go up a node to a parent node γ, starting from the node β, and then find the node α further down by traversing the sub-trees of node γ (refer to Chomsky (1993, p. 166) for a formal definition of c-command). Using binding theory, one can easily explain why certain English sentences are ungrammatical:

(1) Mary_1 saw herself_1.
(2) *Mary_1 saw herself_2.
(3) Mary_1 saw her_2.
(4) *Mary_1 saw her_1.
(5) Mary_1 saw Mary_2.
(6) *Mary_1 saw Mary_1.
(7) Mary_1 noticed that Amanda_2 liked her_1.

(8) *Mary$_1$ noticed that Amanda$_2$ liked herself$_1$.

(9) Mary$_1$ noticed that Amanda$_2$ liked herself$_2$.

Sentences (1) to (6) show examples with c-command. In (1) and (2), the sentence is only grammatical if *Mary* and *herself* are coreferent, since *herself*, as a reflexive pronoun, needs to be bound and it is only bound if *Mary* c-commands *herself* (which it does) and *Mary* and *herself* are coreferent (which they are in (1) but not in (2)). In (3) and (4), the sentence is only grammatical if *Mary* and *her* are not coreferent, since *her*, as a pronominal, must not be bound and in sentence (4) it is bound, since *Mary* c-commands *her* and they are coreferent. Analogously to (3) and (4), only sentence (5) is grammatical, since *Mary* as an R-expression must not be bound and in (5) the two entities called *Mary* are not the same person and thus do not bind each other. Only in sentence (6), where *Mary* and *Mary* are co-referent is the binding principle violated and the sentence is ungrammatical. Sentences (7) and (8) show examples where c-command is absent. Sentence (8) is ungrammatical, since the pronoun *herself* is chosen, which is not bound (it is coreferent with *Mary*, but not c-commanded by *Mary*; notice however that it is c-commanded by *Amanda*, but not co-referent with *Amanda*). The correct choice of pronoun is *her* in sentence (7), since *her* must not be bound and can therefore refer to *Mary* in a situation where c-command is absent. An alternative would be to make *Amanda* and *herself* co-referent, like given in sentence (9), since *Amanda* c-commands *herself* and if both are co-referent, *herself* is properly bound.

Reference to fictional characters Since this thesis is dealing with literary, hence fictional, texts and the coreference of fictional characters, it seems natural to have a look at the linguistic literature regarding the reference to fictional characters. Kamp (2021) argues that the way people refer to fictional characters is similar to the way they refer to non-fictional people and that using fictional names and real-world names share the same properties in his *Mental State Discourse Representation Theory* within the formalism of the *Entity Representation*. In this framework, coreferential relations work the same, irrespective of if they are about real-world entities or fictional entities. While the linguistic features that fictional and non-fictional characters are referred to appear to be identical (Kamp 2021, p. 38), claims have been made that readers of fiction are still able to relate the purely linguistic encoding of fictional characters to real world historical persons (Maier 2017, pp. 23–24) and human experience in general, albeit only as a virtual existence (Cohan 1983). Mead (1990) argues that the stylistics of fictional characters and how the characters are represented stylistically within the fictional texts also plays a

large role in how readers perceive fictional characters. Köppe (2020) identifies anaphoric references as one possible way of reference within a literary text as part of intratextual references and juxtaposes it to intrafictional references, in which characters can refer to other characters within the narrated world.

2.3.2. Coreference Annotation

While the linguistic perspective on CR often operates on singular observation and special and interesting cases, the automatic resolution of coreference requires large datasets with annotated coreferences. To this end, different tools and methods for carrying out this task have been proposed, as well as different datasets annotated. This section gives an overview of some of these methods, tools and datasets dedicated to the annotation of coreference and with the goal of CR in mind.

Datasets

The following gives a non-exhaustive overview of corpora that were deemed relevant for the context of this thesis. We consider a corpus as relevant if it contains coreference annotations for German-language texts, for literary (but not necessarily German) texts, and lastly if it can be used for size comparisons. The corpora are ordered alphabetically.

CRETA An unnamed and unpublished corpus by Rösiger, Schulz, and Reiter (2018) contains coreference annotations on German fairy tales and novellas and will be referred to as *CRETA* corpus throughout this thesis, since it has been developed in the context of the CRETA project[6].

DIRNDL (Björkelund et al. 2014; Eckart, Riester, and Schweitzer 2012) is a corpus of German radio news, containing, next to phonetic annotations, coreference and information status annotations. DIRNDL currently contains 3221 sentences and ca. 50 000 tokens.

DROC (Krug et al. 2018; Krug and Zehe 2018) is a corpus of German novel fragments annotated for coreference, however, only characters were annotated and of these characters, only heads served as markables. DROC comprises of ca. 2000 annotated passages.

GerDraCor-Coref (Pagel and Reiter 2020; Pagel 2022b) will be described in detail in Chapter 4.

[6] `https://www.creta.uni-stuttgart.de/en`

GRAIN (Schweitzer et al. 2018; Schweitzer, Eckart, and Gärtner 2018) is a corpus of radio interviews and contains many annotations including coreference, information status and information structure. The corpus contains 144 interviews with 221 000 tokens.

OntoNotes is one of the largest datasets containing coreference annotations (Pradhan et al. 2007; Weischedel et al. 2013). Next to coreference annotations, OntoNotes contains syntactic and named entity information. OntoNotes' annotated text type are mostly newspaper articles in English, Chinese and Arabic language (2.9 million words in version 5.0).

OpenBoek (Van Cranenburgh and van Noord 2022; van Cranenburgh 2022) contains nine original and translated Dutch novels from Project Gutenberg. In total, the corpus comprises of 103 000 tokens. The texts were preprocessed using the system by van Cranenburgh (2019a) and the resulting mentions and coreferences were manually corrected.

ProppLearner (Finlayson 2017; Winston and Finlayson 2015) is a corpus of Russian folk tales and contains 18 862 tokens, including coreference annotations.

RiddleCoref (Van Cranenburgh 2019a; van Cranenburgh 2019b) comprises of Dutch novels and consists of 21 novels with ca. 107 000 tokens.

TüBa-D/Z (Telljohann, Hinrichs, and Kübler 2004; Hinrichs et al. 2009) is the to date largest German corpus with coreference annotations and is made up of newspaper articles. TüBa-D/Z was primarily conceptualized as a syntactic treebank, but has received a multitude of other annotations over the years, including coreference (Naumann 2007). In version 11.0, TüBa-D/Z comprises of 3816 articles, 104 787 sentences and 1 959 474 tokens.

Table 2.1 gives a quick overview of the discussed corpora, their annotated languages, main domain and main citation.
Three of these corpora (CRETA, DIRNDL and TüBa-D/Z) are later used in Chapter 5 to compare CR results with the GerDraCor-Coref corpus.

Name	**Language(s)**	**Domain(s)**	**# Tokens**	**Citation**
NA (CRETA)	German	Fairytales, Novellas	*NA*	Rösiger, Schulz, and Reiter (2018)
DIRNDL	German	Radio News	50 000	Björkelund et al. (2014)
DROC	German	Novels	393 000	Krug et al. (2018)
GerDraCor-Coref	German	Drama	542 421	Pagel and Reiter (2020)
GRAIN	German	Radio Interviews	221 000	Schweitzer et al. (2018)
OntoNotes	English, Chinese, Arabic	Newspaper	2 900 000	Pradhan et al. (2007)
OpenBoek	Dutch	Novels	103 000	Van Cranenburgh and van Noord (2022)
ProppLearner	Russian	Folktales	18 862	Finlayson (2017)
RiddleCoref	Dutch	Novels	107 000	Van Cranenburgh (2019a)
TüBa-D/Z	German	Newspaper	1 959 474	Naumann (2007)

Table 2.1.: Overview of corpora mentioned in this thesis containing coreference annotations.

Tools

One of the earliest tools that can be used to annotate coreferences is MMAX (later MMAX2) (Müller and Strube 2003; Müller and Strube 2006). The tool is general-purpose in that its purpose is the annotation of general linguistic and multi-level information, but it has been used especially frequently for coreference annotation. MMAX saves annotated markables in a XML format and allows for assigning custom properties to markables. MMAX is implemented in Java.
The OntoNotes dataset has been created using the Callisto annotation tool, which is no longer under active development.[7]
ATHEN is a Java-based annotation tool developed within the Kallimachos project at the University of Würzburg and has been used to annotate the DROC corpus.[8]
CorefAnnotator (Reiter 2018) is especially designed for the annotation of coreferences. The tool distinguishes a text view, in which spans can be annotated and an entity view, in which entities and annotated mentions can be manipulated. CorefAnnotator is implemented in Java and saves annotations in the XML interchange format XMI.
In principle, there are many other multi-purpose annotations tools capable of supporting coreference annotations, for which only an exemplary list is given below for reference and as a pointer for further investigation:

- Slate (Kaplan et al. 2011)
- Anafora (Chen and Styler 2013)
- WebAnno (Eckart de Castilho et al. 2016)
- INCEpTION (Klie et al. 2018)

Methods

Different methods have been developed in order to streamline the task of annotating coreferences, of which two will be discussed below.

The gamification of coreference annotation is an approach carried out by Chamberlain, Poesio, and Kruschwitz (2016), among others.[9] For their Phrase Detectives corpus, the authors let annotators play a game with the objective to find co-references in English texts. Annotators were presented with text snippets from Wikipedia[10] and

[7] `https://mitre.github.io/callisto/`
[8] `https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/Athen`
[9] Chamberlain, Poesio, and Kruschwitz (2016) discuss other, similar approaches of gamified annotation.
[10] `https://www.wikipedia.org`

Project Gutenberg[11] and pre-selected markables within these texts. The annotators were asked to carry out two different tasks: Firstly, deciding which of the presented markables were coreferent and secondly, making a decision on annotations of two other annotators who disagreed in their annotations. These two activities were presented in the context of working as a detective and called "Name the Culprit" and "Detectives Conference", respectively. Annotators were able to receive points for annotating a certain amount of texts and compare their annotation achievements through a high score. This approach is also an example of crowed-sourced coreference annotation. The game can be played (at the time of writing) on a dedicated website[12] and via an integrated Facebook application[13].

Model-based coreference annotation was proposed by Aralikatte and Søgaard (2020) and replaces the task of linking two text spans, i.e. markables, with linking a single text span to a pre-created entity in a knowledge base. This task is otherwise also known as *entity linking* (see for example Mihalcea and Csomai 2007; Zaporojets et al. 2022). The entities are created by parsing the to-be-annotated Wikipedia articles for links referring to NEs. Annotators were then asked to decide if a pronoun in the text refers to one of these Wikipedia pages. This approach is somewhat limited to texts were an easy extraction of pre-defined entities is possible and it is not clear how the approach can be broadened to entities which are not NEs.

2.3.3. Automatic Coreference Resolution

Coreference Resolution is a term for describing the process of (automatically) resolving all or a subset of mentions in a given text and indexing all mentions according to the entity they belong to.
While nowadays the term is generally understood as using computational and automated methods for resolution, it is in principal possible to create algorithms that resolve coreferences "by hand".
A task related to CR is the aforementioned entity linking, where references in a text are linked to entities within a database, for example to entries on Wikidata[14]. In contrast to entity linking, CR is however not concerned with linking references to an external database, but only with identifying which references co-refer within a single text.

[11] `https://gutenberg.org`
[12] `https://anawiki.essex.ac.uk/phrasedetectives`
[13] `https://apps.facebook.com/phrasedetectives`
[14] `https://www.wikidata.org`

Algorithmic approaches An early example of algorithmic coreference resolution is shown in Hobbs (1978). The algorithm only deals with pronoun resolution and uses information of richly annotated constituent parse trees and is shown in Figure 2.5. In the algorithm, one can already observe certain principles that should also later become adopted by other kind of methods, namely rule-based and machine learning based methods:

- Trees are traversed in a left-to-right, breadth-first fashion (compare Raghunathan et al. (2010), who also use this method)
- A certain notion of saliency, i.e. recent nominal phrases are preferred to be antecedents and previous sentences are more likely to contain the correct antecedent than sentences farther away
- Reflexive pronouns underlie certain constraints with respect to the possible parse tree embeddings their antecedents can have (see also *Binding Theory* (Chomsky 1981))

Mention-pair model During the 2000s, the so-called *mention-pair model* gets introduced and adopted by many works on automatic CR. It was originally proposed by Soon, Ng, and Lim (2001).[15] The mention-pair model identifies pairs of co-referring mentions, usually by using machine learning techniques. To achieve this, possible mentions are first identified and then compared according to certain features. The machine learning algorithm decides if the two mentions co-refer or not. In a second step, all pairs are then chained into full coreference chains by identifying which of the pairs co-refer among each other. This second step introduces many problems however, since only two pairs at a time are considered and there is the possibility of inconsistencies within the chain because of this. For example, if the phrases *Barack Obama* and *the President* were marked as coreferent as well as the two phrases *Joe Biden* and *the President*, the merging might falsely put all four mentions as coreferent based on the common phrase *the President*.

Entity-centric model The idea of using global features defined for entities in order to use this information for local classifications of coreferences is not new.
For instance, Gaizauskas et al. (1995) use a "world model", an ontology with associated attributes for each node in the ontology, to check if candidate mention pairs share a common path in the ontology and compute a similarity score between the properties of the candidate mention pair and other previously resolved pairs. Named entities are

[15] However, they do not call it *mention-pair model*. The name was coined by later publications.

1. Begin at the NP node immediately dominating the pronoun.
2. Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.
3. Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which has an NP or S node between it and X.
4. If node X is the highest S node in the sentence, traverse the surface parse tree of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If X is not the highest S node in the sentence, continue to step 5.
5. From node X, go up the tree to the first NP or S node encountered. Call this new node X, and call the path traversed to reach it p.
6. If X is an NP node and if the path p to X did not pass through the $\bar{\text{N}}$ node that X immediately dominates, propose X as the antecedent.
7. Traverse all branches below node X to the *left* of path p in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
8. If X is an S node, traverse all branches of node X to the *right* of path p in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.
9. Go to step 4.

Figure 2.5.: Pronoun resolution algorithm in Hobbs (1978, p. 341).

resolved across the whole text, pronouns within paragraphs and all other mentions within the same sentence or up to two previous paragraphs.
Daumé III and Marcu (2005) use entity-based features for the related task of entity linking, i.e. mapping textual mentions to real-world entities.
Lee et al. (2013) re-introduce rule-based systems into the research discourse,[16] following the rule-based work of Baldwin (1997), and attempt to reduce some of the disadvantages of the mention-pair model by using entity-based features. Their entity-centric model uses the information of all mentions in each coreference chain for each time a decision is made about adding a new mention to an existing chain. Consequently, for their approach they view coreference as a clustering task rather than a task of constructing coreference chains. In their approach, co-referring mentions are grouped into clusters and the information of each cluster is always available when successively adding new mentions into existing and growing clusters. They also introduce a sieve-based approach, for which passes ordered by precision decide on whether a mention is added to a cluster. If a pass accepts a mention as coreferent with a cluster, it is not visible to later passes anymore.

Neural network architectures Wiseman et al. (2015), Wiseman, Rush, and Shieber (2016), and Clark and Manning (2016) use the idea of the previously described entity-centric models and use neural network models to learn entity-level features. However, they rely on external syntactic parsers for mention extraction. Yu, Uma, and Poesio (2020) use the transformer-based BERT architecture to identify singletons and non-referring expressions alongside coreferences.

End-to-end architectures Lee et al. (2017) extend the neural network approaches by introducing the first end-to-end coreference resolution system. End-to-end means that the system is not build in a pipeline fashion where modules receive input from earlier steps. Specifically, the system handles the detection and resolution of mentions in one step, in contrast to the earlier neural network based approaches, for which the detection of mentions is a separate step. Lee, He, and Zettlemoyer (2018) use coarse-to-fine inference in their end-to-end approach, which entails pruning the antecedent-search-space to a manageable size. Joshi et al. (2019) make use of a transformer model, specifically BERT, building up on Lee et al. (2017)'s and Lee, He, and Zettlemoyer (2018)'s work and achieve good improvements on the OntoNotes dataset. Joshi et al. (2020) introduce *SpanBERT*, which modifies the transformer architecture BERT to score random spans

[16] Based on the previous systems by Raghunathan et al. (2010) and Lee et al. (2011).

in an end-to-end approach. Xia, Sedoc, and Van Durme (2020) make improvements on top of Joshi et al. (2020)'s model by reducing memory consumption via an incremental search approach.

Overview In general, the following approaches to coreference resolution can be observed throughout the history of CR: manually applied algorithmic approaches, rule-based automatic approaches, mention-pair models, entity-centric models, neural coreference resolution and end-to-end architectures. A recent overview of the field is provided by Sukthanker et al. (2020).

2.3.4. Metrics for Evaluating Coreference Resolution Systems

Comparing system output and gold annotations for coreference is a complex endeavour, since firstly mention spans need to be determined by both annotators and an automatic systems, leading to potential disagreements down the line and secondly mentions can be potentially assigned to any coreference cluster with the number of clusters not predefined, making the comparisons of coreference clusters a comparison of arbitrary sets with an undefined number of elements.
Consider the following example from Pradhan et al. (2014)[17] for a potential gold cluster K (*key*) and a predicted cluster R (*response*):

$$K = \{a, b, c\}\{d, e, f, g\} \tag{2.1}$$

$$R = \{a, b\}\{c, d\}\{f, g, h, i\} \tag{2.2}$$

where letters a-i represent mentions. R contains three clusters, while K only contains two clusters. Furthermore, both clusters contain mentions that the other cluster does not contain (K contains e, but R does not; R contains h and i but K does not). Lastly, some shared mentions are located in different clusters (in K, c is clustered together with a and b, in R, c is clustered with d and in K, d is clustered with e, f and g). All this poses severe challenges for traditional machine learning evaluation metrics like accuracy or precision, recall and F-score that in contrast operate on predefined data points with labels assigned to these data points.
Multiple metrics have been proposed over time to capture the accuracy of predicted

[17] The labels for the clusters that are present in Pradhan et al. (2014) have been omitted here, as they will not be used in the explanation of the metrics.

coreferences, for which several, often-used metrics will be discussed below.

MUC Vilain et al. (1995) propose a metric for the coreference task in the 6^{th} *Message Understanding Conference* (MUC-6). They count links that are missing from the predicted cluster in order to transform it into the gold cluster as a measure of the recall and links that are missing in order to transform the gold cluster into the predicted cluster as a measure for precision. Here, links can be understood as connections that signify that two mentions belong into the same coreference cluster.
This gives the following formulæ for precision and recall:

$$\text{MUC-RECALL} = \frac{\sum(|S| - |p(S)|)}{\sum(|S| - 1)} \tag{2.3}$$

$$\text{MUC-PRECISION} = \frac{\sum(|S'| - |p'(S')|)}{\sum(|S'| - 1)}, \tag{2.4}$$

where $|S|$ is the number of mentions in a particular gold cluster and $p(S)$ is a function that returns the clusters of the prediction that overlap with this gold cluster plus any singleton clusters containing mentions that are in the prediction, but not in the gold cluster. Doing this for all gold clusters and summing up the individual numbers yields the final recall score. Precision is defined as the inverse, with $|S'|$ being the number of mentions in a predicted cluster and $|p(S')|$ the number of clusters when intersecting the gold clusters with this respective predicted cluster, plus potential singleton clusters of mentions that are in the gold clusters but not in the predicted cluster. Once again, summing over all predicted clusters yields the final precision score.
Although not explicitly discussed in Vilain et al. (1995), the F1-score of precision and recall is commonly defined as the harmonic mean of the two and thus calculated as

$$\text{MUC-F1-SCORE} = \frac{2 \times \text{MUC-PRECISION} \times \text{MUC-RECALL}}{\text{MUC-PRECISION} + \text{MUC-RECALL}} \tag{2.5}$$

B^3 Baldwin et al. (1998) and Bagga and Baldwin (1998) present a new coreference scoring metric, which is also described in Pradhan et al. (2014). B^3 precision is calculated by assigning scores to predicted mentions, which are determined by dividing the number of predicted mentions in a predicted cluster contained in a gold cluster by the overall number of mentions contained in this predicted cluster. For B^3 recall, the roles of predicted and gold mentions and clusters are reversed (Pradhan et al. 2014, pp. 33–34).[18]

[18]This is a similar process as for MUC.

The equations for retaining B^3 precision and recall are as follows:

$$\text{B}^3\text{-RECALL} = \frac{\sum_{i=1}^{N_r}\sum_{j=1}^{N_s}\frac{|R_i \cap S_j|^2}{|R_i|}}{\sum_{i=1}^{N_r}|R_i|} \tag{2.6}$$

$$\text{B}^3\text{-PRECISION} = \frac{\sum_{i=1}^{N_r}\sum_{j=1}^{N_s}\frac{|R_i \cap S_j|^2}{|R_i|}}{\sum_{i=1}^{N_s}|S_i|} \tag{2.7}$$

$$\tag{2.8}$$

R is a set representing a gold cluster and S a set representing a predicted cluster. N_r is the number of all gold clusters and N_s the number of all predicted clusters.
Like for MUC, the F1-score is the harmonic mean of precision and recall:

$$\text{B}^3\text{-F1-SCORE} = \frac{2 \times \text{B}^3\text{-PRECISION} \times \text{B}^3\text{-RECALL}}{\text{B}^3\text{-PRECISION} + \text{B}^3\text{-RECALL}} \tag{2.9}$$

CEAF$_m$ and CEAF$_e$ Luo (2005) identifies flaws in the MUC and B^3 scores and proposes a *Constrained Entity-Aligned F-Measure* (CEAF). Luo (2005) introduces a similarity function $\phi(\cdot)$ that can be set to calculate different forms of similarity between entities. The general formula for CEAF is:

$$\text{CEAF-RECALL} = \frac{\phi(g^*)}{\sum_i \phi(R_i, R_i)} \tag{2.10}$$

$$\text{CEAF-PRECISION} = \frac{\phi(g^*)}{\sum_i \phi(S_i, S_i)} \tag{2.11}$$

$$\text{CEAF-F1-SCORE} = \frac{2 \times \text{CEAF-RECALL} \times \text{CEAF-PRECISION}}{\text{CEAF-RECALL} + \text{CEAF-PRECISION}}, \tag{2.12}$$

where R_i is a cluster from a set of gold clusters R, S_i a cluster from a set of predicted clusters S[19] and g^* is a one-to-one mapping between the clusters in R and S which maximises similarity. Since $\phi(\cdot)$ denotes similarity between clusters, $\phi(R_i, R_i)$ is simply the number of mentions in R_i (and analogously $\phi(S_i, S_i) = |S_i|$) and $\phi(R, R)$ is the number of entities in R (and analogously $\phi(S, S) = |S|$).
The metric defines a mention-based calculation (CEAF$_m$) and an entity-based calculation (CEAF$_e$), by setting $\phi(\cdot)$ to be $\phi(R, S) = |R \cap S|$ in the former and $\phi(R, S) = \frac{2|R \cap S|}{|R|+|S|}$ in

[19] In Luo (2005), R stands for *response* and S for *system*.

the latter case. The mention-based metric's precision and recall is thus defined as

$$\text{CEAF}_m\text{-RECALL} = \frac{\sum_{j \times k \in g^*} |R_j \cap S_k|}{\sum_i |R_i|} \tag{2.13}$$

$$\text{CEAF}_m\text{-PRECISION} = \frac{\sum_{j \times k \in g^*} |R_j \cap S_k|}{\sum_i |S_i|} \tag{2.14}$$

,
and the entity-based metric's precision and recall as

$$\text{CEAF}_e\text{-RECALL} = \frac{\sum_{j \times k \in g^*} \frac{2 \times |R_j \cap S_k|}{|R_j| + |S_k|}}{|R|} \tag{2.15}$$

$$\text{CEAF}_e\text{-PRECISION} = \frac{\sum_{j \times k \in g^*} \frac{2 \times |R_j \cap S_k|}{|R_j| + |S_k|}}{|S|} \tag{2.16}$$

The indices j and k are the clusters of R and S which maximise similarity according to g^*. $\sum_i |R_i|$ and $\sum_i |S_i|$ are simply the sum of the number of all mentions in all clusters of R and S and $|R|$ and $|S|$ the number of clusters in R and S, respectively.
The calculation of the F1-scores CEAF_m-F1-SCORE and CEAF_e-F1-SCORE follows from Equation 2.12.

CoNLL/MELA The CoNLL score is an attempt to unify popular metrics, and is the average of the MUC, CEAF_e and B^3 scores:

$$\text{CoNLL-SCORE} = \frac{\text{MUC-SCORE} + \text{B}^3\text{-SCORE} + \text{CEAF}_e\text{-SCORE}}{3} \tag{2.17}$$

The metric was used in the context of the CoNLL-2011 (Pradhan et al. 2011) and CoNLL-2012 (Pradhan et al. 2012) shared tasks on CR. Note that the score was originally called MELA (Denis and Baldridge 2009), but is referred to as CoNLL-Score by many publications.

BLANC Recasens and Hovy (2011) attempt to redefine the Rand index (Rand 1971) for coreference evaluation and call their metric *BiLateral Assessment of Noun-Phrase Coreference* (BLANC). They note that applying the Rand index as it is leads to unintuitive results due to the potential outnumbering of singletons compared to clustered mentions in

CR. They redefine the Rand index as a balancing between coreference and non-coreference links and correct and wrong links. In their nomenclature, c is a link between two mentions that are coreferent, n is a link between two mentions that are not coreferent, r is a link for which the gold clusters and predicted clusters tell the same outcome (either both say the mentions have a c link or a n link) and w is a link for which gold and predicted clusters do differ. Precision and recall are defined as the mean between the precision and recall of coreferent and non-coreferent links and the F1-score as the average of the harmonic means of the coreferent and non-coreferent precision and recall:

$$\text{BLANC-RECALL} = \frac{\frac{rc}{rc+wn} + \frac{rn}{rn+wc}}{2} \tag{2.18}$$

$$\text{BLANC-PRECISION} = \frac{\frac{rc}{rc+wc} + \frac{rn}{rn+wn}}{2} \tag{2.19}$$

$$\text{BLANC-F1-SCORE} = \frac{\frac{2 \times \frac{rc}{rc+wn} \times \frac{rc}{rc+wc}}{\frac{rc}{rc+wn} + \frac{rc}{rc+wc}} + \frac{2 \times \frac{rn}{rn+wc} + \frac{rn}{rn+wn}}{\frac{rn}{rn+wc} + \frac{rn}{rn+wn}}}{2} \tag{2.20}$$

Luo et al. (2014) note that this measure does not work if the mentions in the gold and predicted outputs are not identical and propose a variation of BLANC that also holds for cases when mentions do not completely overlap due to the system finding other mentions than are annotated in the gold clusters. They introduce new counts $mc1$ (missing coreference links that are in gold, but not in the prediction), $mc2$ (missing coreference links that are in the prediction but not in gold), $mn1$ (missing non-coreferent links that are in gold but not in the prediction) and $mn2$ (missing non-coreferent links that are in the prediction but not in gold)[20].
The updated BLANC scores are then as follows:

$$\text{BLANC-RECALL} = \frac{\frac{rc}{rc+wn+mc1} + \frac{rn}{rn+wc+mn1}}{2} \tag{2.21}$$

$$\text{BLANC-PRECISION} = \frac{\frac{rc}{rc+wc+mc2} + \frac{rn}{rn+wn+mn2}}{2} \tag{2.22}$$

The calculation of the updated F1-score is analogous to the calculation in Equation 2.20.

[20]Nomenclature of Luo et al. (2014) changed to match the notation of Recasens and Hovy (2011).

Since the updated BLANC scores are the ones used in the implementation of Pradhan et al. (2014), they are also the ones reported in this thesis.

LEA The *Link-based Entity-Aware* metric (LEA) by Moosavi and Strube (2016) juxtaposes a score for the importance of an entity with a score for the goodness of resolution of this entity. The general formula for computing the LEA-score is given as

$$\frac{\sum_{e_i \in E}(\mathit{importance}(e_i) \times \mathit{resolution\text{-}score}(e_i))}{\sum_{e_k \in E} \mathit{importance}(e_k)}. \tag{2.23}$$

importance$(\cdot)$ and *resolution-score*$(\cdot)$ are functions returning a value for the importance given to an entity and a value for the goodness of resolving the entity, respectively. Moosavi and Strube (2016) set *importance*$(\cdot)$ to be the size of an entity, i.e. the number of mentions within this entity:

$$\mathit{importance}(e) = |e| \tag{2.24}$$

and *resolution-score*$(\cdot)$ as the ratio of the number of correctly resolved links between two clusters and the number of links in one of the clusters:

$$\mathit{resolution\text{-}score}(b_i) = \sum_{a_j \in A} \frac{\mathit{link}(b_i \cap a_j)}{\mathit{link}(b_i)}, \tag{2.25}$$

where A is a set of clusters, b_i is a specific entity from another set of clusters and *link*$(\cdot)$ a function retrieving the number of links between mentions in an entity. Depending on if precision or recall should be calculated, the roles of the two clusters a_j and b_i are set to be either gold or predicted clusters.
By setting the values of Equation 2.24 and 2.25 into Equation 2.23, the LEA-scores for precision and recall can be computed:

$$\text{LEA-RECALL} = \frac{\sum_{k_i \in K}(|k_i| \times \sum_{r_j \in R} \frac{\mathit{link}(k_i \cap r_j)}{\mathit{link}(k_i)})}{\sum_{k_z \in K} |k_z|} \tag{2.26}$$

$$\text{LEA-PRECISION} = \frac{\sum_{r_i \in R}(|r_i| \times \sum_{k_j \in K} \frac{\mathit{link}(r_i \cap k_j)}{\mathit{link}(r_i)})}{\sum_{r_z \in R} |r_z|}, \tag{2.27}$$

$$\tag{2.28}$$

where K is a set of gold clusters and R a set of predicted clusters. As for other metrics, the F1-score for LEA is the harmonic mean of precision and recall:

$$\text{LEA-F1-SCORE} = \frac{2 \times \text{LEA-PRECISION} \times \text{LEA-RECALL}}{\text{LEA-PRECISION} + \text{LEA-RECALL}} \tag{2.29}$$

See Pradhan et al. (2014) for a further discussion of all these metrics (except for the LEA score) with example calculations and a link to a reference implementation for computing them (including the LEA score). Cai and Strube (2010) present and discuss variations for the B^3 and CEAF metrics which can be used for evaluating end-to-end coreference resolution systems where annotated mentions are not available.

Die Ähnlichkeit soll, hör ich, unverkennbar sein.
(The similarity, I hear, is unmistakable.)

Baron in Hugo von Hofmannsthal's "Der Rosenkavalier"

3 Related Work

This chapter gives an overview of relevant literature and studies directly related to the topics of this thesis. The following list of publications does not claim to provide an exhaustive or complete list of all relevant literature, but contains some of the most relevant works and can be used as a starting point for further investigations. The topics covered in this literature review which deem most relevant to the contents of this thesis are literary coreference annotation, coreference resolution for literary texts and classification of character types, including protagonists. The publications discussed here are ordered chronologically within sections.

3.1. Literary Coreference Annotation

There are several works that include literary texts as a subset in their coreference annotations, but do not specifically alter their annotation guidelines towards this domain or provide detailed analysis on the differences to other domains in their data (cf. Dipper, Lüdeling, and Reznicek 2013; Chamberlain, Poesio, and Kruschwitz 2016) and are therefore not discussed in detail. For Dipper, Lüdeling, and Reznicek (2013), using the same guideline for all domains is a deliberate decision, as it enables better comparability between annotations of different domains. Chamberlain, Poesio, and Kruschwitz (2016) notice a much higher sentence length for their texts from Project Gutenberg, i.e. for literary texts, and a lesser need to edit markables in post-correction.
Following are works that specifically address literary texts as a domain of coreference annotation. Furthermore, some papers present annotations for coreference on literary data

together with experiments on resolving the coreferences. In this case, the annotations are only described in Section 3.2 together with the experiments on CR, instead of additionally being listed separately in Section 3.1.

Finlayson (2017) presents an annotated corpus of Russian folktales called *Propp-Learner* with the goal to test the formalist theory of Vladimir Propp (cf. Propp 1968) regarding re-occurring structures in Russian folktale computationally. Next to other layers of annotations, the corpus contains annotations marking all referring expression as well as co-reference information. Annotations were carried out by in total twelve annotators and underwent an adjudicating process. An inter-annotator agreement (IAA) study yielded an F1-score of 0.91 for the annotation of referring expressions (mentions) and a chance-adjusted Rand score (after Hubert and Arabie 1985) of 0.85.

Rösiger, Schulz, and Reiter (2018) provide guidelines for creating annotation schemes for coreference resolution catered specifically towards the annotation of literary texts. They point out a number of unique aspects of literary texts that influence the annotation of coreference, for example: i) Different levels of narrativity with separate layers of entities (typically between a narrator and the characters of a text, but also nested narrative levels are not uncommon in literary texts), ii) a high rate of switching between generic and specific use of entities, i.e. characters often use generic expressions when referring to a concrete person, but since this connection is only inferential, it strictly is not coreferential, iii) a higher length of the texts makes it difficult for annotators to keep track of all introduced entities, iv) authors may introduce true ambiguity of reference, where coreference is purposefully not supposed to be resolved to serve the narrative. They annotate several German-language literary texts of different types, such as novellas, plays and fairy tales and use a self-developed tool that assigns mentions to groups, i.e. entities, instead of annotating links between markables, as traditionally done in coreference annotation (cf. Müller and Strube 2003).

Bamman, Popat, and Shen (2019) provide a dataset of entity annotations for 100 literary documents from Project Gutenberg. The data is not annotated for coreference or linked markables in general, but rather presents a variation of NE annotation that also includes common noun phrases that denote NEs and nested phrases. The task is therefore only loosely related to coreference annotation, but can be seen as a preliminary step in identifying possible mentions. They face issues in applying categories build for newspaper data to literary texts, including: i) metaphors, ii) personification and iii)

metonymy. The most common type of entity in the data is PER (person), followed by FAC (facility) and LOC (location). This is in contrast to a distribution on news data, where GPE (geo-political entity) and ORG (organization) constitute the second and third most frequent categories.

Bamman, Lewke, and Mansoor (2020) build upon the work in Bamman, Popat, and Shen (2019) and provide actual coreference annotations. Their dataset comprises of 100 literary works in English language, already featured in Bamman, Popat, and Shen (2019). They follow the OntoNotes guidelines for annotating coreference, but include singletons and only annotate mentions that received an NER tag in Bamman, Popat, and Shen (2019). Furthermore, they assume the annotators to already have knowledge of the full text before annotating and thus, identities that are only revealed during the course of the text are already annotated with the knowledge of the full text in mind. They report several statistical properties of the mention and entity distribution in the texts: Many entities only span small portions of the texts while a few entities span almost the entirety of the texts; these entities with a long span contain most of the mentions; a skewness in the ways that these entities with long spans are mentioned, i.e. there are time spans when these entities are not mentioned and spans where they are mentioned a lot; the distance to the closest antecedent is much shorter for pronouns than for common or proper nouns. They identify the entities that span large portions of the texts to be important characters and the other mentions to be minor characters, generic and generally known entities and entities not important for the broader plot or discourse. A neural model based on Lee et al. (2017) performs better on their data than on OntoNotes for gold mentions, with a CoNLL score of 79.3%. The same result can be found for the performance on predicted mentions, for which the score falls by 11.2 percentage points.

Van Cranenburgh and van Noord (2022) present a corpus called *OpenBoek*, which contains nine Dutch novels with a total of 103 000 tokens. The annotations were carried out using the *dutchcoref* system (van Cranenburgh 2019a) and manual corrections in CorefAnnotator (Reiter 2018) by two annotators. Non-referring and time-related mentions were excluded. The authors also performed extensive automatic and manual spelling correction and normalization. Since in literary texts, readers might not be aware of certain coreferences during a first-time reading due to plot-related twists, so like Bamman, Lewke, and Mansoor (2020) they opted to annotate assuming the annotators to be omniscient readers. They found that performing automatic spelling normalization substantially

helps to improve CoNLL scores when using *dutchcoref* on the texts. The authors also perform an evaluation using *dutchcoref* with only its rule-based components and some additional neural-based modules (neural mention identification, gender, animacy and number identification and pronoun resolution) on a single novel, achieving a top CoNLL F1 score of 67.60 when using all three neural modules and manual spelling normalization.

3.2. Literary Coreference Resolution

At the time of writing, there are only a few published works dealing with coreference resolution on literary texts, for the languages English, German and Dutch, described below.

Bamman, Underwood, and Smith (2014) are primarily interested in detecting character types in English narrative texts (described in more detail in Section 3.3 below), but use CR as a means to create more coherent and feature-reach representations of character types. On a dataset of ca. 15 000 English narrative texts, including novels, plays and poetry, they perform proper name as well as pronominal CR. For resolving proper names to entities, they create a list of possible character names as well as possible variations on this name (e.g. *Tom Sawyer*, *Tom*, *Sawyer*, *Mr. Tom Sawyer*, etc.). Using this list, they traverse through the text and assign all mentions of a name or its variants to the respective entity. They note that the proportion of pronominal mentions is relatively high for narrative texts; they give a value of 74% of pronominal mentions for their data. Hence, they also perform CR for pronouns, by using a logistic regression classifier with different features, such as labels from dependency parsing, salience and POS, on a self-annotated corpus of three novels. The model achieves an accuracy of 82.7% in 10-fold cross validation. It is important to note that this value only captures accuracy for pairs of pronouns with potential antecedents. The accuracy of chains emerging from these pairs is not evaluated.

Krug et al. (2015) develop a rule-based system for resolving literary characters in German novels, based on Lee et al. (2011). The rules of their system are similar to the ones in Lee et al. (2011), extended by rules that deal with references in direct speech, detecting nicknames and handling German titles[1]. They evaluate their system on annotated

[1]Note that Krug et al. (2015) claim Lee et al. (2011) to have 7 passes, which they extend to 11; however Lee et al. (2011) present 13 passes, build upon a previous work from Raghunathan et al. (2010), which indeed presents only 7 passes. Apart from the passes dealing with direct speech, nicknames

fragments from 48 German novels of the 19^{th} century. Notably, only references to literary characters are annotated and resolved, i.e. there is no full coreference resolution. Mentions are therefore gathered by applying named entity recognition (NER), only keeping named entities (NEs) that denote a person. In contrast to data based on newspaper articles, the average sentence length of the novels is higher than for articles (24.2 vs. 16.3 tokens on average). The authors also notice a fewer amount of entities that get mentioned frequently, compared to newspaper data, where more entities occur, but are mentioned less often. Furthermore, there is a higher number of pronouns in the novels, given as 70% of all named entities, which poses certain problems for the resolution of references in German, because many German pronouns are ambiguous and may refer to multiple previously introduced entities. They run evaluation on two independent test sets, in order to show the unbiasedness of their system towards the data, and report a MUC F1 score of 85.5 and a B^3 F1 score of 56.0 for the first evaluation and a MUC F1 score of 86.0 and a B^3 F1 score of 55.5 for the second evaluation. They also perform error analysis and note that the majority of mistakes come from semantically complex contexts where knowledge about the real world or the world of the novel would be required, as well as mistakes coming from encapsulated contexts like thoughts or letters that are not modeled by their system.

Van Cranenburgh (2019a) develop a rule-based system for Dutch novels. The annotations were performed by post-correcting the output of the developed system. Like Krug et al. (2015), it is build similar to the system by Lee et al. (2011) and evaluated on fragments of novels, rather than whole texts. In contrast to Krug et al. (2015), van Cranenburgh (2019a) not only consider literary characters to be part of the task, but also objects. However, the task is not full coreference resolution, as events and abstract entities are not included. For mention detection, all noun phrases from parse trees are extracted and filtered for being a person or object. Van Cranenburgh (2019a) report several values on different metrics, including CoNLL and LEA scores, and run their system on their annotated data of Dutch novels, as well as data of shared tasks (Dutch WikiNews articles and Flemish magazines). The system scores much higher on the novels (66.7 CoNLL F1), compared to WikiNews (41.2 CoNLL F1) and magazine (48.4 CoNLL F1) data. They also report a detailed error analysis on a sub-sample of their annotations and present several findings: i) there are no mistakes

and titles, the other passes of Krug et al. (2015) are already occurring (identical or in modified form) in Lee et al. (2011).

in resolving names, ii) the most common mistakes when merging or dividing clusters (i.e. when one cluster in the gold data corresponds to multiple clusters in the system output or vice versa) occur with pronouns iii) half of all errors of assigning wrong spans to mentions come from German phrases that the used parser is not trained to handle.

Poot and van Cranenburgh (2020) compare the rule-based system from van Cranenburgh (2019a) to a end-to-end neural system for Dutch texts. They find that the rule-based system outperforms the end-to-end system for three out of five novels. When comparing novels with newspaper texts, the rule-based system outperforms the end-to-end system on the novels, but gets outperformed for the newspaper texts.

Van Cranenburgh et al. (2021) present a hybrid system with a rule-based component coming from van Cranenburgh (2019a) called *dutchcoref* and additional neural-based modules dealing with mention detection, gender, animacy and number detection and pronoun resolution. In their hybrid system, automatically generated parse trees and BERT embeddings are fed into the mention detection and gender, animacy and number modules, after which the rule-based component suggests possible coreferences and lastly pronouns are resolved using another neural module. They find an improvement of around three percentage points from using all three neural modules over only using the rule-based system on the development set (66.55 vs. 69.37 CoNLL F1 score), however, on the test set, the rule-based component without any neural modules performs almost identical to the setups with neural modules (70.90 vs. 71.00 CoNLL F1 score). They find that the development set has a lower out-of-vocabulary rate than the test set, as well as a lower number of mentions and entities and a higher number of names. The authors assume that these circumstances as well as a different distribution of genre in the two sets are possible explanations for the observed differences in results.

Han et al. (2021) present *FantasyCoref* which mainly contains English translations of the Grimm fairy tales. They adopt an omniscient reader's point-of-view, in which characters that are for instance disguising themselves are still annotated as coreferent to their non-disguised mentions. They also mark entities within prophecies as coreferent with later actual occurrences of this entity. They achieve high inter-annotator agreement scores with CoNLL scores of up to 87.04% and good results with an end-to-end system, leading to CoNLL scores of up to 76.88%.

Schröder, Hatzel, and Biemann (2021) adapt a family of systems by Lee et al. (2017), Lee, He, and Zettlemoyer (2018), and Joshi et al. (2019) to German and test on the DROC corpus (Krug et al. 2015), which contains German novels annotated for coreference. They achieve state-of-the-art results by improving on previous results on up to 30 percentage points. They also utilize incremental learning of coreference clusters (Xia, Sedoc, and Van Durme 2020; Toshniwal et al. 2020), which allows to train and fine-tune on in principle arbitrarily long texts in a neural coreference resolution setup. This is especially useful for literary texts, which are usually much longer than the commonly used newspaper texts.

Schmidt, Krug, and Puppe (2022) present experiments for CR on two types of texts, German historic novels and German fairy tales, with two types of architectures, rule-based and neural. They adapt the rule-based system by Krug et al. (2015) to also handle family relation designations and reflexive pronouns. The neural network system is based on Lee, He, and Zettlemoyer (2018) and trained on both the novels and the fairy tales and tested on both domains as well. They find that the rule-based system is more stable across domains, while the neural network system achieves better results on the domains on which it was trained, but performs significantly worse on cross-domain testing. A neural network system trained on the novels and fine-tuned on the fairy tales performs best.

Dönicke et al. (2022) present *MONAPipe*, which is a pipeline implementation for the popular Python package *spaCy* focused on literary texts. The pipeline contains a coreference resolution component, for which the authors present evaluation results. The component reimplements the rule-based system by Krug et al. (2015), but uses universal dependencies as parse tree inputs. They achieve CoNLL F1 scores between 25.18 to 47.37% on GerDraCor-Coref (Pagel and Reiter 2020), depending on different setups like only using heads of mentions, only using NPs or using gold mentions.

Hicke and Mimno (2024) use Large Language Models (LLMs) to generate coreference markups for plain text sentences. The models not only need to generate the markup but also re-generate the input sentence. All models are fine-tuned on plain-text-input markup-output pairs. The T5 models are able to reliably replicate the input text and achieve high F1 scores (highest score 80.16%) for correctly generating coreference markup, while Pythia-based models neither able to generate any coreference markup nor replicate the input text due to "hallucinated" generated text.

Applied coreference resolution Several works use pre-existing tools for resolving coreferences on literary texts in order to use the acquired information for other means.
Vala et al. (2015) use Stanford's CoreNLP pipeline to detect and group mentions of character names and refine the output by using several heuristic rules. They report F1 scores between 0.4478 and 0.7579 for four groups of 88 texts.
Vala et al. (2016) make use of the Stanford sieve system by Lee et al. (2011) and evaluate the system on several annotated chapters of a single novel. They report an F1 score of 0.542 for the Stanford system, compared to further annotations by non-experts, which yield an F1 score of 0.975 when evaluated against the original expert annotation.
Iyyer et al. (2016) use the tool *BookNLP*, which emerged from Bamman, Underwood, and Smith (2014), for resolving coreferences and use the information to derive trajectories of character relations in literary texts. As they do not have manual annotations of coreference for their data, they do not report evaluation scores for applying the system.

3.3. Automatic Detection of Character Types

Automatic detection of character types is a relatively new task which emerged in CLS. In this thesis, the term is understood as determining the plot-related roles of literary characters and encompasses roles such as *protagonist*, *title character* and more abstract roles like *father*, *daughter*, *schemer*, *messenger* and other character types that have been worked out by, and proved useful in, literary studies. Below is an outline of some of the works dealing with this or similar tasks.

Bamman, O'Connor, and Smith (2013) operate on movie scripts and attempt to learn latent character roles which are only defined by their main actions. They compare these latent character roles created by their system to a manually curated list of 72 stereotypical character roles, including the corrupt corporate executive, the jerk jock or the surfer dude (Bamman, O'Connor, and Smith 2013, p. 356). The latent character roles are created by performing Latent Dirichlet Allocation (LDA, Blei, Ng, and Jordan 2003). The authors also create a second model which incorporates external meta information, like genre and characters' age and gender. When comparing to what extent the automatically generated clusters match the gold clusters, they find that increasing the number of clusters in the automatic system and the number of roles in the gold clusters also increases system performance, which they contribute to the model being able to capture fine-grained character roles with an increase of clusters. However, they find that using external meta

information does not help with performance.

Bamman, Underwood, and Smith (2014) continues the work of Bamman, O'Connor, and Smith (2013) and experiment on ca. 15 000 English novels. They employ a Bayesian-based model to learn latent character roles and compare it to hypotheses made by a literary scholar, which include statements about the similarity of characters between and within different authors and authors' works. They are then able to compare the model's judgements about the similarities of its automatically induced character roles to the judgements made by a literary scholar. They find that a regression model from Bamman, O'Connor, and Smith (2013) performs best for matching the hypotheses about characters being more similar to each other within one author's work as compared to characters from another author, but the Bayesian-based model performs best to distinguish characters within an author's work. The authors notice furthermore that the model learns to automatically predict gendered character roles and that it might be useful to treat more complex, exploratory literary hypotheses which stem more from a certain subjective point of view, as a separate phenomenon. The authors also note that the predicted latent character types are still not very similar to types that literary scholars would usually work with, despite the good performance with regard to the hypotheses, but aligning better with literary genres (Bamman, Underwood, and Smith 2014, p. 377).

Valls-Vargas, Zhu, and Ontañón (2014) work on 10 Russian folktales translated into English. On these texts, they attempt to identify seven character roles introduces by Vladimir Propp (Propp 1968): hero, villain, dispatcher, donor, (magical) helper, sought-for-person[2] and false hero (Valls-Vargas, Zhu, and Ontañón 2014, p. 189). They use an end-to-end system called *Voz* to extract coreference chains from the stories on which character identification is performed. The role identification is performed by creating character action matrices, which contain all actions (in the form of action verbs) performed by characters and performed on other characters. These matrices are then compared via a similarity calculation to manually created role action matrices that represent what the authors assume which actions prototypically to belong to a certain role. The system achieves its highest performance when using gold coreference annotations and gold action verb extraction and compared to a role action matrix that only contains the hero and villain role with a match of 44.12%. The lowest result comes from a setup

[2]The *sought-for-person* corresponds to the term *princess* used by Jahan, Mittal, and Finlayson (2021), which is described further down in this text. Both terms are used interchangeably in Propp (1968, p. 79).

where the system receives no gold information and is compared to a role action matrix that contains action verbs coming from Propp's narrative functions, resulting in a match of 11.54%, 2.75 percentage points below a random baseline. A clear advantage of the approach is that it can be carried out in an unsupervised fashion, since only the general role action matrices need to be created against which to evaluate.

Jannidis et al. (2016) take curated summaries for 58 German novels and extract all mentioned characters from them, assuming them to be the most central characters of a novel. From the novels, they extract the number of occurrences in coreference chains for each character, the number of how often characters are involved in direct speech and lastly the weighted degree one two types of networks, one constructed from characters appearing together in a span of text and the other constructed from characters occurring together in directed speech as either the speaker or the addressee. They rank the characters mentioned in the summaries by number of mentions and occurrence and the characters extracted from the novels by their feature values and check if any of the top five or top ten characters from a novel occurs in the top ten rankings of the summaries. The highest score is achieved with the top ten characters of the text-span network feature compared to the summary ranking which is based on the count of mentions, with a matching percentage of 51.6%. The top five characters from the direct speech count feature perform worst when compared to the occurrence-based summary ranking with a percentage of 37.5%. In general, comparing to the occurrence-based summary ranking leads to lower scores than comparing to the mention-count-based summary ranking. When allowing the matching to be with any of the summary characters, the highest score is 64.7% for the text span network and the count of occurrences in coreference chains features. In this setup, comparing to the occurrence-based summary ranking now yields higher scores than comparing to the count-based summary ranking.

Skowron et al. (2016) take a corpus of 212 action movie scripts and annotate characters appearing in it with different roles: hero, antagonist, spouse/partner/lover, sidekick, supporting character, mentor, power in the background and law representatives. They deploy a multitude of features, outlined below. One annotator annotated linguistic features, categorized into expressivity of characters, such as sentiment and use of interjections and social-relational, such as dialogue act, addressing dialogue partners, non-standard English use, etc. Another set of features comes from training a skip-thought model which provides sentence vector representations. Additionally, the authors created co-presence

networks and computed betweenness centrality, closeness centrality, clustering coefficient, squared clustering coefficient, in- and out-degree, and a binary feature for the character with the highest betweenness centrality in a script. They give all these features into a Support Vector Machine (SVM) and train it for detecting the different annotated roles. They find the features based on the literal content of utterances and on the networks to overall work best. An SVM trained with these features resulted in an F1-score of 0.77 for detecting heroes, 0.42 for detecting antagonists, 0.45 for detecting characters supporting the hero and an overall F1-score of 0.43. However, this model was not able to detect many roles at all, in particular the mentor role, characters supporting the antagonist, characters representing a power in the background and lovers of the antagonist. These roles were also never classified correctly by any of the other models using different features.

Algee-Hewitt (2017) takes a corpus of 3568 plays and constructs co-presence networks on these plays, as well as calculating betweenness centrality and eigenvector centrality for all networks. He takes the upper quantile of a distribution of Gini coefficients divided by the eigenvector centrality scores to be a cut-off point for protagonism, i.e. any character in a play who is within this top quantile will be a protagonist, and plays with a higher score will have a single dominating protagonist, while plays with a low score will have multiple equal protagonists. Algee-Hewitt then goes on to compare different time periods of his corpus with regards to how the plays in them distribute protagonism and observes a high amount of plays with a few strong protagonists in the sixteenth and seventeenth century and a trend towards multiple equal protagonists or groups of characters with their own protagonists when approaching the eighteenth and nineteenth century.

Fischer et al. (2018) compare different feature groups in terms of their ability to capture dominance relationships in German plays. They calculate count-based features, in particular number of scenes, utterances and tokens per character, and network-based features, in particular degree, weighted degree, betweenness centrality, closeness centrality and eigenvector centrality, and show that the top quantile of count-based and network-based distributions gives a similar distribution; however, the network-based features allocate more characters into the top percentile than the count-based features. They propose that such a multi-dimensional approach is needed in order to capture all possible dominance relationships within plays, i.e. possible protagonists, as count-based and network-based features capture different aspects of character relations.

Jahan, Mittal, and Finlayson (2021) build upon previous work introduced by Jahan, Chauhan, and Finlayson (2017), Jahan and Finlayson (2019), and Jahan et al. (2020), which dealt with detecting animacy of mentions and linking it to the coreference chains it belongs to in the former case and with a general detection of characters by assigning them to coreference chains in the two latter cases. They use a catalogue of the seven roles developed by Propp, containing the roles hero, villain, helper, donor, princess, false hero and dispatcher. As data, they use the extended ProppLearner corpus by Jahan et al. (2020) containing coreference chains with attached information about animacy and characterhood. On these coreference chains, they perform k-means clustering in order to group character chains into groups of similar roles, using a series of features such as TF-IDF, sparse vector representations of tokens occurring in the coreference chains and different mappings of characters to pre-annotated plot functions. Using different number of clusters, they found a cluster size of seven to be optimal, which corresponds to the seven Proppian character roles. They found a feature which encodes if there is a match between the string of a coreference chain and a sentence containing a certain plot function to be the overall best performing feature, with an F1-score of 0.58.

3.4. Limitations of the Related Work

The presented related work shows some limitations and gaps that this thesis aims to fill out.

The most apparent limitation in terms of literary coreference annotation and literary CR is that none of the presented approaches tackle the text type of dramatic texts. However, it can be argued that dramatic texts are especially interesting in terms of coreference resolution as the indication of speaker tags offers text-given informations that can be utilized by CR systems (see Chap. 5). Furthermore, dramatic texts feature mentions of plot-relevant objects that can span the entire texts and are crucial to track and resolve in order to properly analyze the story (see Chap. 4).

Secondly, in the realm of character type detection, while there is work on dramatic texts (Algee-Hewitt 2017; Fischer et al. 2018) and the arguably related text type of movie scripts (Bamman, O'Connor, and Smith 2013; Skowron et al. 2016) in terms of protagonist detection, more complex character types like schemers or tasks related to protagonist detection like title character detection are not covered by the existing research literature (see Chap. 6).

Thirdly, no other research has tried to utilize coreference information to improve character

type detection. Yet, it stands to reason that this type of information would be highly useful for character type detection, as for instance the number of times a character is mentioned or what other characters say about a character is vital information for classifying the type of a character (see Chap. 7).

Siehst du? Siehst du mehr, ob das kein Luderleben ist? und dabei bleibt man frisch und stark, und das Korpus ist noch beisammen, und schwillt dir stündlich wie ein Prälatsbauch [...].
(Do you see? Do you see, is this not a wanton life? and at the same time you remain fresh and strong, and the corpus is still together, and swells hourly like a prelate's belly [...].)

Spiegelberg in Friedrich Schiller's "Die Räuber"

4

GerDraCor-Coref: A Corpus of Coreference Annotations for German Theatre Plays

This chapter presents GerDraCor-Coref (Pagel 2022b), a corpus of selected German theatre plays annotated with coreference information. Theatre plays are an interesting resource to research coreferences on since they are in dialogical form, quite long texts with a complex structure (usually acts, scenes, utterances and stage directions) and stylistically and linguistically complex. These are all properties usually not found in the typically researched types of texts. In addition to an introduction of the annotation tool used (Section 4.1), a brief overview of the annotation guidelines (Section 4.2), an Inter-Annotator Agreement study (Section 4.3) and a statistical analysis comparing GerDraCor-Coref's coreference annotations with other German corpora (Section 4.4), the chapter also explores long coreference chains as well as coreference chains which contain long gaps in between their mentions (Section 4.5). This is of great interest since the plays exceed the length of documents usually annotated for coreference and coreference chains that span large portions or the entire text are often relevant for literary interpretations (e.g. main characters or important objects). In the end, the annotated coreferences of the corpus are used to explore the relationship of characters of selected plays as well as frequently mentioned entities (Section 4.6).
Some of the information presented in this chapter has already been published in Pagel

and Reiter (2020) and Pagel and Reiter (2021).

4.1. Annotation Tool

CorefAnnotator (Reiter 2018; Reiter, Kiss, and van Cranenburgh 2022) has been used for all annotations in GerDraCor-Coref.[1] CorefAnnotator has been chosen since it focuses specifically on the annotation of coreferences and has been successfully used in other coreference annotation efforts (van Cranenburgh 2019a; Han et al. 2021) The tool allows to work on a text and on the entity structure at the same time in a split view (Fig. 4.1). While the text is displayed in a window on the left side, which allows to read and annotate the text directly, all annotated entities are displayed in a window on the right-hand side. Here, entities can be manipulated by removing entities and mentions, setting flags for entities and mentions or renaming entities, among other things. The entity view also allows annotators to easily keep track of all entities and mentions already annotated. Entities are named automatically after the string of their first annotated mentions, but can be renamed by the annotator. Entities are colour-coded which also determines the colour of the underline of mentions in the text view. The colours can also be changed for each entity individually. Entities can be grouped if a mention refers to two existing entities at the same time. The resulting group entity keeps track of the individual entities it is composed of (see also Figure 4.8). CorefAnnotator allows to export annotations into different formats, most notably into (compressed) XMI, CoNLL format and a custom CSV format containing offsets, mentions and flags.

4.2. Annotated Phenomena

Coreference

The major contribution of the corpus is the annotation of coreferences. In general, all referring noun phrases were considered to be mentions. Also, all dramatic text levels described by Pfister (1988, see Section 2.2.1) have been annotated, namely the dramatis personæ, speaker tags, stage directions and utterances.
Figure 4.2 shows an example for all coreferent and potentially referring mentions in a snippet from TextGrid Repository (2012b): *Hofmannsthal, Hugo von. Der Rosenkavalier.*

[1] There were early annotations put directly into the TEI of GerDraCor and annotations using the tool WebAnno (Eckart de Castilho et al. 2016). All of these annotations have been transferred to CorefAnnotator.

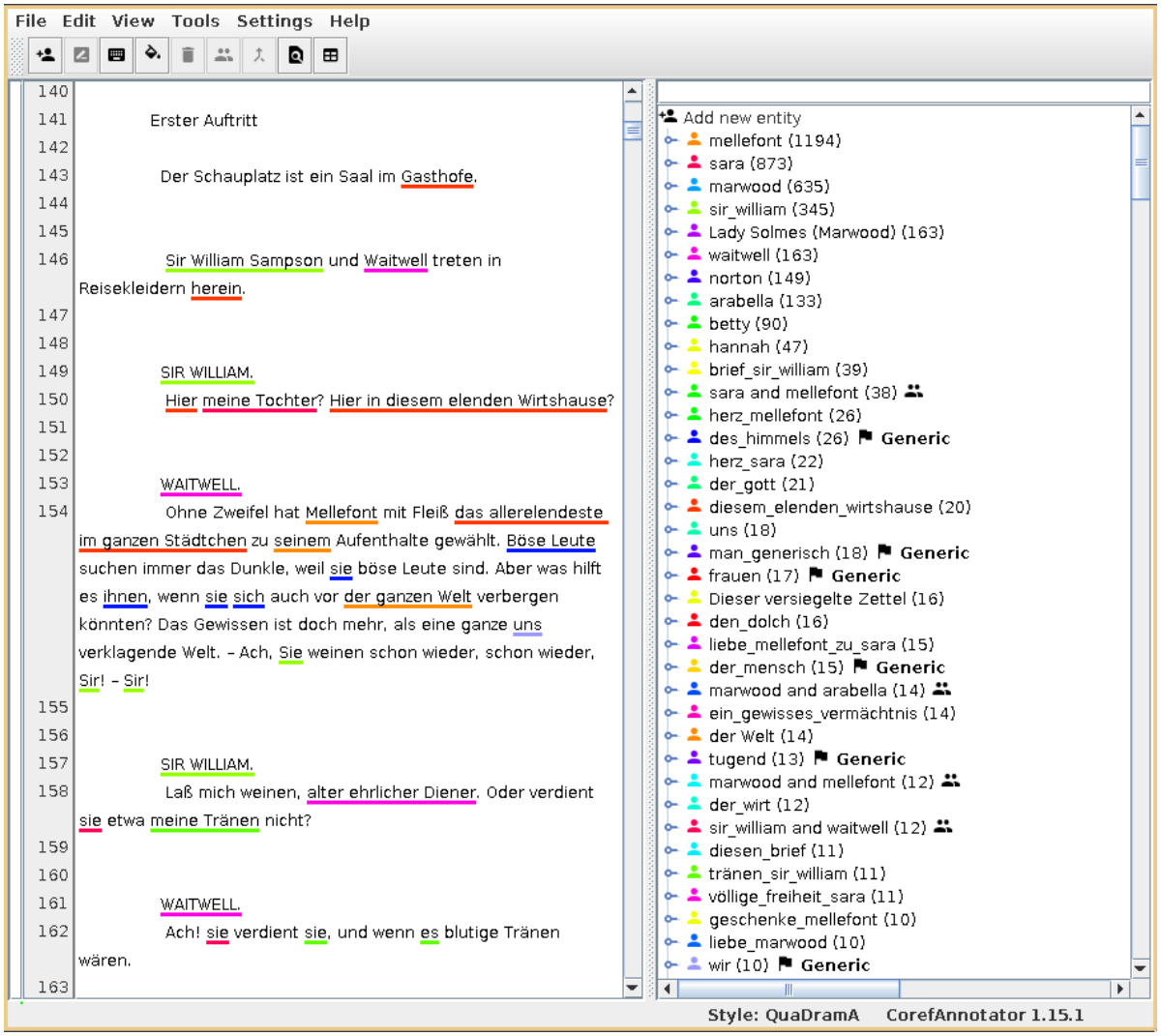


Figure 4.1.: Screenshot of CorefAnnotator version 1.15.1 with a snippet of an annotation of Lessing's *Miß Sara Sampson*. The left window shows an excerpt of the text with underlined mentions, the right window the beginning of a list of entities. Flags are marked in bold next to an entity's label. The numbers in brackets after an entity's label are the number of mentions of that entity.

Noun phrases which are referring are surrounded by square brackets and mentions which are coreferent are marked by a subscript. In this example, referring noun phrases are definite (*ich*, *du*, *dein*, *dir*, *er*, *der*) and indefinite (*es*, *was*, *was anders*, *s*) pronouns, complex noun phrases (*ein Lärm im Hof*, *dein Mann*), nested noun phrases (*Hof* inside the larger noun phrase *ein Lärm im Hof*, *dein* inside the phrase *dein Mann*), named entities (*Raitzenland*, *Esseg*) and conjunctions (both constituents *Pferd'* and *Leut'* as well as the whole conjunction *Pferd' und Leut'* are referring).

1 [Octavian]$_1$.
2 [Der Feldmarschall]$_2$?
3
4 [Marschallin]$_3$.
5 Es war [ein Lärm] im [Hof]$_4$ von [[Pferd'] und [Leut']] und [er]$_2$ war da.
6 Vor [Schreck] war [ich]$_3$ auf einmal wach, nein schau nur,
7 schau nur, wie kindisch [ich]$_3$ bin: [ich]$_3$ hör noch immer [den Rumor im [Hof]$_4$]$_5$.
8 [Ich]$_3$ bring[s]$_5$ nicht aus [dem Ohr]. Hörst [du]$_1$ leicht auch [was]?
9
10 [Octavian]$_1$.
11 Ja, freilich hör [ich]$_1$ [was]$_6$, aber muß [es]$_6$ denn [[dein]$_3$ Mann]$_2$ sein!
12 Denk [dir]$_3$ doch, wo [der]$_2$ ist: im [Raitzenland]$_7$, noch hinterwärts von [Esseg].
13
14 [Marschallin]$_3$.
15 Ist [das]$_7$ sicher sehr weit?
16 Na dann wird[s]$_5$ halt [was anders] sein. Dann is ja gut.

Figure 4.2.: Snippet from TextGrid Repository (2012b): *Hofmannsthal, Hugo von. Der Rosenkavalier*, extended with markup showing coreference relations. For a translation, see Figure C.3

Expletives were not annotated, since they do not refer and only coreferences were annotated. Hence, every third person pronoun which is not annotated could automatically be labeled as expletive if this information is needed in later studies. In Figure 4.2, the pronoun *Es* (in English *There*) in line 5 would be an example for such a non-referring expletive and is accordingly not surrounded in brackets. Bridging anaphors were also not considered for annotation (see Section 2.3.1).
Since a large amount of references in dramatic texts are references to literary characters,

an obvious possibility is to perform entity linking instead of full coreference resolution. However, CR allows to annotate and resolve references to entities that are not characters, which provides the possibility to analyze and utilize information of entities beyond characters, such as feelings of characters or items in the possession of characters. Studies that only use annotated references to characters such as Krug et al. (2015) and Andresen and Vauth (2018a) are therefore closer to entity linking than the annotations and experiments presented within this thesis.

Non-nominal antecedents

The corpus also includes annotations of non-nominal antecedents (see Section 2.3.1). In general, all mentions were considered to refer to non-nominal antecedents if the reference could not be resolved to a noun phrase, but only to a verbal or clausal phrase. Figure 4.3 shows examples for an antecedent that is a verbal phrase (*in die Hand schreiben*, eng. *writing sth. on the hand*) and for clausal antecedents (*wo ist Philipp*, eng. *where is Philipp*; *Es hilft*, eng. *It helps*). Note that for verbal phrases, the exact antecedent is often speculative. In this example, one could make the case that Sylvius is referring to the fact that specifically Agnes is writing into his hand, in which case the whole clause would be annotated (i.e. *ich schreib's dir in die Hand*, eng. *I'll write it on your hand*), rather than just the verbal phrase, which is referring to the statement that writing something into the hand helps in general (with remembering things). In such cases, the shorter phrase is annotated if equally plausible. Each non-nominal mention has been marked with the special flag *Non-Nominal* in CorefAnnotator.

```
1  Sylvius.
2  Agnes, [wo ist Philipp]1?
3
4  Agnes.
5  Du lieber Gott, ich sag['s]1 dir alle Tage,
6  Und schrieb['s]1 dir auf ein Blatt, wärst du nicht blind.
7  Komm her, ich [schreib['s]1 dir in die Hand]2.
8
9  Sylvius.
10 Hilft [das]2?
11
12 Agnes.
13 [[Es]2 hilft]3, glaub mir['s]3.
14
```

15 Sylvius.
16 Ach, [es]$_2$ hilft nicht.
17
18 Agnes.
19 Ich meine,
20 Vor dem Vergessen.

Figure 4.3.: Snippet from TextGrid Repository (2012c): *Kleist, Heinrich von. Die Familie Schroffenstein*, extended with markup only showing coreference chains which contain a non-nominal mention. For a translation, see Figure C.4

Generics

Noun phrases that refer to a general class rather than an individual instance of a class (see Section 2.3.1) are marked with the flag *Generic* in CorefAnnotator. The flag is applied on the entity level, since all mentions of a generic entity have to also be generic themselves. Figure 4.4 gives an example of a generic entity and its mentions (*Der überlegene Mann*, eng. *The superior man*). Note that the subsequent uses of the masculine singular pronouns refer to a concrete person (a character from the play called August), and not a generic notion of a "superior man" that Albertine referred to before.

1 Albertine.
2 Mutter! haben Sie in seiner Miene nichts von dem feinen Spott bemerkt,
mit dem [der überlegene Mann]$_1$ die Versuche des Weibes – selbst
wenn [er]$_1$ sie nicht ganz verdammen kann, so gerne lächerlich macht?
3
4 Madame Wölbing.
5 Nein, die Anerkennung deines Talents kam aus den[*sic!*] Herzen.
6
7 Albertine.
8 *schnell.*
9 Ich fürchte nicht seinen Tadel, nur seinen Spott! Er soll streng, aber er soll
redlich seyn. Seine Rüge soll mich belehren, aber sein Witz soll mich
nicht er bittern[*sic!*]. Können Sie das von ihm erwarten, so bringen
sie[*sic!*] ihn.

Figure 4.4.: Snippet from TextGrid Repository (2012h): *Weißenthurn, Johanna von. Das Manuscript*, extended with markup showing a coreference chain representing a generic entity. Other coreferences have not been marked. Unexpected spellings that occurred in the source have been marked with [*sic!*] (likely, the intended spellings are *dem*, *erbittern* and *Sie*). For a translation, see Figure C.5.

Predicates

Mentions are always marked with the flag *Predicate* (see Section 2.3.1) if the mention is in a predicate position, but not all predicates co-refer and thus not every predicate is part of an entity. While it is not often the case that a predicate is referred to later on, this usually happens if the predicate describes a property by using a generic entity. Figure 4.5 shows a snippet containing the generic mention *eine Verbrecherin* (eng. *a (female) criminal*), with which Sara describes herself in a predicate construction and in the next sentence uses the same generic expression again, once again in a predicate construction.

1 [Sara]$_1$.
2
3 [...]
4
5 So soll [ich]$_1$ [mein]$_1$ Vaterland als [eine Verbecherin]$_{2,generic,predicate}$ verlassen? Und als [eine solche]$_{2,generic,predicate}$, glauben Sie, würde [ich]$_1$ Mut genug haben, [mich]$_1$ der See zu vertrauen?

Figure 4.5.: Snippet from TextGrid Repository (2012e): *Lessing, Gotthold Ephraim. Miß Sara Sampson*, extended with markup showing coreference and a predicate which is part of a coreference chain. For a translation, see Figure C.6.

Predicates that are not part of an entity are not marked. In Figure 4.6, the first occurrence of *böse Leute* (eng. *evil people*) is coreferent with the 3rd person plural pronoun *sie* (eng. *they*), while the second occurrence of *böse Leute* is in a predicate position (*sie sind böse Leute*, eng. *they are evil people*) and not coreferent with the first occurrence.

1 Waitwell.
2
3 [...]
4

5 [Böse Leute]$_1$ suchen immer das Dunkle, weil [sie]$_1$ [böse Leute]$_{predicate}$ sind.
Aber was hilft es [ihnen]$_1$, wenn [sie]$_1$ [sich]$_1$ auch vor der ganzen Welt verbergen könnten?

Figure 4.6.: Snippet from TextGrid Repository (2012e): *Lessing, Gotthold Ephraim. Miß Sara Sampson*, extended with markup showing coreference and a predicate which is not coreferent. For a translation, see Figure C.7.

Grouped entities

Plural mentions that were previously referenced via individual, singular antecedents pose specific challenges towards coreference annotation and CR (Eschenbach et al. 1989; Kamp and Reyle 1993). If a plural is used to refer to two mentions that also exist as separate entities, it is not correct to mark the plural expression coreferent with each mention individually, as this would imply that the two existing entities are coreferent. Take for example the situation in Figure 4.7. Two entities, denoting the characters *Saladin* and *Sittah*, are referred to by the plural pronoun *wir* (eng. *we*). It would not be correct to make the mentions of *wir* coreferent with the mentions of Saladin and Sittah, as this would then encode that the mentions of Saladin and Sittah also refer to the same entity, which they do not. At the same time, there is an obviously strong relationship between the plural entity and the entities denoting the two characters.

1 [Saladin]$_1$.
2 Ei sieh! so hättest [du]$_2$ ja wohl, wenn [du]$_2$
3 Verlorst, mit Fleiß verloren, [Schwesterchen]$_2$?
4
5 [Sittah]$_2$.
6 Zum wenigsten kann gar wohl sein, daß [deine]$_1$
7 Freigiebigkeit, [[mein]$_2$ liebes Brüderchen]$_1$,
8 Schuld ist, daß [ich]$_2$ nicht besser spielen lernen.
9
10 [Saladin]$_1$.
11 [Wir]$_3$ kommen ab vom Spiele. Mach ein Ende!
12
13 [...]
14
15 [Sittah]$_2$.
16 Ach so

17 Willst [du]$_1$ den Stachel des Verlusts nur stumpfen.
18 Genug, [du]$_1$ warst zerstreut; und mehr als [ich]$_2$.
19
20 [Saladin]$_1$.
21 Als [du]$_2$? Was hätte [dich]$_2$ zerstreuet?
22
23 [Sittah]$_2$.
24 [Deine]$_1$
25 Zerstreuung freilich nicht! – [O Saladin]$_1$,
26 Wenn werden [wir]$_3$ so fleißig wieder spielen!

Figure 4.7.: Snippet from TextGrid Repository (2012f): *Lessing, Gotthold Ephraim. Nathan der Weise*, extended with markup showing coreference relations for *Saladin* (index 1), *Sittah* (index 2) and the plural referring to both (index 3). For a translation, see Figure C.8.

In CorefAnnotator, this issue is solved by introducing a grouped entity, containing the plural mentions and the information that this entity refers to two or more singular entities. Figure 4.8 shows how the three entities of Saladin, Sittah and the plurals are encoded in CorefAnnotator.

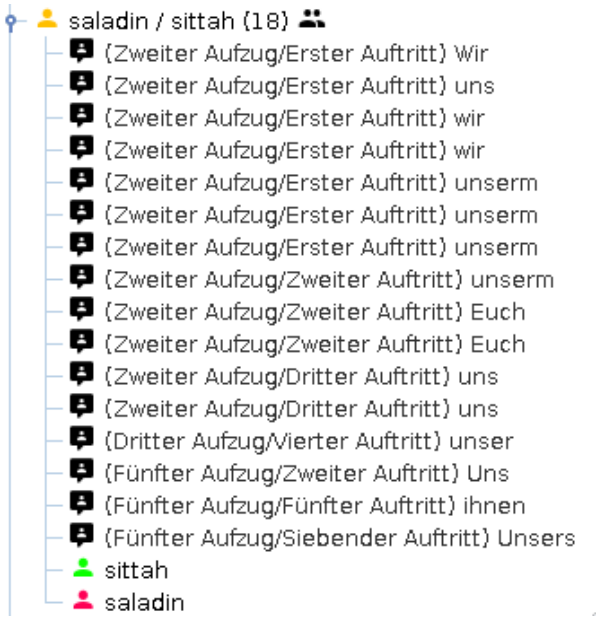


Figure 4.8.: Screenshot from CorefAnnotator, showing the group entity containing the plural mentions as well as references to the stand-alone entities of Saladin and Sittah.

Sometimes, it is not clear which entities are included in a plural expression, or a plural expression refers to a variable group of people. For instance, in Schiller's *Die Räuber*, it is not clear which people are part of the group of bandits, as this changes throughout the course of the play and many members are not mentioned by name, while others are. In this case, the entity denoting the group of bandits is not treated as a group entity, but simply as its own entity, even though singular entities in the play are technically part of this group.

4.3. Inter-Annotator Agreement

In order to assess the quality of the annotations, an IAA study is performed. The better the score of agreement, the more the annotations, the annotation task and the annotation guidelines can be said to be inter-subjective and consistent rather than random (cf. Pustejovsky and Stubbs 2012, p. 126; Kübler and Zinsmeister 2015, chap. 2). To this end, multiple acts have been annotated multiple times by different annotators to allow for comparisons. In total, four plays with twelve acts were annotated by at least two annotators.[2]

Popular methods for comparing agreements of annotations are Cohen's κ (Cohen 1960) or Fleiss' κ (Fleiss 1971). However, it is not straightforward to apply these measures for computing the agreement of coreference annotations, since these measures operate on binary comparisons, while coreference chains can be thought of as sets that need to be compared as a whole and item by item (see also Artstein and Poesio 2008). Therefore, one way to compute the agreement of multiple coreference annotations is to use the established metrics for evaluating the output of a system to gold annotations, such as the MUC score[3]. This was carried out by creating pairs of annotations and declaring one annotation to be the "gold" annotation and the other the "system" annotation and measuring the respective metric. The average of the resulting values gives the IAA score[4]. In order to allow for a wide range of possible comparisons with other works, agreement will be reported using the MUC, BLANC, CEAF_e, CEAF_m, B^3, LEA and CoNLL scores, calculated using the reference scorer of Pradhan et al. (2014) and Pradhan,

[2] Table A.1 gives an overview of the plays and acts that were annotated in parallel by multiple annotators and by which annotator.

[3] See also Artstein and Poesio (2008) for a survey of other ways to compute agreement for anaphoric annotations and especially the therein mentioned Passonneau (2004).

[4] While the F1 score of all metrics is symmetric, i.e. the same value irregardless of which annotation is set as gold or prediction, the precision and recall values are always reversed between gold and prediction annotation.

Luo, and Recasens (2016). Finlayson (2017) reports the chance-adjusted Rand index for quantifying the IAA of the coreference annotations in his corpus; however, the Rand index can only be applied if the mentions of the gold annotations are identical to predicted output and since annotators could choose mention spans freely, the Rand index cannot be reported here[5].
Table 4.1 shows the scores as mean values, as well as the standard deviation.

	Metric	Mean	SD
Mention Span		0.61	0.31
Coreference	B^3	0.49	0.32
	BLANC	0.48	0.36
	$CEAF_e$	0.32	0.22
	$CEAF_m$	0.54	0.31
	CoNLL	0.47	0.29
	LEA	0.47	0.32
	MUC	0.60	0.34
	Mean	0.48	0.31

Table 4.1.: Inter-Annotator agreement on mention spans and coreference. All scores are F1 scores.

It can be seen that the average over all metrics for agreement on coreference lies at 0.48. The average SD is relatively high (0.31), which suggests that some plays have much higher agreement and some much lower. This might indicate that some plays are more difficult and some more easy to annotate than others. Furthermore, the agreement on mentions is relatively low with 0.61 F-Score, which serves as an upper boundary for coreference agreement. The values reported here are lower than the ones reported in Versley (2006), who report a MUC value of 83.0, the ones in Krug et al. (2018) with a MUC score of 88.5 and a B^3 score of 69.0, the ones in Han et al. (2021) who report a CoNLL score of 87.04% and the ones in Finlayson (2017), who presents a Chance-Adjusted Rand index of 0.85, which is comparable to the BLANC score. Chamberlain, Poesio, and Kruschwitz (2016) report an average κ score of 0.90, however, this is not directly comparable to the scores reported in Table 4.1. This might suggest that dramatic texts are more difficult to annotate for coreference than other types of text, although the sample sizes are too small

[5]Note however that the BLANC metric is a variant of the Rand index which, in a modification of Luo et al. (2014), also applies when mentions in the responses are not identical.

to know with certainty (see Section 4.4 for a comparison of the size of GerDraCor-Coref with other corpora). Furthermore, different annotation guidelines were used in all of these studies.
Since annotators could freely choose the spans for the non-nominal antecedents, it is also interesting to look at the overlap of that. In order to achieve this, two strains of IAA studies are performed. Firstly, the agreement is measured if two annotators chose the exact same text span for an annotation labeled as *non-nominal*. Secondly, the condition is relaxed and two annotations count as agreements if at least the beginning or the end of a text span of two annotations match, or if one text span is included within the other. The former case will be called strict non-nominal IAA and the latter relaxed non-nominal IAA. Table 4.2 shows the results for both the strict and the relaxed setup. Shown are the percentages of matches between two annotations, per play as well as combining the annotations of all plays (*Total*).

Setup	Play	Percentage
strict	Der sterbende Cato	5.21
	Emilia Galotti	37.70
	Die Räuber	13.28
	Miß Sara Sampson	5.17
	Total	9.67
relaxed	Der sterbende Cato	12.50
	Emilia Galotti	52.46
	Die Räuber	28.85
	Miß Sara Sampson	8.62
	Total	18.66

Table 4.2.: Inter-Annotator agreement on choosing the span of a non-nominal antecedent.

It can be seen that overall, the agreement is quite low, with only 9.67% overlap for the strict setup and 18.66% overlap for the relaxed setup. However, the percentage depends heavily on the annotated play, as for the play *Emilia Galotti* in the relaxed setup, more than half of the time, the spans overlap at least partially and for 37.70% of the cases there is an exact match. This would let to conclude that for some plays it is easier to agree on the non-nominal antecedents, however, it is not clear what kind of linguistic properties *Emilia Galotti* has compared to the other plays that would explain this.

4.4. Statistical Analysis

Pagel and Reiter (2020) presented statistical analyses of an older version of GerDraCor-Coref with 31 annotated texts. The following analyses are an update of the analyses presented in Pagel and Reiter (2020) and are based on the current version of GerDraCor-Coref (v. 1.5.0) which encompasses a total of 47 annotated texts, which can be further divided into 84 acts[6] and 542 scenes. For a full list of these plays, see Appendix B, Table B.1.

		Num. of Documents	Unit	Total Count	Mean	SD
GerDraCor-Coref						
Drama			tokens	542 421	6457.39	2207.45
	Act	84	mentions	119 812	1426.33	467.70
			entities	14 369	171.06	78.46
			tokens	476 686	879.49	886.74
	Scene	542	mentions	103 184	190.38	181.19
			entities	18 279	33.73	28.45
TüBa-D/Z						
Newspaper			tokens	1 565 620	467.35	478.22
		3350	mentions	144 785	43.22	48.55
			entities	39 682	11.85	11.91
DIRNDL						
Radio News			tokens	38 634	702.44	212.68
		55	mentions	2832	51.49	21.00
			entities	1178	21.42	8.91
GRAIN						
Radio Interviews			tokens	42 324	1840.17	153.45
		23	mentions	6832	297.04	40.63
			entities	1771	77.00	8.29

Table 4.3.: Overall count of documents, tokens, mentions and entities in the German-language corpora GerDraCor-Coref (acts and scenes), TüBa-D/Z, DIRNDL and GRAIN, as well as mean values and standard deviation (SD) for tokens, mentions and entities.

In Table 4.3, we can see that GerDraCor-Coref comprises of roughly 542 000 tokens, 120 000 mentions and 14 000 entities when looking at all acts and 470 000 tokens, 100 000 mentions and 18 000 entities when looking at the scenes. These numbers include all annotated levels of the plays, i.e. dramatis personæ, utterances, stage directions and speaker tags. The numbers for splitting the acts into scenes are different, since there are acts which are not further subdivided into scenes. These acts are not counted for

[6]The number of acts is only about twice the number of plays since only 10 plays have been fully annotated (with most of them containing five acts), while for the remaining texts, only a single act has been chosen randomly for annotation.

the scene-based statistic. The number of entities is higher for scenes compared to acts, because entities that would usually be counted as one across an entire act are split up for the scene-based setup and counted separately. Looking at the mean values shows large differences between acts and scenes, since the length of acts and scenes differs greatly. While acts are 6500 tokens long on average, the average scene is 880 tokens long. Comparing the standard deviations (SD) gives more insight, since it shows that there are scenes that differ widely from the mean (SD of 886.74 with a mean of 879.49), while acts distribute closer around their mean length (SD of 2207.45 with a mean of 6457.39). The picture is similar for mentions and entities, with scenes having a more unequal distribution regarding their number of mentions and entities than acts. For comparison, the table also shows the corresponding values for the corpora TüBa-D/Z, DIRNDL and GRAIN, which are German corpora containing coreference annotations (see Section 2.3.2). While TüBa-D/Z is much larger than GerDraCor-Coref (40 times more documents and 3 times the number of tokens), the number of mentions is actually comparable, demonstrating the density of mentions in GerDraCor-Coref (see also Table 4.4).
Since the mean values are of limited value when comparing acts with scenes due to the extrem difference in length, we can also look at different ratios in order to get a better understanding of how token, mention and entity numbers relate to each other. One metric is the ratio of mentions to tokens, MTR, which is the number of mentions divided by the number of tokens:

$$MTR = \frac{Number\ of\ mentions}{Number\ of\ tokens} \tag{4.1}$$

This measure normalizes the lengths of the texts and allows for a direct comparison between acts and scenes in terms of their number of mentions.
Another possible ratio is the ETR, the entity-token-ratio:

$$ETR = \frac{Number\ of\ entities}{Number\ of\ tokens} \tag{4.2}$$

In parallel to MTR, this ratio normalizes the length of the texts and allows for a direct comparison of the number of entities.
EMR is the ratio of entities to mentions:

$$EMR = \frac{Number\ of\ entities}{Number\ of\ mentions} \tag{4.3}$$

This value shows how many entities there are compared to the number of mentions and

is a measure of the density of entities in a text.
Lastly, we look at the ratio of mentions to entities, MER:

$$MER = \frac{Number\ of\ mentions}{Number\ of\ entities} \tag{4.4}$$

MER captures how many mentions are there in an entity on average.
The values for each ratio for acts and scenes as well as for the corpora TüBa-D/Z, DIRNDL and GRAIN are shown in Tab. 4.4. When normalizing over the number of tokens (MTR and ETR), acts and scenes contain roughly the same number of mentions on average and scenes contain slightly more entities than acts do. Furthermore, the values of mentions per entity (MER) is higher for GerDraCor-Coref than for the other corpora. Looking at the other corpora, it becomes also clear that GerDraCor-Coref has the largest density of mentions, both when looking at acts and scenes (MTR). The number of entities is relatively equally distributed across the corpora (ETR). Lastly, GerDraCor-Coref has the lowest ratio of entities to mentions (EMR), suggesting that there are a lot more mentions than entities in GerDraCor-Coref (to which the high MTR value already pointed).

Corpus		MTR	ETR	EMR	MER
GerDraCor-Coref	Act	0.2209	0.0265	0.1199	8.3382
	Scene	0.2165	0.0383	0.1771	5.6449
TüBa-D/Z		0.0925	0.0253	0.2741	3.6486
DIRNDL		0.0733	0.0305	0.4160	2.4041
GRAIN		0.1614	0.0418	0.2592	3.8577

Table 4.4.: The different ratios, MTR, ETR and EMR, on acts and scenes, as well as on other corpora.

Looking at the number of times a certain flag was annotated in the corpus, Table 4.5 shows that only around 3.5% of mentions are flagged as *generic*, 0.8% as *non-nominal* and 1.5% as *predicate*.
There are not many studies to compare these values with.
Andresen et al. (2018) report numbers on predicates found by two parsers in the German novels *Corpus Delicti* by Juli Zeh and *Aus guter Familie* by Gabriele Reuter. They let two annotators post-correct the parsers' output and only 80% of the predicates predicted by either parser to be correct. Given this, they arrive at 192 correct predicates for the one and 121 correct predicates for the other parser. Andresen et al. (2018) do not report

how many mentions are annotated for these two novels, however, Andresen and Vauth (2018b), mentioned in Andresen and Vauth (2018a), contains 6336 mentions annotated for *Corpus Delicti*. From this, we can calculate that either 3% or 1.9% of all mentions are in a predicate construction, depending on the parser, which is similar to the 1.42% in GerDraCor-Coref. However, the numbers are not directly comparable, as Andresen et al. (2018) can only report on found predicates, hence the actual maximal recall is unknown, and they only annotated characters for their coreference annotations.
Chen, Su, and Tan (2010) report that 19.97% of 6187 anaphora in OntoNotes 2.0 are non-nominal, which is quite a bit higher than the 0.8% in GerDraCor-Coref; however, this number only reflects the percentage of non-nominal antecedents in GerDraCor-Coref, while the number from Chen, Su, and Tan (2010) includes all mentions involved in a non-nominal coreference chain. Still, less than 1% of non-nominal antecedents seems low, but it is not clear if this is a particularity of the type of text or due to annotation errors, especially given that the agreement on annotating non-nominal antecedents is not very high (see Section 4.3).
Reiter and Frank (2010) report 13.2% of the annotated entities in the ACE-2 corpus (Mitchell et al. 2003) to be generic entities. However, this number is not directly comparable to the 3.47% of GerDraCor-Coref, as this number represents the percentage of generic mentions, not entities.

Flag	Percent
Generic	3.47
Non-Nominal	0.81
Predicate	1.42

Table 4.5.: Percentages of number of times the flags *generic*, *predicate* and *non-nominal* were annotated in GerDraCor-Coref.

A look at the Part-of-Speech (PoS) distribution of the mentions of the different corpora in Figure 4.9 reveals some further particularities of GerDraCor-Coref. The texts were automatically tagged with PoS tags using the Mate Tools tagger (Bohnet and Nivre 2012).[7] Shown are the percentages of PoS tags in the different corpora, but only for tokens inside mentions. The PoS categories *ADJ* (adjective), *ART* (article), *NE* (named entity), *NN* (normal noun), *PRON* (pronoun) and *PUNCT* (punctuation) are listed explicitly, while the *Other* category contains all other PoS labels. GerDraCor-Coref's

[7] `http://code.google.com/p/mate-tools/`

mentions contain a considerable amount of punctuation (8.03%, compared to 0.03%, 0.06% and 0.04% in the other corpora) since full stops are often part of a speaker tag, which was included into the annotation. Furthermore, it is notable that compared to the other corpora, GerDraCor-Coref contains much more pronouns (34.67%) and less adjectives (4.15%). The extensive use of pronouns might point to the deictic nature of dramatic texts, but the little use of adjectives is puzzling since one would expect dramatic texts to refer to characters with descriptive expressions.
A possible explanation is that GerDraCor-Coref contains less NP mentions compared to the other corpora, and therefore, adjectives have overall less opportunities to be used. Maybe another explanation could be that dramatic texts make little use of adjectives outside of character descriptions, whereas newspaper texts or radio news use descriptive noun phrases also for places and/or objects.

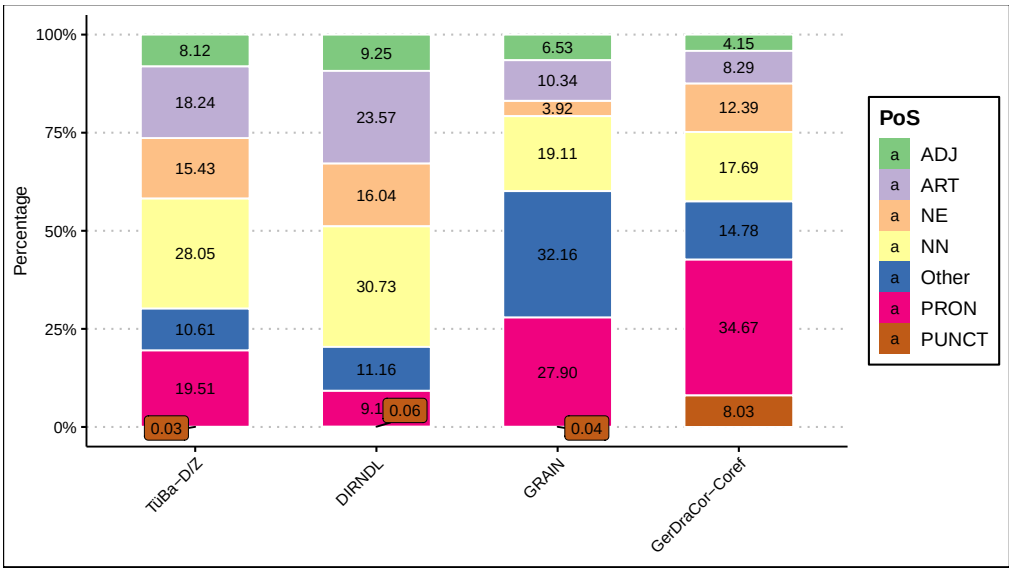


Figure 4.9.: Overview of PoS distribution on corpora TüBa-D/Z, DIRNDL, GRAIN and GerDraCor-Coref. *Other* includes all PoS categories not explicitly named.

Figure 4.10 attempts to provide an answer to this question. Shown are the percentages of adjectives across different NE categories, namely *LOC* (location), *PER* (person) and *ORG* (organization). For automatically determining the NE tags, the Stanford Named Entity Recognizer (Finkel, Grenager, and Manning 2005) has been used.[8] If adjectives are actually primarily used for character descriptions in GerDraCor-Coref, then adjectives should primarily be seen for the *PER* label and less for the *LOC* and *ORG* labels.

[8] `https://nlp.stanford.edu/software/CRF-NER.shtml`

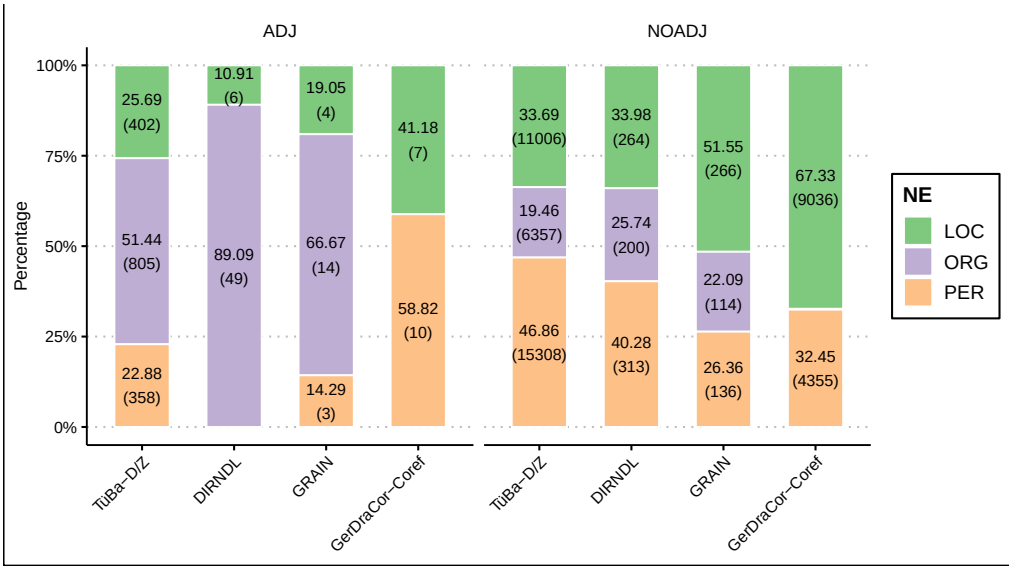


Figure 4.10.: Percentage of mentions containing adjectives and not containing any adjective and were tagged with a certain NE label in the corpora TüBa-D/Z, DIRNDL, GRAIN and GerDraCor-Coref. Absolute numbers are given in brackets below the percentages.

Figure 4.10 seems to confirm the hypothesis that GerDraCor-Coref uses adjectives primarily in character mentions, while for the other corpora, adjectives are much more prevalent in mentions referring to locations or organizations. The lack of mentions labeled as *ORG* in GerDraCor-Coref increases this difference. The figure also shows mentions that do not contain an adjective for comparison. This confirms that GerDraCor-Coref contains more *LOC* mentions than *PER* mentions, but much more *PER* mentions with adjectives than *LOC* mentions with adjectives.

Overall, a couple of differences became apparent that distinguish dramatic texts from commonly researched types of text such as newspaper texts:

- Entities contain more mentions on average in dramatic texts
- Dramatic texts contain more pronouns
- Mentions in dramatic texts contain less adjectives
- Dramatic texts contain more mentions to persons when containing an adjective

The following sections dive deeper into two particularities of dramatic texts: (i) coreference chains spanning long texts and (ii) references to literary characters.

4.5. Analysis of Long and Distant-Mention Coreference Chains

Theatre plays are special with regards to coreference in that they feature very long coreference chains not found in other commonly researched types of texts like newspapers. Furthermore, the length of the texts enables the use of entities which are mentioned infrequently and with mentions that have a greater distance to each other. For example, consider the excerpts from Lessing's *Miß Sara Sampson* in Fig. 4.11

Zweiter Aufzug

Siebender Auftritt

Marwood.
Du erinnerst mich, daß ich nicht gegen den Rechten rase. Der Vater muß voran! Er muß schon in jener Welt sein, wenn der Geist seiner Tochter unter tausend Seufzern ihm nachzieht. *Sie geht mit einem Dolche, den sie aus dem Busen reißt, auf ihn los.* Drum stirb, Verräter!

[...]

Vierter Aufzug

Dritter Auftritt

Mellefont.
Sieh, dieses Mördereisen riß ich ihr aus der Hand, *Er zeigt ihm den Dolch, den er der Marwood genommen.* als sie mir in der schrecklichsten Wut das Herz damit durchstoßen wollte.

[...]

Fünfter Aufzug

Zehnter Auftritt

Mellefont.
Nicht so, Sir! Diese Heilige befahl mehr, als die menschliche Natur vermag!

> Sie können mein Vater nicht sein. — Sehen Sie, Sir, *Indem er den Dolch aus dem Busen zieht.* dieses ist der Dolch, den Marwood heute auf mich zuckte. Zu meinem Unglücke mußte ich sie entwaffnen. Wenn ich als das schuldige Opfer ihrer Eifersucht gefallen wäre, so lebte Sara noch. Sie hätten Ihre Tochter noch, und hätten sie ohne Mellefont. Es stehet bei mir nicht, das Geschehene ungeschehen zu machen; aber mich wegen des Geschehenen zu strafen — das steht bei mir! *Er ersticht sich, und fällt an dem Stuhle der Sara nieder.*

Figure 4.11.: TextGrid Repository (2012e). *Lessing, Gotthold Ephraim. Miß Sara Sampson.* For a translation, see Figure C.9

Der Dolch, the dagger with which Marwood attempts to stab Mellefont and with which Mellefont eventually stabs himself is mentioned in Act 1, Scene 7, in Act 4, Scene 3 and in Act 5, Scene 10 and there also only very briefly. In between those occurrences, the dagger is never mentioned. Between the mention in Act 1 and 4 lie 9669 tokens and 191 utterances and between the mention in Act 4 and 5 lie 5520 tokens and 136 utterances. I will refer to this type of coreference chains that have significant gaps between their mentions as *distant-mention chains* in the following.

One goal of the following experiments is to find and analyze similar occurrences of distant-mention chains: objects or other types of plot devices that are mentioned near the beginning of a play and mentioned again later in the plot where they take on an important role.

A first natural question is how to define at what distance a chain can be considered a distant-mention chain. Table 4.6 shows some statistics about the distances in coreference chains in GerDraCor-Coref, TüBa-D/Z, DIRNDL and GRAIN. Distances are measured in either tokens, i.e. the number of tokens between mentions of an entity, or sentences, i.e. the number of sentences between the mentions of an entity. From the distribution of all these distance values, several summary statistics are calculated: the minimum, maximum and mean values, the median, as well as the first and third quartile.[9] It can be seen that the distances are quite different between the different corpora. While GerDraCor-Coref has quite large distances between mentions on average (905.16 for tokens and 99.91 for sentences), the values are much lower for the other corpora.

[9] While the median marks the data point that separates the first 50% of the data from the last 50%, the first quartile is the data point separating the first 25% from the last 75% of the data and the third quartile is the data point separating the last 25% from the first 75% of the data. Together with minimum and maximum value, this gives an idea about the coarse distribution of all data points.

	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
GerDraCor-Coref						
Tokens	1	8	31	905.16	279	37 665
Sentences	1	2	5	99.91	38	3986
TüBa-D/Z						
Tokens	1	8	24	76.58	74	3703
Sentences	1	1	2	5.04	5	208
DIRNDL						
Tokens	1	9	20	25.44	36	162
Sentences	1	1	1	1.93	3	9
GRAIN						
Tokens	1	5	20	85.55	73	1772
Sentences	1	1	2	6.84	6	123

Table 4.6.: Summary of the distances in coreference chains, showing minimum (Min.) and maximum (Max.) values, mean value, as well as the median and the first (1st. Qu.) and third (3rd. Qu.) quartiles.

We opt for the value of the 3rd quartile as the minimum distance for mentions to be considered a distant-mention chain, since it represents the value most distances fall into in the top 75% of all values. Doing so, we get the values shown in Table 4.7, where the minimum value is set to the value of the 3rd quartile of Table 4.6 and all values come only from those mentions with a distance of this value or higher. Shown are once again the minimum, maximum and mean values, the median, as well as the first and third quartile for the distant-mention chains. We can once again see that compared to the other corpora, mentions in GerDraCor-Coref have a much larger distance on average, also when only considering distant-mention chains. TüBa-D/Z and GRAIN roughly have the same amount of distances on average, even though GRAIN's average document length is much larger than TüBa-D/Z's (compare Tab. 4.3). This suggests that TüBa-D/Z's chains span large parts of the document, while the distances in GRAIN are more local. The distances in DIRNDL are quite small.

Another type of coreference chain featured in drama are *long chains*, usually coreference chains of characters. With the term *long*, I am referring to the number of mentions inside the chains as well as the distance it spans from its first to its last occurrence. For example, chains of characters usually span the entirety of the play, especially main characters, and

	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
GerDraCor-Coref						
Tokens	279	647	1597	3501.93	4336.00	37 665
Sentences	38	85	190	381.93	489.00	3986
TüBa-D/Z						
Tokens	74	106	159	244.56	277.00	3703
Sentences	5	6	9	13.56	15.00	208
DIRNDL						
Tokens	36	43	52	55.98	62.75	162
Sentences	3	3	3	3.65	4.00	9
GRAIN						
Tokens	73	105	179	287.16	363.50	1772
Sentences	6	8	13	19.02	24.00	123

Table 4.7.: Summary of the distances in distant coreference chains, showing minimum (Min.) and maximum (Max.) values, mean value, as well as the median and the first (1st. Qu.) and third (3rd. Qu.) quartiles.

the density of mentions in the chains is very high, meaning that characters are mentioned very frequently throughout the play.

Long chains pose a problem for CR, since computational models are notoriously bad at keeping track of long-distance phenomena and the handling of long documents and phenomena often needs to be addressed and implemented after a particular method for short documents has already been established (cf. Hochreiter and Schmidhuber 1997; Beltagy, Peters, and Cohan 2020; Hudson and Al Moubayed 2022; Xia, Sedoc, and Van Durme 2020).

Table 4.8 displays an overview of the length of coreference chains in different corpora, while length is measured either in tokens or sentences. This means that with a value of 0, no tokens or sentences lie between the first and last mention (i.e. singletons in the case of tokens or the whole entity is mentioned in only one sentence in the case of sentences) and a value of 1 would mean that one token or sentence lies between the first and last mention, and so on. A comparison reveals that once again, GerDraCor-Coref behaves differently from the other corpora in that it features much longer coreference chains on average. The distributions of the length of chains are generally comparable to the distribution of distant-mention chains in Table 4.7. Since for all corpora the mean values are much higher than the median values, there are only a few extremely long chains,

however for GerDraCor-Coref this is more the case than for all others and its long chains are longer than in the other corpora.

	Min.	1st. Qu.	Median	Mean	3rd. Qu.	Max.
GerDraCor-Coref						
Tokens	0	16	47	3433.73	1790.00	49 460
Sentences	0	1	4	357.20	184.00	5158
TüBa-D/Z						
Tokens	0	7	29	160.86	170.00	4597
Sentences	0	0	1	7.89	8.00	244
DIRNDL						
Tokens	0	16	31	37.30	55.00	169
Sentences	0	1	2	2.13	3.00	9
GRAIN						
Tokens	0	11	27	184.01	124.75	2044
Sentences	0	0	1	9.80	6.00	130

Table 4.8.: Summary of the lengths of coreference chains, showing minimum (Min.) and maximum (Max.) values, mean value, as well as the median and the first (1st. Qu.) and third (3rd. Qu.) quartiles.

The observations in Table 4.7 and 4.8 go against the observation by Toshniwal et al. (2020, p. 8519) that "[i]n practice, we find that most entities have a small spread (number of tokens from first to last mention of an entity) [...]"; at least for long texts that are dramatic. Disregarding long chains in computational models, accepting small losses in accuracy, would also be problematic for any kind of CLS analysis, since long chains are often most interesting from a literary point of view. The topic of the computational handling of long and distant-mention chains will be re-visited in Chapter 5.

4.6. Using Coreference Annotations to Examine Literary Characters and Topics

Table 4.10 shows the most mentioned entities across plays and annotators, sorted by frequency. Since annotators were able to assign custom names to entities, nouns with different grammatical properties in the assigned names were adjusted manually. Generic entities are mentioned most often (*der Mensch*, eng. *the human*; *die Welt*, eng. *the*

world; *die Natur*, eng. *the nature*), but also locations (*der Himmel*, eng. *the sky* or *the heaven*; *Hölle*, eng. *hell*), feelings (*die Liebe*, eng. *the love*; *Hoffnung*, eng. *hope*), body parts (*mein Herz*, eng. *my heart*; *die Hände*, eng. *the hands*) and important props (*der Brief*, eng. *the letter*; *die Tür*, eng. *the door*; *Gift*, eng. *poison*). The table also gives the number of plays a certain entity appears in, giving a feeling for the distribution of concepts across plays. *Die Welt* appears in the largest number of plays (19 plays), followed by *die Menschen* and *der Himmel* (11 plays). Most other frequently mentioned entities appear in four to five plays on average.

	Entity Denomination	Count	Num. of Plays
1	Die Menschen	181	11
2	Die Welt	168	19
3	Der Himmel	158	11
4	Mein Herz	156	4
5	Das Volk	71	4
6	Der Teufel	64	9
7	Das Leben	57	8
8	Die Natur	49	5
9	Hölle	35	4
10	Die Augen	31	8
11	Der Liebe	27	5
12	Die Hand	22	5
13	Deutschland	21	4
14	Ein Weib	21	5
15	Meine Seele	21	4
16	Hoffnung	20	5
17	Den Brief	19	4
18	Die Zeit	19	4
19	Die Tür	18	4
20	Das Gesetz	16	4
21	Die Hände	16	5
22	Den Tisch	15	4
23	Das Wort	14	5
24	Die Sonne	14	5

25	Ein Glück	14	4
26	Ein Mann	14	5
27	Dein Herz	13	4
28	Den Kopf	13	4
29	Das Herz	12	4
30	Die Menschheit	12	4
31	Die Arme	11	4
32	Die Nacht	11	4
33	Mut	11	4
34	Die Wahrheit	9	4
35	Gift	7	4

Table 4.10.: Names given by the annotators for entities that occur in at least four plays, the count of the times this entity is mentioned and the number of plays it occurs in. Entities that only consist of stopwords were filtered out.

Given coreference information, a number of analyses becomes possible. For instance, the relationship of character utterances and character mentions can be compared. Figures 4.12, 4.13 and 4.14 display the appearing characters of the plays *Die Familie Schroffenstein* by Heinrich von Kleist, *Miß Sara Sampson* by Gotthold Ephraim Lessing and *Die natürliche Tochter* by Johann Wolfgang Goethe and the number of their utterances as well as the number of mentions of these characters by other characters, over the course of the play. Especially interesting are characters that do not speak much but are frequently mentioned by others, since they seem to play an important role in the plot of the play but would be missed if only looking at utterances. The use of mentions in addition to the number of utterances provides information about these characters that would be lost if only utterances would be considered. Some characters appear at the beginning of a play, but disappear afterwards to only be mentioned by other characters (The *Aldöbern* in Fig. 4.12, *Arabella* in Fig. 4.13 or the *König* in Fig. 4.14); some characters disappear temporally, but are still mentioned regularly during this disappearance (*Rupert* in Fig. 4.12, *Sir William* in Fig. 4.13 or *Eugenie* in Fig. 4.14) and many characters only appear once and are also never mentioned outside of this appearance.

We can also more directly look at which character mentions which other character more

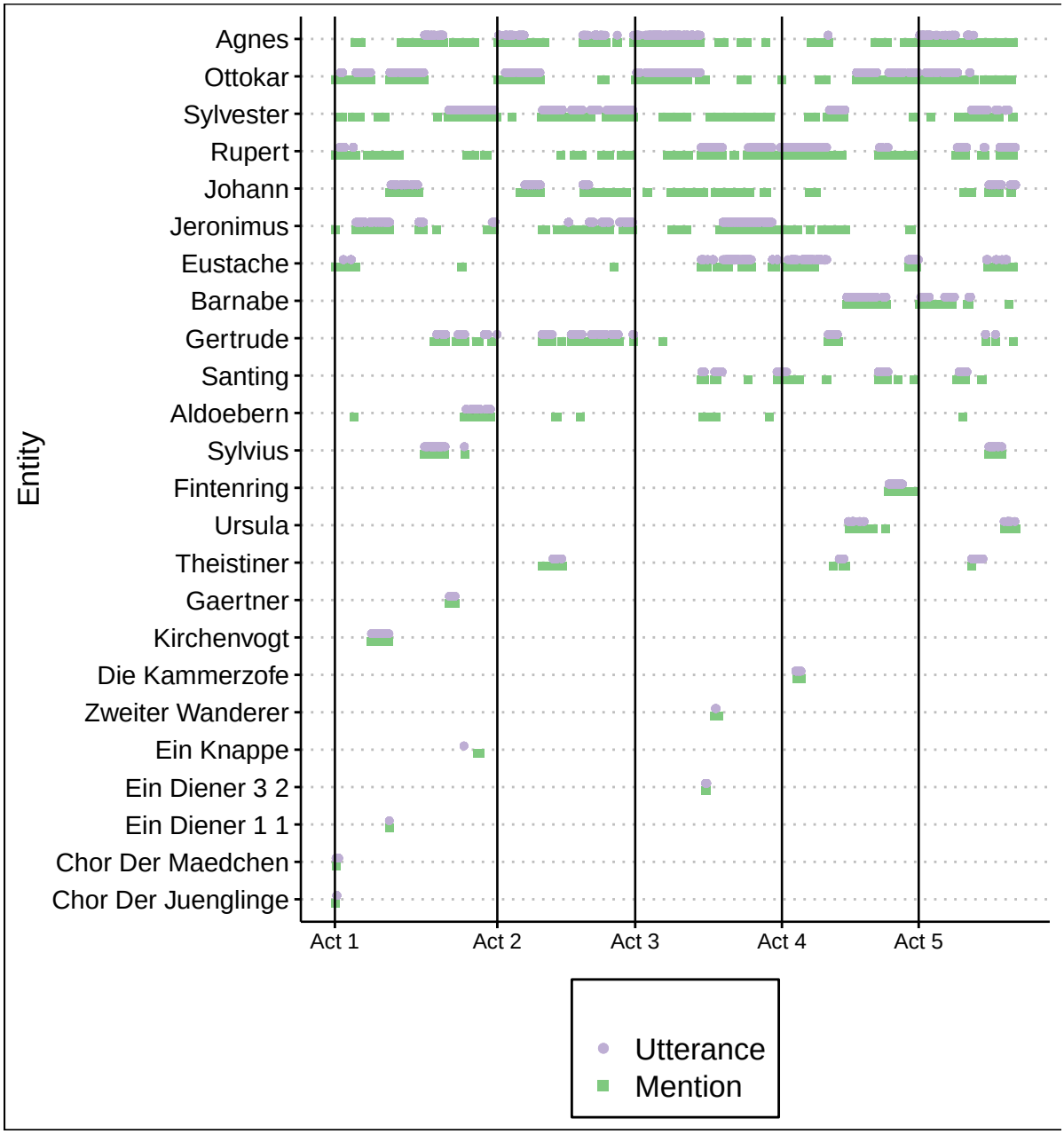


Figure 4.12.: Utterances and mentions for characters in Heinrich von Kleist's *Die Familie Schroffenstein*. Each utterance is depicted by a purple-coloured dot, while each mentions is represented by a green square. The position on the x-axis corresponds to the relative position in the text. Act boundaries are marked by vertical lines. Only entities which appear as characters are included. Characters are sorted in decreasing order by the number of their mentions.

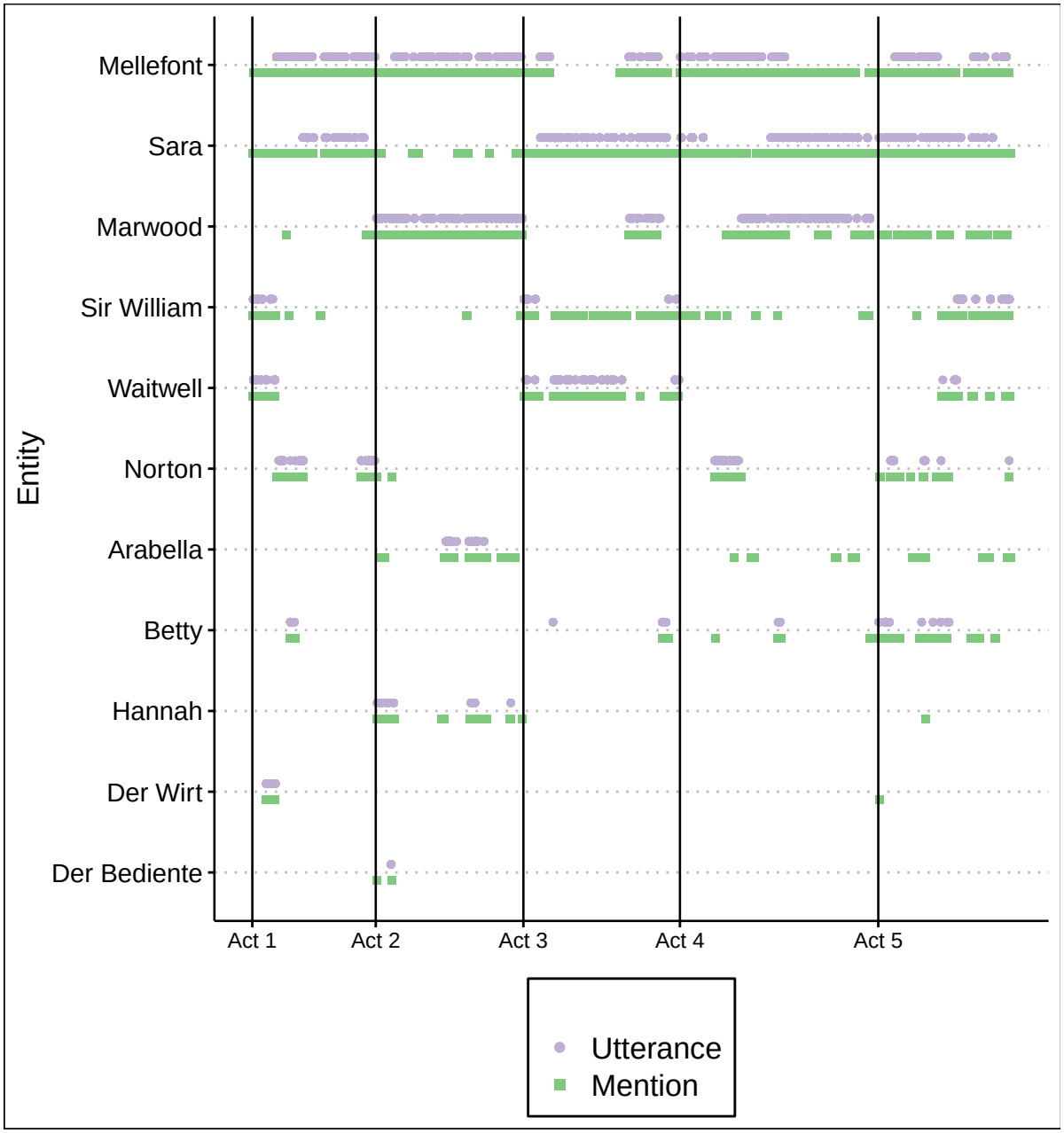


Figure 4.13.: Utterances and mentions for characters in Gotthold Ephraim Lessing's *Miß Sara Sampson*. Each utterance is depicted by a purple-coloured dot, while each mentions is represented by a green square. The position on the x-axis corresponds to the relative position in the text. Act boundaries are marked by vertical lines. Only entities which appear as characters are included. Characters are sorted in decreasing order by the number of their mentions.

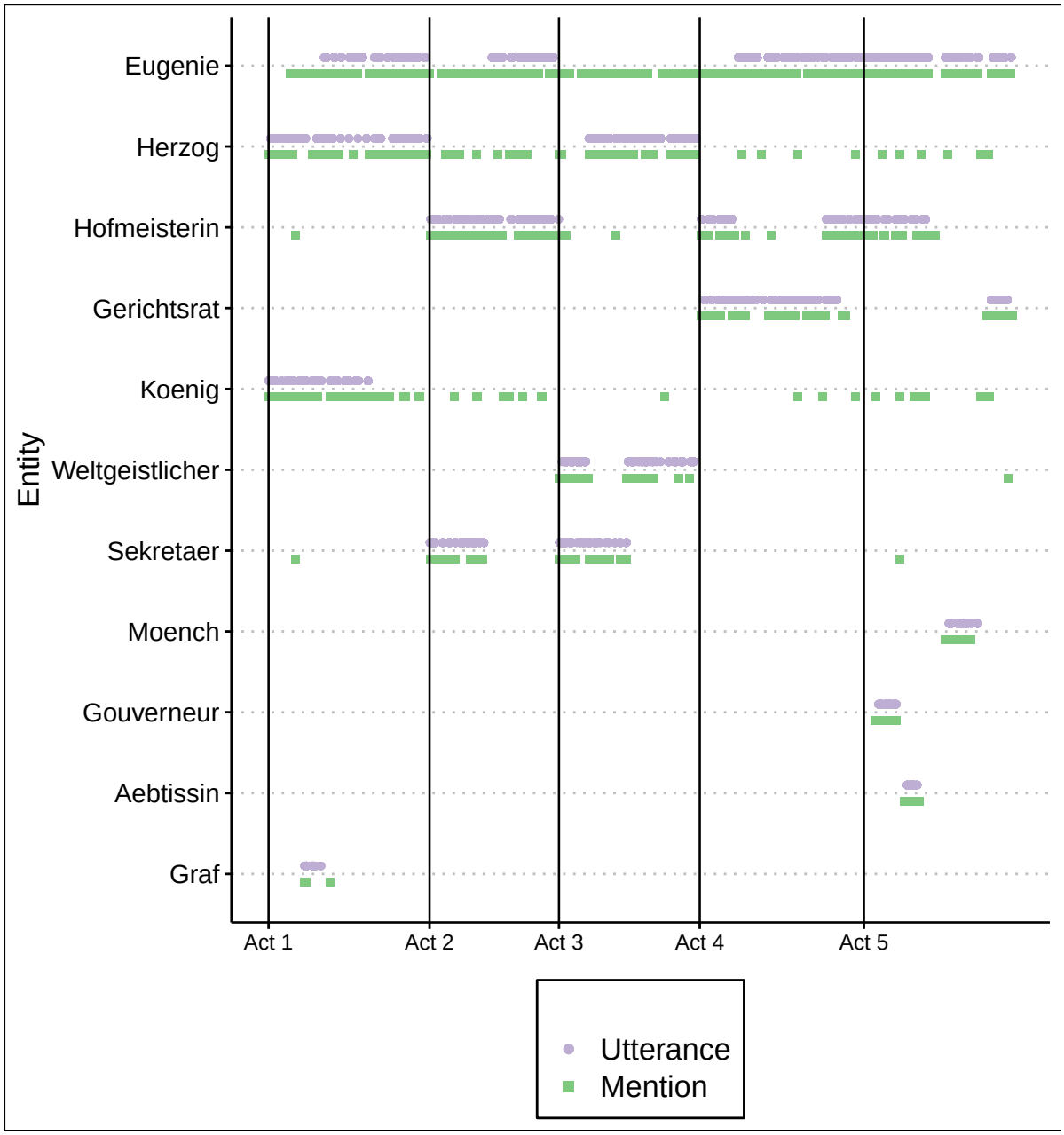
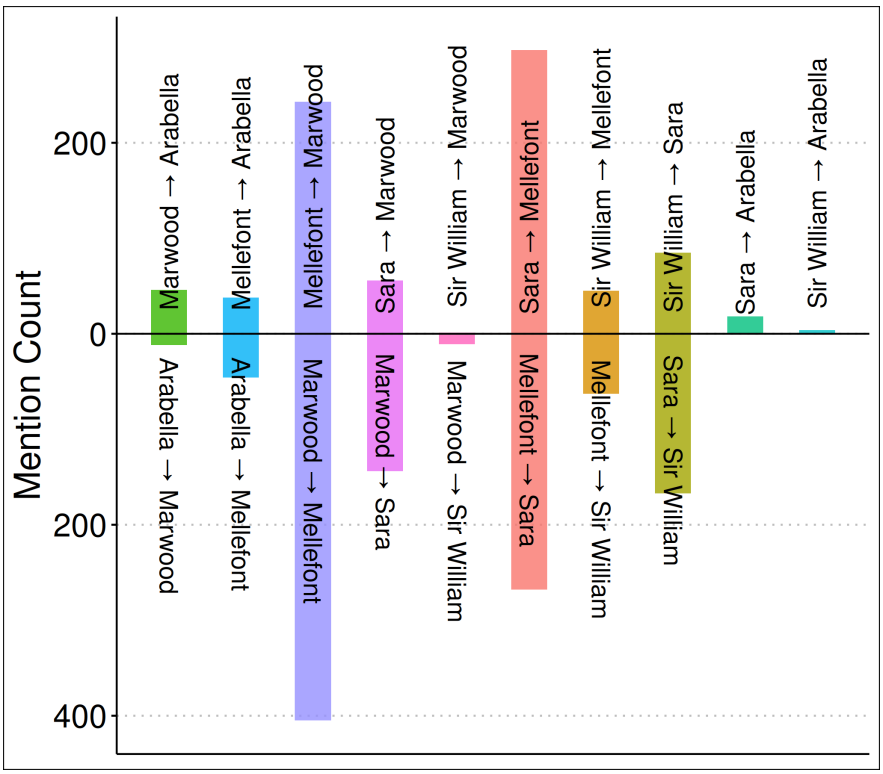


Figure 4.14.: Utterances and mentions for characters in Johann Wolfgang Goethe's *Die natürliche Tochter*. Each utterance is depicted by a purple-coloured dot, while each mentions is represented by a green square. The position on the x-axis corresponds to the relative position in the text. Act boundaries are marked by vertical lines. Only entities which appear as characters are included. Characters are sorted in decreasing order by the number of their mentions.

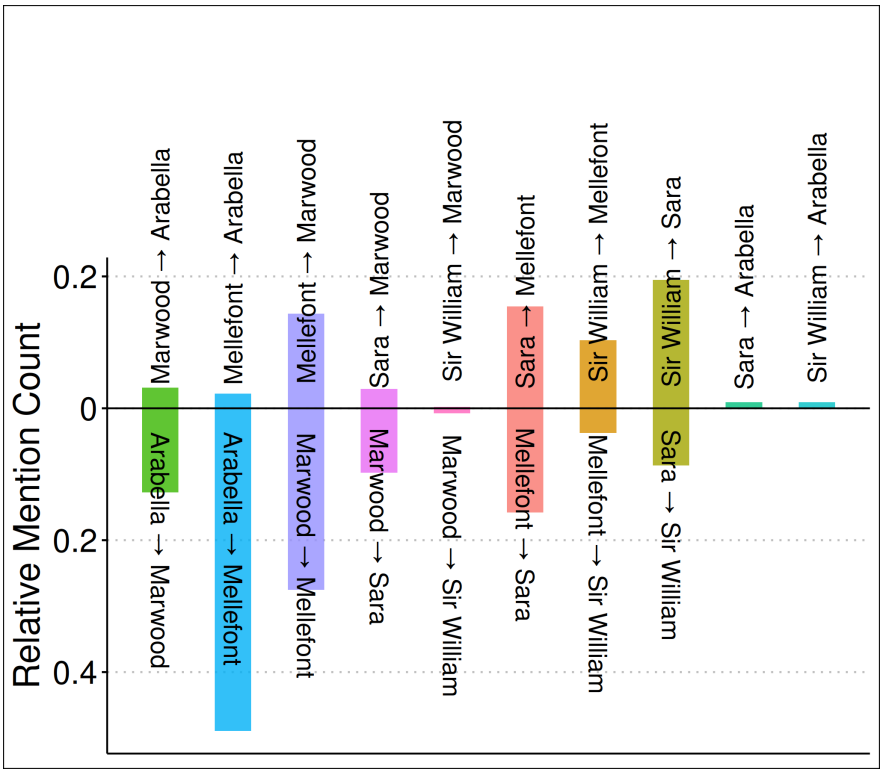
often in direct comparison. Figure 4.15a shows the main characters of Lessing's *Miß Sara Sampson* and how often they mention each other.
The following is a short synopsis of the play. Sara and Mellefont eloped and are searched for by Sara's father Sir William, who wants to make amends with the two. Marwood is a former lover of Mellefont and attempts to win him back and to get the better of her new rival Sara. Furthermore, Marwood and Mellefont have a mutual child called Arabella, who is used as leverage by Marwood against Mellefont.
There are some obvious imbalances in Figure 4.15a: Marwood mentions Mellefont much more often than Mellefont mentions Marwood; the same goes for Marwood and Sara, Sara and Mellefont as well as Sara and Sir William. In general, it seems that female characters mention male characters much more often than the other way around. The only exception seems to be the pair Mellefont–Arabella, which is relatively equal. Sara mentions Arabella, but is never referred to by her (the same for Sir William and Arabella). Looking at the relative counts, where "relative" means relative to the amount a character mentions any other entity (note: not only other characters), we can see that the apparent equal pair of Mellefont and Arabella is not very equal for the relative counts: almost half of all mentions that Arabella uses in the play are for Mellefont.
We can examine if the observation for *Miß Sara Sampson* — that female characters mention male characters much more often than male characters mention female characters — holds for the whole corpus. Figure 4.16 shows the absolute and relative counts of how often male and female characters mention each other. In Figure 4.16a we can see that male characters mention other male characters much more often than female characters mention other female characters. In terms of absolute numbers, female characters mention male characters almost as often as male characters mention female characters, which speaks against the hypothesis that female characters mention male characters more often than the other way around. However, Figure 4.16b shows that when looking at the relative numbers, the distribution is as speculated: relative to the amount of mentions they make in general, female characters do mention male characters much more often than the other way around.

4.7. Summary

This chapter presented a corpus of German theater plays annotated for coreference, called GerDraCor-Coref. The corpus, while smaller than other German-language corpora like TüBa-D/Z in terms of tokens, contains a competitive number of mentions and entities.

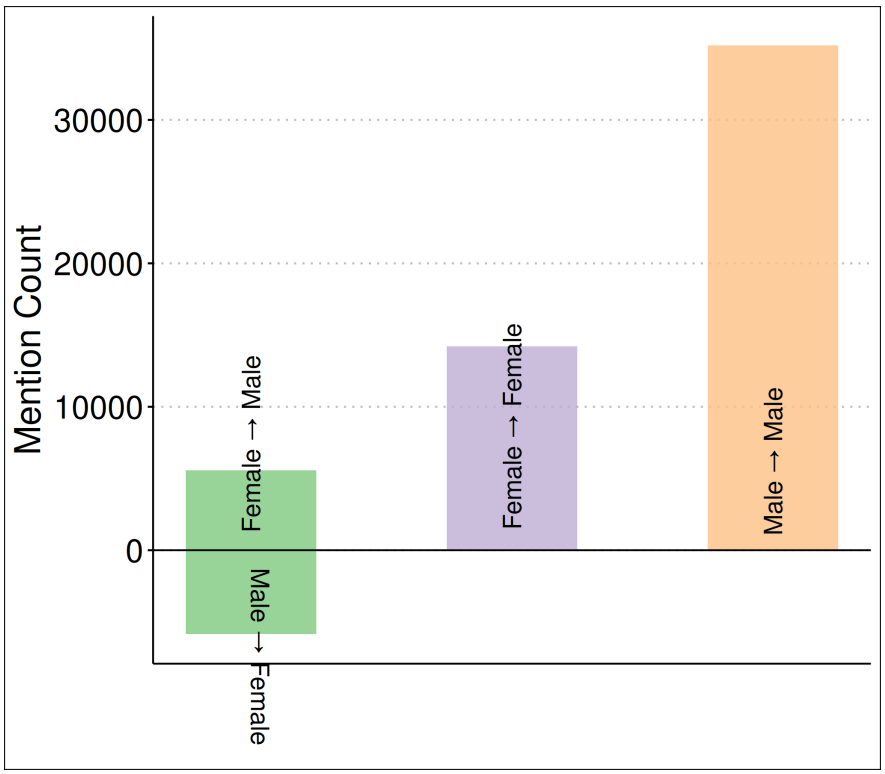


(a) Absolute mention count

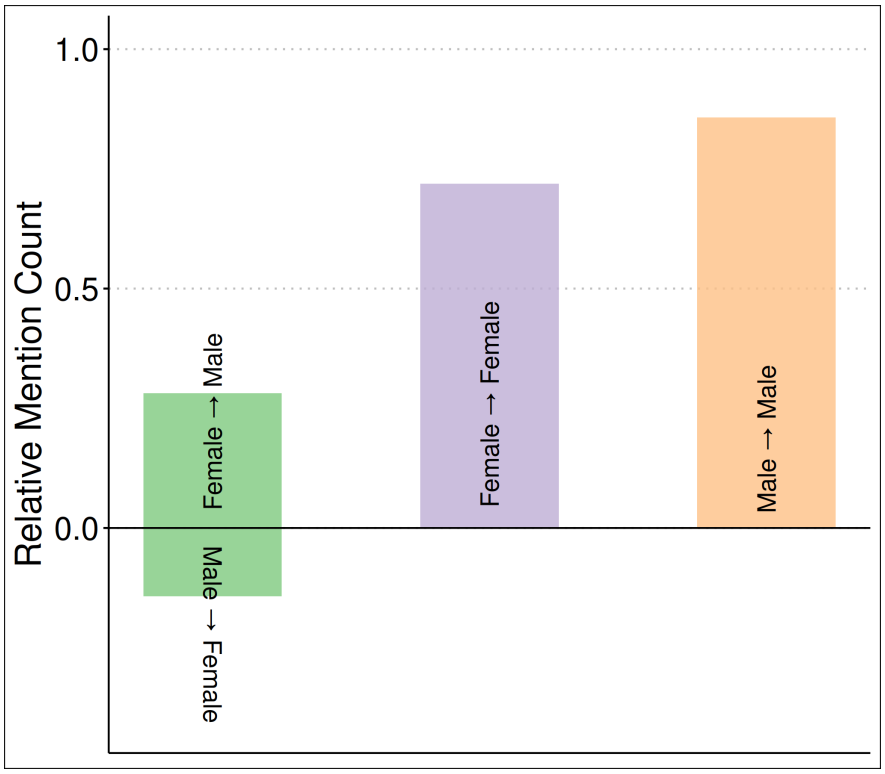


(b) Relative mention count

Figure 4.15.: Counts of how often the main characters of Lessing's *Miß Sara Sampson* mention each other. Both absolute counts (a) as well as counts relative to the number of mentions a character makes in general (b) are given.



(a) Absolute mention count



(b) Relative mention count

Figure 4.16.: Mention counts of male and female characters. Both absolute counts (a) as well as counts relative to the number of mentions a character makes in general (b) are given.

The corpus was annotated using the CorefAnnotator (Reiter 2018) and is available in different formats (XMI, CoNLL, TEI). An inter-annotator study showed that agreement scores are lower for GerDraCor-Coref than for other corpora, wich might hint at plays being more difficult to annotate than other types of (literary) texts. The average number of mentions is much larger in GerDraCor-Coref compared to other corpora, however, interestingly the number of entities is much smaller on average. This might hint at a fundamental difference of dramatic texts compared to other types of texts: less coreference chains overall, but much longer and much more dense chains. An analysis of long and distant coreference chains showed that long text formats such as plays also contain different coreference structures than corpora containing other types of text, which should be considered when attempting to resolve the references. The annotations can be used to explore pan-corpus entities, such as generics and important plot devices present in many plays, as well as perform single play analyses about the references to characters by other characters, delivering potentially new insights into the structure of the plays in a quantitatively tangible manner. It could be shown that overall, female characters mention male characters much more frequently than male characters mention female characters and that male characters mention other male characters more frequently than female characters mention other female characters. Moreover, the number of times, a character of a certain gender mentions a character of the same gender is much higher than a character of a certain gender mentioning a character of the opposite gender. All this combined may hint at certain social dynamics in the plays: Characters of a certain gender stay among themselves, at least in terms of mentions, but when characters mention a character of the opposite gender, female characters are more dependent on male characters than the other way around.

Wie soll meine Seele alle diese Rätsel auflösen?
(How is my soul to resolve all these riddles?)

Lucie in Johann Gottlob Benjamin Pfeil's "Lucie Woodvil"

5

Coreference Resolution for Theatre Plays

This chapter explores the possibilities of automatically resolving coreferences in dramatic texts. Coreference denotes a linguistic phenomenon in which two or more expressions, so called mentions, refer to the same real or fictional world entity. For instance, in the sentence "The cat eats its meal", the expressions or mentions *the cat* and *its* refer to the same entity, in this case the same animal. As the data source, the corpus GerDraCor-Coref is used, which has been described in the previous Chapter 4. Next to GerDraCor-Coref, other corpora with non-dramatic language are used to compare the performances to GerDraCor-Coref and to reveal potential unique challenges that dramatic texts might pose. In Section 5.2, a rule-based system to perform coreference resolution on dramatic texts is presented, called DramaCoref. DramaCoref resolves coreferences by applying a series of passes, which are rules that decide if two mentions could potentially refer to the same entity. The mentions are extracted using constituent parsers component, whose performance is also compared to using a neural network for mention extraction (Section 5.1). Passes are ordered by their precision so that more precise passes decide first on the mention's affinity to an entity. Once all mention- and cluster-pairs are judged by DramaCoref, the result can be compared to the manually annotated coreferences of GerDraCor-Coref and evaluation scores can be computed (Section 5.3). Different setups, either data-wise or for different settings of DramaCoref, are also evaluated in order to investigate strengths and weaknesses of DramaCoref and to learn more about the data

sources. The performance of DramaCoref is also compared to other available coreference resolution systems which were designed to work on newspaper data. Lastly, mistakes by DramaCoref are analysed to get a more qualitative and detailed image of potential issues and phenomena (Section 5.4).
The results presented in this chapter have been partially published in Pagel and Reiter (2021) and have been enriched with new experiments and results.

5.1. Mention Detection

The rule-based system called DramaCoref described in the following section (5.2) requires already detected mentions on which it applies its passes. Apart from this, mention detection is also an interesting task by itself, since the question which phrases in a sentence refer to an entity is not always obvious, since syntactic placeholders like expletives or idioms are typically considered to not refer.
For this chapter, two mention detection setups were experimented with: (i) a setup in which mentions are detected by using all NPs of the output of a syntactic parser, and (b) a setup in which mentions are detected using a neural transformer model. The goal is to find the setup which yields the best results and which can be used in the following experiments on DramaCoref. While the use of syntactic parsers are a traditionally often used way to receive the mentions for CR (Soon, Ng, and Lim 2001; Rösiger and Kuhn 2016; Tuggener 2016), neural network models are a promising alternative, as they enable to apply mention detection end-to-end without relying on the pipeline input of other linguistic analyses and enable to tackle to task of mention detection directly without for example interpreting all noun phrases returned by a syntactic parser as mentions. Yu, Bohnet, and Poesio (2020) were able to show that their LSTM-based model was able to outperform other state-of-the-art mention detection systems and that given the output of their neural mention detection system as the input for other CR systems yielded a small improvement in evaluation scores. Hence, next to the output of a syntactic parser, this chapter experiments with using a transformer model to detect the mentions of GerDraCor-Coref.

5.1.1. Mention detection using syntactic parsers

Two different syntactic parsers are used to predict mentions in this setup: The Berkeley parser (Petrov et al. 2006; Petrov and Klein 2007) and the Stanford parser (Klein and

Manning 2002; Rafferty and Manning 2008) with their respective German models. Both parsers are statistical constituent parsers. The implementation for both parsers comes from the UIMA-based DKPro tool (Eckart de Castilho and Gurevych 2014) in version 1.7.0.[1] In both cases, all NPs and pronouns[2] predicted by the two parsers are considered to be mentions.

5.1.2. Neural mention detection

For the neural mention detection setup, several models are considered which are all downloaded from the HuggingFace platform [3]. All models are based on the transformer architecture (Vaswani et al. 2017). One model is the German version of BERT (Devlin et al. 2019) in the HuggingFace implementation (110M parameters)[4]. Another model is based on the German version of DistilBERT (Sanh et al. 2019) with 67.4M parameters[5]. A third model is a German ELECTRA model (Clark et al. 2020), which is a modification of the BERT algorithm (110M parameters)[6]. The final model is a German BERT model which was fine-tuned on German literary and historical texts (110M parameters)[7].

5.1.3. Experiments

Experiments using the Syntactic Parsers

For this experiment, the documents of GerDraCor-Coref are split into a 80% train and 20% test set. While the parsers would not need such a split, since they do not need to be trained again, this setup is chosen in order to compare the performance of the parsers to the self-finetuned neural networks, described further below.
For each token, the parser output is evaluated regarding if it correctly predicted a mention boundary, the continuation of a mention or a token not belonging to a mention.
Table 5.1 shows the results of evaluating the Stanford and Berkeley parsers on GerDraCor-Coref. Reported are the accuracy, precision, recall and F1 scores.

[1] The model files for both the Berkeley parser and Stanford parser can be found under `https://web.archive.org/web/20221013144037/https://dkpro.github.io/dkpro-core/releases/2.2.0/docs/model-reference.html`.
[2] The output of both parsers does not label pronouns as NPs, therefore they need to be accounted for separately.
[3] `https://huggingface.co`
[4] `https://huggingface.co/bert-base-german-cased`
[5] `https://huggingface.co/distilbert-base-german-cased`
[6] `https://huggingface.co/german-nlp-group/electra-base-german-uncased`
[7] `https://huggingface.co/severinsimmler/literary-german-bert`

Parser	Accuracy	F1	Precision	Recall
Berkeley	0.32	0.32	0.45	0.49
Stanford	0.70	0.53	0.50	0.57

Table 5.1.: Results for the mention detection using different parsers.

The Stanford parser performs clearly better than its Berkeley counterpart. With 0.53 F1-score, the performance is however still relatively low and shows that mention detection is a challenging task to perform on GerDraCor-Coref.

Experiments using the Neural Models

The setup is almost identical to the before described experiment with the two parsers: the models are fine-tunded on an 80%-20% train-test split and the models are evaluated according to if they correctly predicted a mention boundary, continuation or a non-mention token. However, additionally, the train set from GerDraCor-Coref is concatenated with the entirety of the TüBa-D/Z corpus in order to allow for more training data for the neural network to train on. All models are finetuned for four epochs.
Table 5.2 shows the results of evaluating the neural models, with accuracy, precision, recall, F1 score.

Model	Accuracy	F1	Precision	Recall
DistilBERT-German-Cased	0.76	0.29	0.25	0.33
Literary-German-BERT	0.76	0.29	0.25	0.33
BERT-German-Cased	0.76	0.29	0.25	0.33
ELECTRA-German-Uncased	0.76	0.29	0.25	0.33

Table 5.2.: Results for the mention detection using different transformer models.

It can be seen that the performances do not differ between the models. The high accuracy hints at a bias towards true negatives, i.e. the model correctly classifies many tokens to not belong to a mention but falls short on correctly predicting tokens belonging to a mention. Overall, the neural models perform worse than the parsers, suggesting that the amount of training data was not enough to successfully finetune the models. This leads to the conclusion that the parser outputs offer more robust results for this smaller dataset. In the following experiments with the rule-based system, the mentions which are fed into the system are therefore the output of the Stanford parser.

5.2. Rule-based Coreference Resolution System: DramaCoref

The rule-based system DramaCoref (Pagel 2020) is based on previous work by Raghunathan et al. (2010), Lee et al. (2011) and Lee et al. (2013)[8], and in part on the work by Krug et al. (2015), who themselves based their work on the aforementioned three papers. If not indicated otherwise, the following description of the system presents original additions to the ideas suggested by this previous work.

While rule-based systems have been largely replaced in NLP by statistical methods such as machine learning and deep learning techniques, they provide a number of advantages especially useful for CLS experiments (see also Krug et al. (2015) and van Cranenburgh (2019a) for some similar arguments):

- Rule-based systems work well on low-resource settings, as they require only a small held-out dataset to develop rules and determine the order of rules
- Rule-based systems are easy to interpret, as the reasons for the decision making of each rule are maximally transparent
- Rule-based systems are robust towards unseen domains and data as they do not require special training for new domains (however they might require the development of new rules)
- It is possible to directly use expert knowledge to develop rules specifically catered towards literary texts

In addition to the aforementioned points, the experiments presented in Section 5.1 clearly highlighted how neural models performed worse than their machine-learning-based counterparts for mention detection on dramatic texts. In light of this, this chapter opts to utilize a rule-based system over alternatives, especially since interpretability and the consideration of expert knowledge are important factors when working within the interdisciplinary field of computational literary studies.

5.2.1. Mention detection and ordering

The system cannot detect mentions by itself and relies on the input of other resources. To this end, the output of the Stanford constituent parser from Section 5.1 is taken and all noun phrases detected by the parser are taken as input for the rule-based system. All mentions are ordered by occurrence in the texts, with the first occurring mentions being

[8]The last two papers were an incremental improvement of the system and the experiments presented in Raghunathan et al. (2010).

processed first. In case of overlapping mentions, the mentions which occur first when traversing the constituent trees in a breadth-first fashion are processed first.

5.2.2. Passes and sieve

The core component of DramaCoref is the so-called *multi-pass sieve*, in which single "passes" or rules determine if a potential mention-antecedent pair belongs in the same coreferential cluster. The totality of the passes is the sieve, with the metaphorical idea of all mention-antecedent pairs being "sieved" and only valid pairs coming out of the process (Raghunathan et al. 2010, pp. 492–93). The passes are ordered by precision, with the precision of a pass being determined on a held-out dataset.[9]
An overview of all passes used in DramaCoref can be found in Table 5.3.

Pass ID	Short Name	Source
1	ExactMatch	[1]
2a	Acronyms	[1]
2b	Appositions	[1]
2c	RelPron	[1]
2d	ReflexivePron	New
3	StrictHeadMatch	[1]
4	StrictHeadMatchVar1	[1]
5	StrictHeadMatchVar2	[1]
6	HeadEntail	[1]
7	Pron3rdPers	in [1], this pass handles all pronouns
9a	SpeakerPron1stPers	No. 2 "Discourse Processing" in [2], modified
9b	SpeakerPron2ndPers	No. 2 "Discourse Processing" in [2], modified
10	RelaxedStringMatch	[2]
11	ProperHeadWordMatch	[2]
11a	PoperHeadWordMatchVar1	New
12a	LexicalSynonym	[2], modified
12b	LexialHyponym	[2], modified
14	ExactLemmaMatch	New

Table 5.3.: Overview of the passes that are implemented in DramaCoref and their sources. [1] refers to Raghunathan et al. (2010), [2] refers to Lee et al. (2011).

[9] This approach is similar, but not exactly identical, to the one found in Lee et al. (2013, pp. 905–06), who first order the passes based on linguistic intuition, and later find that determining the order of the passes automatically yields more or less the same results as ordering manually.

Pass 1: Exact Match This pass matches a mention-candidate pair if their surface forms are exactly identical. Only non-pronominal noun phrases are considered. Figure 5.1 shows the mentions pass 1 is able (and not able) to resolve. Note that *meine Mutter* (*my mother*) and *meiner Mutter* (*my mother's*) are incorrectly detected as belonging to different entities by this pass, since the surface forms are not completely identical. However, passes ordered to be applied later might correctly group these two mentions into the same entity.

Siebenter Auftritt
Emilia. Odoardo.

EMILIA.
Wie? Sie hier, [mein Vater]$_0$? — Und nur Sie? — Und [meine Mutter]$_1$? nicht hier? — Und [der Graf]$_2$? nicht hier? — Und Sie so unruhig, [mein Vater]$_0$?

ODOARDO.
Und du so ruhig, [meine Tochter]$_3$?

EMILIA.
Warum nicht, [mein Vater]$_0$? — Entweder ist nichts verloren: oder alles. Ruhig sein können, und ruhig sein müssen: kömmt es nicht auf eines?

ODOARDO.
Aber, was meinest du, daß der Fall ist?

EMILIA.
Daß alles verloren ist; — und daß wir wohl ruhig sein müssen, [mein Vater]$_0$.

ODOARDO.
Und du wärest ruhig, weil du ruhig sein mußt? — Wer bist du? Ein Mädchen? und [meine Tochter]$_3$? So sollte der Mann, und [der Vater]$_4$ sich wohl vor dir schämen? — Aber laß doch hören: was nennest du, alles verloren? — daß [der Graf]$_2$ tot ist?

EMILIA.
Und warum er tot ist! Warum! — Ha, so ist es wahr, [mein Vater]$_0$? So ist sie wahr die ganze schreckliche Geschichte, die ich in dem nassen und

> wilden Auge [meiner Mutter]$_5$ las? – Wo ist [meine Mutter]$_1$? Wo ist sie hin, [mein Vater]$_0$?

Figure 5.1.: Snippet from TextGrid Repository (2012d): *Lessing, Gotthold Ephraim. Emilia Galotti*, showing the coreferences that pass 1 is able to find. For a translation, see Figure C.10

Pass 2a-c: Acronyms, Appositions, Relative and Reflexive Pronouns This group of passes handles fixed morphological and syntactic constructions, in particular acronyms, appositions, reflexive and relative pronouns. Acronyms are defined as sequences of at least two characters which are all upper case and covered by pass 2a. In Raghunathan et al. (2010), appositions are understood as mentions which modify the antecedent directly following, for example *[[chancellor]$_1$ Angela Merkel]$_1$*. In GerDraCor-Coref, the modifier, for example *chancellor*, is not annotated separately, i.e. the whole phrase *chancellor Angela Merkel* would just be one mention. Since DramaCoref can also be applied to corpora other than GerDraCor-Coref, the pass has been implemented nevertheless and is called 2b. While there are no acronyms or appositions in GerDraCor-Coref, there are plenty examples of relative pronouns. For instance, in the sentence "Es ist solches ein kleiner Trost in dem Verdrusse, den sie mir dadurch verursacht, daß sie noch nicht von mir scheiden will." (eng. *This is a small consolation in the frustration she causes me by not wanting to part with me yet.*), spoken by *Der Graf* (eng. *the count*) in Johann Christian Krüger's play *Die Candidaten*, the relative pronoun *den* (eng. *which*) is coreferent to the embedding mention *dem Verdrusse, den sie mir dadurch verursacht ...*: *[dem Verdrusse, [den]$_1$ sie mir dadruch verursacht ...]$_1$*. This pass requires a i-within-i relationship in order to function. The responsible pass for relative pronouns is pass 2c. Reflexive pronouns are quite ubiquitous in German, for example *[Die Kinder]$_1$ schmiegen [sich]$_1$ an sie* (eng. *The children snuggle up to her.*, note however that the reflexive pronoun *sich* (eng. *themselves* is not used in the English translation)) in Karl Johann Braun von Braunthal's play *Faust*. They are handled by pass 2d, which assigns every occurrence of a reflexive pronoun, taken from a pre-compiled list, to the closest preceding mention which agrees in number. This pass was added since such constructions occur in German but not in English and was therefore not present in Raghunathan et al. (2010).

Pass 3: Strict Head Match Pass 3 matches mention-candidate pairs which share the same head noun. Only non-pronominal noun phrases are considered. Since matching

phrases with identical heads gives room for many mistakes, pass 3 employs a number of restrictions in order to keep precision as high as possible. The restrictions are:

1. No i-within-i
2. All modifiers in the mentions and the candidate must match
3. All non-stop words in the mention must also occur in the cluster the candidate is already part of

Pass 4: Strict Head Match – Variation 1 Passes 4 and 5 are closely related to pass 3. While pass 3 imposes certain constraints to the possible candidates, pass 4 and 5 drop one of these constraints, respectively. This is to ensure that the constraints of pass 3 are not to strict and reject many candidates as non-matches when they actually are. Pass 4 drops the constraint that all modifiers in mention and candidate must match.

Pass 5: Strict Head Match – Variation 2 Pass 5 drops the constraint that all non-stop words in the mention must also occur in the cluster of the candidate.

Pass 6: Head Entailment For pass 6 to report a match, the head of the mention is entailed in the tokens of the first mention in a candidate cluster. It applies the not-i-within-i constraint and only considers nominal mentions. For example, this pass matches the mention *Rosa* to the full name *Rosa Fiebig* in Erich Mühsam's play *Judas*.

Pass 7: Third Person Pronouns Raghunathan et al. (2010) implement a pass to handle all types of pronouns based on a number of constraints like person, gender, number, animacy and NE category. DramaCoref splits the handling of pronouns of different persons into separate passes; pass 7 only handles third person pronouns. Additionally, the pass checks if the third person pronoun and a candidate agree in gender and number[10].

Pass 9a-b: First and Second Person Pronouns Lee et al. (2011) introduce a discourse processing pass that identifies speakers and assigns first and second person pronouns to them. While not completely clear from the paper, it appears that they still use pass 7 from Raghunathan et al. (2010) (pass no. 13 in Lee et al. (2011)) in order to process first and second person pronouns where no speaker is available. Since for drama it is always possible to assign a speaker, DramaCoref repurposes this pass and makes it handle every first and second person pronouns. Pass 9a handles first person pronouns, by matching

[10]DramaCoref does not implement a check on the agreement for animacy or NE status like in Raghunathan et al. (2010).

all first person pronouns that have the same speaker. Pass 9b matches second person pronouns if the speaker is the same and the speaker of the following and/or preceding utterance is also the same (this speaker will however be different to the speaker of the utterance in which the second person pronouns occur).

Pass 10: Relaxed String Match This pass checks if the string of a mention and the string of a candidate are identical after dropping all tokens following the head of both mentions. While the presence of different modifiers usually indicates that noun phrases refer to different entities, this does not necessarily need to be the case. By dropping all possible trailing modifiers, this pass can increase the recall.

Pass 11: Proper Head Word Match This pass checks if the heads of a mention and a candidate are identical and if they do not contain modifiers that would suggest that they belong to different entities. To this end, the pass checks if the two mentions contain differing proper nouns or differing strings marked as location by the NER component of the pre-processing. Furthermore, it checks if the two mentions contain differing strings marked as numerals by the POS tagging component of the pre-processing. Additionally, the pass only allows not-i-within-i constructions and only handles nominal mentions.

Pass 11a: Proper Head Word Match – Variation 1 This pass functions the same as pass 11, but adds the constraint that the modifiers cannot contain a first person pronoun if the speakers of mention and candidate differ, or cannot contain a second person pronoun if the speakers of the following or preceding utterances differ.

Pass 12a: Lexical Synonymy This pass reads in the information given by GermaNet (Hamp and Feldweg 1997) and checks if the head of a mention is a synonym of the head of a candidate, according to the synsets of GermaNet. Additionally, only not-i-within-i constructions are permitted.

Pass 12b: Lexical Hyponymy This pass also utilized the informations of GermaNet by checking if the head of a mention is either in a hyponym or hyperonym relationship with the head of a candidate. This pass additionally adds the constraint that the nodes in GermaNet cannot be further apart than 4 and that the distance of the mention and candidate in terms of sentences apart cannot be larger than 3. This pass also requires not-i-within-i.

Pass 14: Exact Lemma Match Pass 14 checks if the lemmas for all tokens in a mention and a candidate are identical. This pass is therefore almost identical to pass 1, but adds the consideration of lemmas in order to cater better to the requirements of German where lexically identical words can take on many different grammatical forms, in contrast to English[11].

Post-processing Post-processing is implemented as a final pass. The post-processing pass merges clusters if they were identified to represent identical characters. This way, if for instance one cluster contains only pronouns which could be connected to a certain character and another cluster which contains proper names associated to the same character, these two clusters can be merged.

5.3. Coreference Resolution using DramaCoref

A set of experiments has been performed in order to examine DramaCoref's predictive power on theatre plays as well as on other types of texts, namely newspaper texts, novellas and broadcast news. In particular, the experiments can be categorized as follows:

1. Varying Domains: Plays, fairy tales, newspaper, radio news, radio interviews
2. Gold mentions vs. predicted mentions
3. Acts vs. scenes
4. Single pass performance and cumulative pass performance
5. Pronoun-only and cast-member-only
6. Use of information from dramatis personæ

The first category applies DramaCoref on domains different from plays in order to see if the system is able to generalize well outside of its intended domain. In the second category, the system is evaluated using gold mentions, i.e. using the mentions manually annotated, in order to see the upper bound on the performance of coreference on its own and to see how much the performance drops due to wrongly predicted mentions. For the third category, coreference is predicted on whole acts as before, but also on single scenes from these acts. The difference lies in the length of the document, as the system only needs to predict a chain for a single scene and might potentially make less mistakes due to error propagation of long chains. The fourth setup checks the performance of single passes, i.e. how the system performs if it consists only of a single pass instead of using multiple passes in tandem. In a different setup, the cumulative performance is evaluated,

[11]English of course also has this phenomenon, but much less pronounced.

which is gathered by applying the best performing pass first, then the best and second best performing pass together, and so on. The fifth category evaluates the performance if the system were to only resolve pronouns or only characters from the cast list, which should be an easier task than full coreference resolution. Lastly, a system is used which incorporates information from the dramatis personæ in order to merge coreference chains of cast members which otherwise might be separate.

5.3.1. Data

The main data are the plays from GerDraCor-Coref. Experiments to test the performance on other domains and types of text include material from Naumann (2007, TüBa-D/Z, newspapers)[12], Rösiger, Schulz, and Reiter (2018, novellas and fairy tales), and Björkelund et al. (2014, DIRNDL, radio news)[13].

5.3.2. Experimental Setup

The set of documents is split into two parts:

1. A development set DEV-DRAMACOREF (20% of all documents)
2. A test set TEST-DRAMACOREF (80% of all documents)

DEV-DRAMACOREF is used to order DramaCoref's passes according to their precision on DEV-DRAMACOREF. TEST-DRAMACOREF is used to evaluate the performance of DramaCoref. All evaluation results in this section are reported for TEST-DRAMACOREF, if not specified otherwise. Also, the ordering of passes for experiments on TEST-DRAMACOREF always follows the order determined on DEV-DRAMACOREF, as described before.
The system gets a file in CoNLL format as input, containing the documents of the respective set and information about the NPs present in the text. For all NPs, the system makes a decision into which coreference entity they belong and outputs a document in CoNLL format, containing a column with the coreference information. This generated CoNLL file is then compared to the original gold file and evaluated using the "Reference Coreference Scorer" (Pradhan et al. 2014)[14]. Grouped entities and non-nominal antecedents have

[12] `https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/resources/corpora/tueba-dz/`

[13] `http://www.ims.uni-stuttgart.de/data/dirndl`

[14] The scorer has been slightly modified in order to easily output results in a table format, see `https://github.com/pagelj/reference-coreference-scorers`. Furthermore, an unreleased version of the code has been used, which contains an implementation of the LEA score, see `https://github.com/conll/reference-coreference-scorers/tree/LEA-scorer`.

been excluded from the evaluation since they are not covered by the scorer.

5.3.3. Results

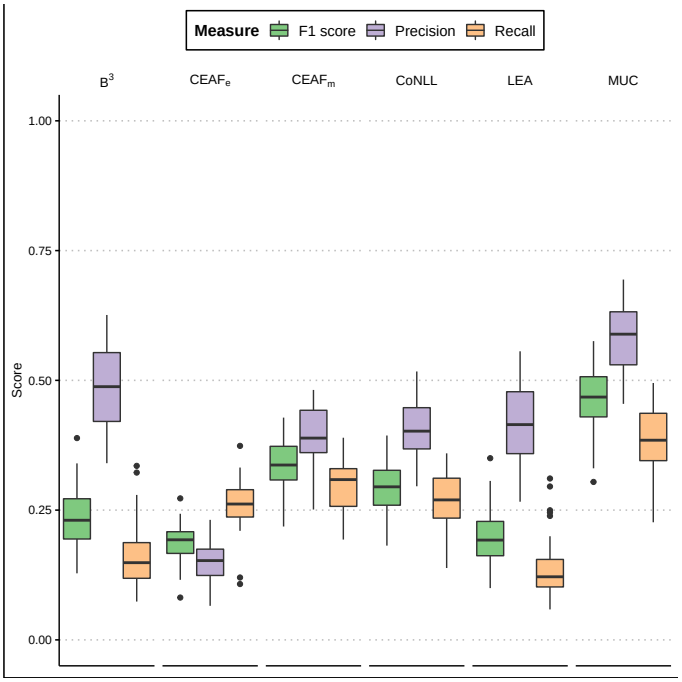


Figure 5.2.: Boxplots showing scores for metrics B^3, $CEAF_e$, $CEAF_m$, CoNLL, LEA and MUC. Shown are F1 score, precision and recall. The scores were gathered by applying DramaCoref on TEST-DRAMACOREF.

Metrics Figure 5.2 shows results for applying DramaCoref on GerDraCor-Coref, for all evaluation metrics described in Section 2.3.4 and further subdivided into precision, recall and F1 score. Since multiple documents are evaluated, instead of showing the average values, boxplots are shown for each setup. A boxplot shows general characteristics of a distribution, in particular the second and third quartile in a box, the median as a bold line in the middle of the box and the first and fourth quartile as lines leaving the box. Outliers are represented as dots. Boxplots therefore allow to visually compare high-level information of distributions of different categories or groups. It can be seen that the scores behave differently in their estimation of the goodness of prediction. $CEAF_e$ is the only score for which the precision is lower than the recall. This can be explained by $CEAF_e$ operating on the level of entities instead on the level of single mentions. MUC achieves the overall highest scores, but it also frequently gets criticized for being too

lenient in its evaluation. For simplicity reasons, the following evaluation will be carried out using the standard CoNLL score, which the average of the MUC, B^3 and $CEAF_e$ score and should therefore capture the different aspects these different scores represent in a single metric.

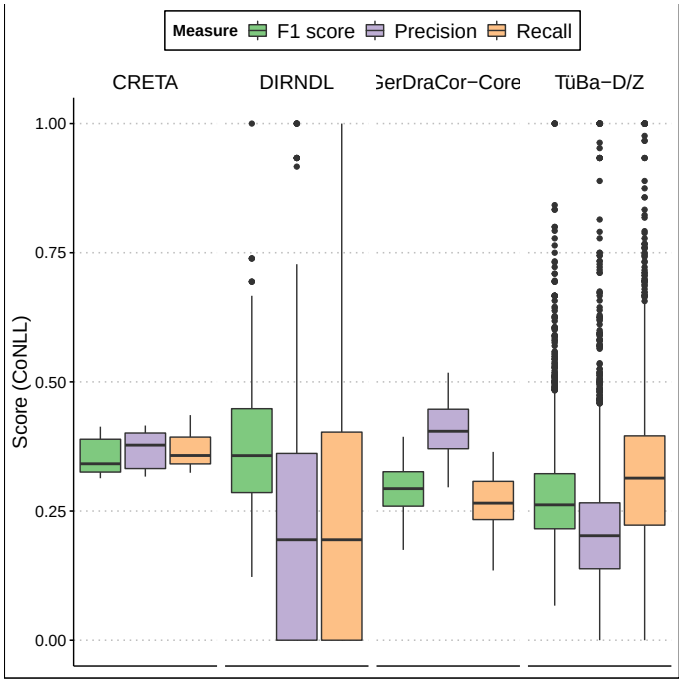


Figure 5.3.: CoNLL scores for applying DramaCoref on different corpora: the CRETA corpus, DIRNDL, TüBa-D/Z and GerDraCor-Coref. Shown are F1 score, precision and recall.

Corpora Figure 5.3 makes a comparison of the performance of DramaCoref on different corpora. This way, it can be evaluated if DramaCoref is fully domain-dependent or if it is capable to generally resolve coreferences. It can be seen that DramaCoref achieves its highest F1 performance on the CRETA corpus, which is a literary corpus. The lower recall of DramaCoref on GerDraCor-Coref compared to CRETA could hint to the fact that the coreferences in the CRETA corpus are more homogeneous when compared to GerDraCor-Coref, as DramaCoref's passes are not able to detect many coreferences in GerDraCor-Coref, but are able to retrieve a larger amount of coreferences in CRETA. The scores for DIRNDL are very spread out, suggesting that DIRNDL is the most heterogeneous when it comes to its documents. Precision is lowest on TüBa-D/Z, which can be explained by TüBa-D/Z being the domain-wise farthest from drama.

For GerDraCor-Coref, DramaCoref achieves the highest precision, which reflects the domain-specificity.

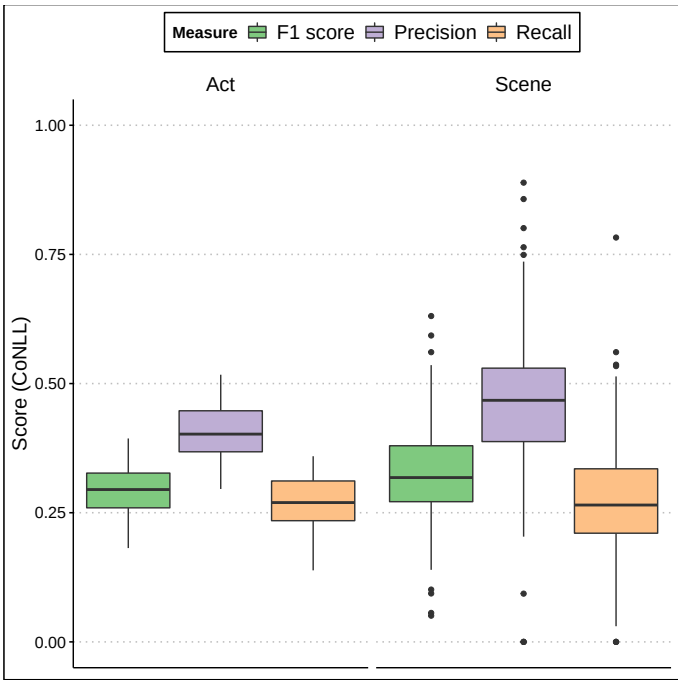


Figure 5.4.: CoNLL F1, precision and recall scores for applying DramaCoref on the acts and scenes of TEST-DRAMACOREF, respectively.

Acts and Scenes Figure 5.4 starts a series of analyses with the goal to further investigate the performance of DramaCoref on GerDraCor-Coref. It shows the difference in performance when applying DramaCoref on full acts or only single scenes of GerDraCor-Coref. It becomes clear that the resolution is easier on scenes than on acts, which makes sense since scenes are shorter and there is less opportunity for the system to lose track of very long chains. On the other hand, the performance on acts is not much lower and the standard deviation is relatively small, which speaks for the robustness of DramaCoref and its ability to handle long chains relatively well. As before, precision is much better than recall.

Post-processing Figure 5.5 shows the performance gain of the post-processing pass. It can be seen that post-processing has only little effect on the performance with a median F1 score of 0.295 for using post-processing vs. a median F1 score of 0.287 for not using post-processing. The amount of character entities which can be correctly classified using

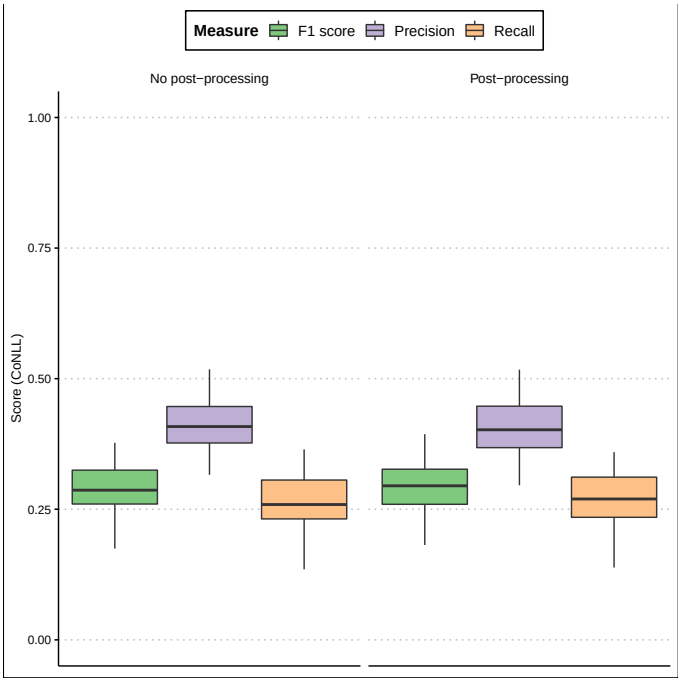


Figure 5.5.: CoNLL F1, precision and recall scores for applying DramaCoref on TEST-DRAMACOREF, one time applying the post-processing pass and one time leaving this pass disabled.

the post-processing pass is probably not large enough to make a large enough impact. Still, it has at least some small benefit for recall performance (+1.1 percentage points).

Gold Mentions Figure 5.6 shows the performance on automatically detected mentions and gold mentions. Gold mentions are the mentions which were annotated by the annotators. In the gold mention setup, the system therefore only needs to correctly predict all coreferences and is not limited by receiving wrongly classified mentions. The gold mention setup therefore represents an upper bound of performance for DramaCoref under the ideal condition that all mentions were predicted correctly. The figure shows that mention detection plays a huge role in the performance of DramaCoref.

Cross Validation Figure 5.7 shows the F1 CoNLL scores for applying DramaCoref on the acts of GerDraCor-Coref in a 10-fold cross validation situation. For this, the documents were put into different sets 10 times so that each drama was used exactly once. This way, it can be investigated if the test split used before had an impact on the classification, for example if by chance many easy or difficult to classify documents happened to land in the test split. The results show that this is not the case, the

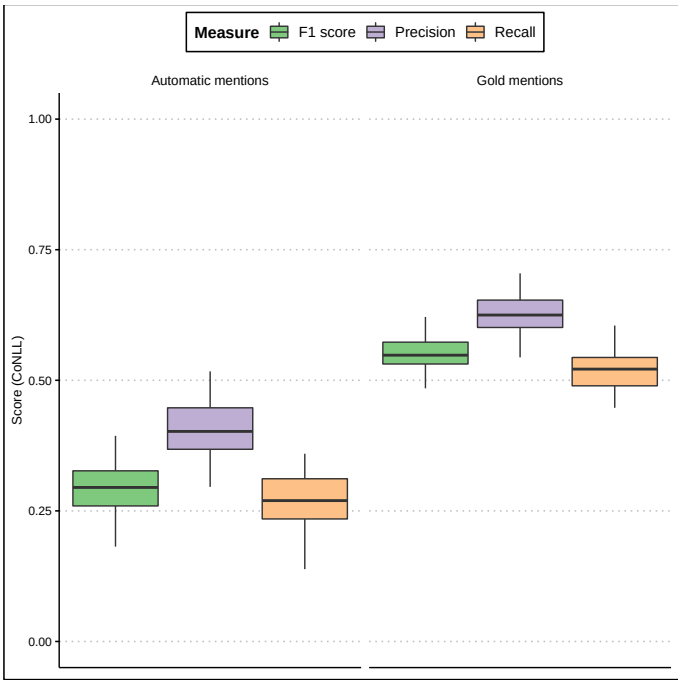


Figure 5.6.: CoNLL F1, precision and recall scores for applying DramaCoref on TEST-DRAMACOREF, one time using gold mentions and one time using automatically determined mentions.

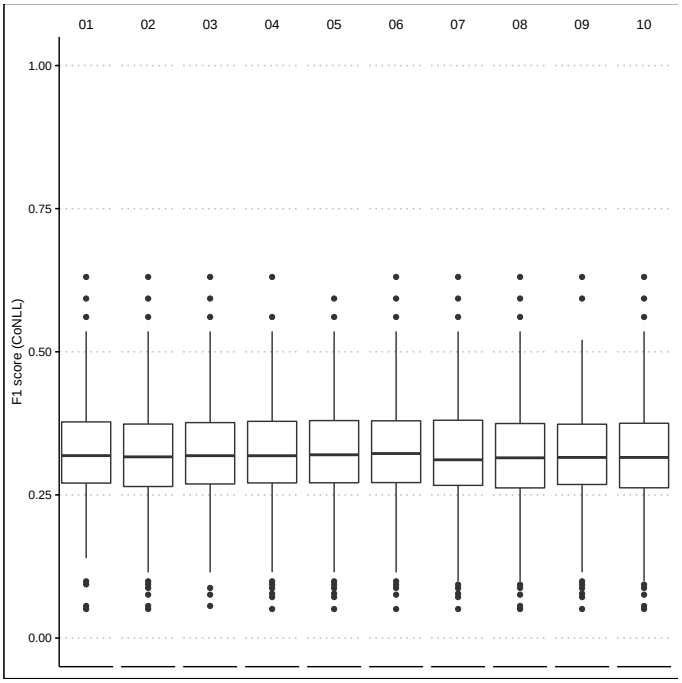
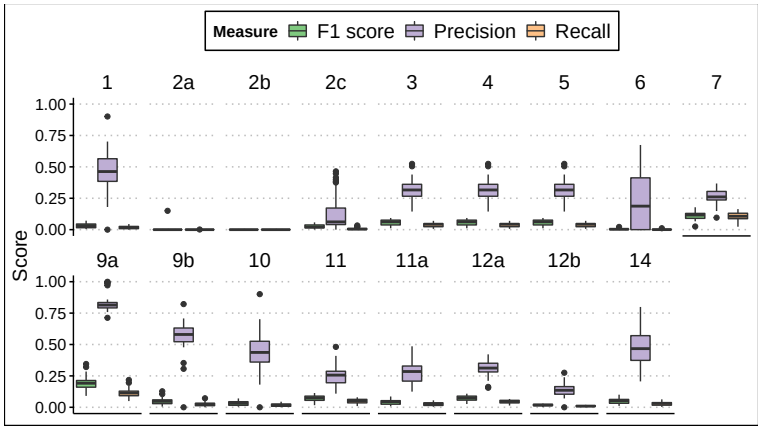
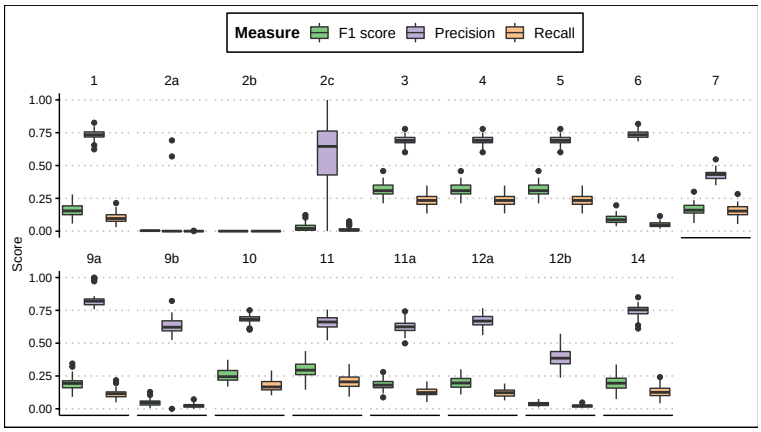


Figure 5.7.: CoNLL F1 scores for applying DramaCoref on GerDraCor-Coref in a 10-fold cross validation setup. Shown are the results for each of the 10 folds.

performance is almost identical for each of the splits. This also means that DramaCoref is relatively robust and able to deliver consistent results for any of the plays.



(a) Automatic mentions



(b) Gold mentions

Figure 5.8.: CoNLL F1, precision and recall scores for applying the single passes of DramaCoref on DEV-DRAMACOREF, with automatic mentions (a) and gold mentions (b).

Pass Performance Figure 5.8a shows the performance of single passes of DramaCoref with automatically generated mentions. It can be seen that most passes display a high precision but often lack in recall. This is to be expected since most passes are designed to be precise and to achieve high precision. On the other hand, passes which are supposed to retrieve a majority of mentions with less regard to precision do not achieve a high recall either (e.g. pass 10). When comparing this to the performance of the passes on

gold mentions (Fig. 5.8b), the recall of all passes is much higher This means that a major culprit in the low recall performance is the fact that the passes often do not have the correct mentions available. However, it is still the case that the precision is always higher than the recall for all passes. The following Section 5.4 attempts to elicit possible reasons for this observation. Best performing pass is pass 9a which retrieves first person personal pronouns. This is not surprising, as first person pronouns are usually covered by speaker tags. Cases in which this pass fails are also examined in the following Section 5.4.

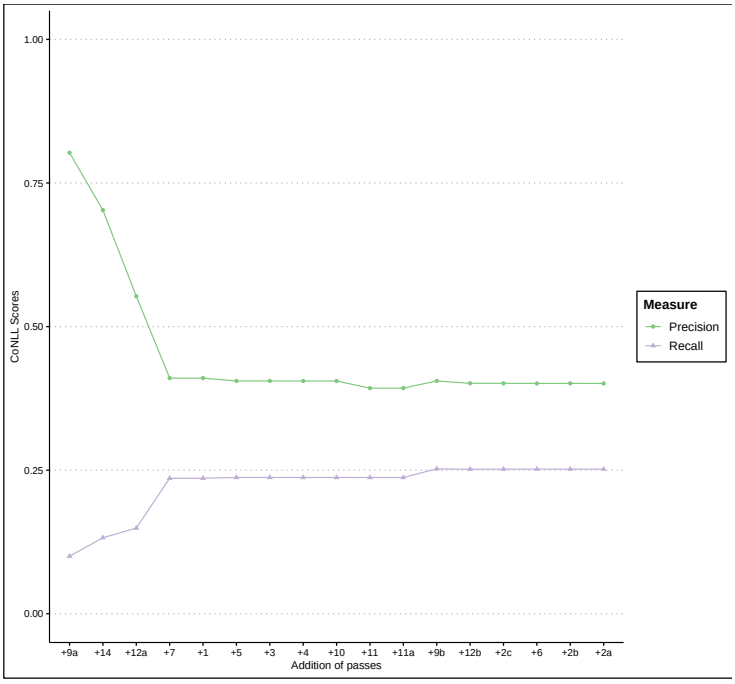


Figure 5.9.: CoNLL precision and recall scores (averaged) for applying the passes of DramaCoref on DEV-DRAMACOREF, one after another, always keeping previously applied passes. Passes are ordered by their precision on DEV-DRAMACOREF.

Cumulative Results Figure 5.9 shows the development of precision and recall on the DEV-DRAMACOREF if passes are added one after another, starting with the pass with the highest precision. It can be seen that passes 9a, 14 and 12a contribute the most to the performance. Afterwards, all other added passes do not alter precision or recall much.
To conclude the results section, the results of other parsers on GerDraCor-Coref are shown, as well as a comparison of other systems on different datasets for coreference resolution experiments with literary data.
Table 5.4 shows the performance for two other CR systems, applied on GerDraCor-Coref: CorZu (Tuggener 2016) and IMS HotCoref DE (Rösiger and Kuhn 2016). CorZu is a

rule-based system operating on mentions coming from dependency parses. IMS HotCoref DE is build upon the English language HotCoref (Björkelund and Kuhn 2014) and utilizes perceptron models.

System	MUC	B^3	$CEAF_e$	CoNLL
CorZu	52.14	17.56	21.26	30.32
IMS HotCoref DE	56.55	14.98	14.84	28.79
DramaCoref	42.54	19.87	18.97	27.12

Table 5.4.: Comparison in performance between DramaCoref, CorZu and IMS HotCoref DE.

The results show that DramaCoref is the best model in terms of its B^3 score, which is a rather conservative score and difficult for most systems to achieve a high performance in. It also outperforms IMS HotCoref DE in the $CEAF_e$ score. However, CorZu and IMS HotCoref DE perform slightly higher on the average CoNLL score. This shows that there still need to be improvements made to the passes of DramaCoref, which the previous results also showcased. On the other hand, none of the three system achieves scores higher than 30% CoNLL, suggesting that the coreferences of the data are difficult to resolve in general.

Results of Related Work Lastly, Table 5.5 gives an overview of the performance of other papers which evaluated coreference resolution on literary data (Krug et al. 2015; van Cranenburgh 2019a), as well as the results of Lee et al. (2011) on which DramaCoref was build upon.

Paper	Mentions	MUC	B^3	$CEAF_e$	CoNLL
Lee et al. (2011)	auto	0.61	0.69	0.45	0.58
	gold	0.65	0.71	0.48	0.61
Krug et al. (2015)	auto	0.86	0.56	NA	NA
	gold	NA	NA	NA	NA
van Cranenburgh (2019a)	auto	0.71	0.62	0.67	0.67
	gold	0.78	0.72	0.78	0.76

Table 5.5.: Comparison of the results on coreference resolution of the papers Lee et al. (2011), Krug et al. (2015), and van Cranenburgh (2019a).

The results of the different papers are not directly comparable, since they each used a

different system and dataset, but it can be seen that usually the difference in performance between auto and gold mentions is not that large, which is different for DramaCoref on GerDraCor-Coref. Otherwise, the performance on gold mentions for DramaCoref on GerDraCor-Coref (avg. 0.55 F1 CoNLL score) is comparable to the performance of the other systems for gold mentions.

5.4. Error analysis for DramaCoref

In the following, an error analysis is performed for a number of selected passes, using examples from GerDraCor-Coref.

Pass 9a loses out on precision mostly because of letters which are read out loud.

VRONI.
's Siegel is eh schon ganz verbröckelt, [ich]$_1$ mach 'n auf!

JAKOB.
Tu's, is jetzt dein' Sach'!

VRONI *öffnet den Brief.*
Er is vom Vater sein'm Bruder, vom Kreuzweghofbauer! — Heiliger Gott!

JAKOB.
Du verschreckst ein'n!

VRONI.
Um Gottes will'n, Bruder, los zu, los nur zu, was er 'm Vater g'schrieb'n hat: »Lieber Jakob! Dein Testament, worin Du die Burger Vroni und ihre zwei Kinder als Erben von all Dein Hab und Gut einsetzt, hab [ich]$_1$ erhalten. Es ist nit schön, daß Du [mich]$_1$ und [meine]$_1$ Kinder so g'ring drein abfertigst ...«

Figure 5.10.: Snippet from TextGrid Repository (2011): *Anzengruber, Ludwig. Der Meineidbauer*, extended with markup showing coreference mistakes made by pass 9a. For a translation, see Figure C.11

In this example, the first person personal pronouns *ich* (*I*), *mich* (*me*) and *meine* (*my*) in Vroni's last utterance actually refer to the "Kreuzweghofbauer", but since Vroni is

reading the letter out loud, pass 9a will assign the pronouns as referring to Vroni. This example is also a case for a false positive for the pass 9b, since the letter also uses second person pronouns (*Du*, *you*; *Dein*, *your*) which refer to the father of Vroni and Jakob, but are assigned to refer to Jakob by pass 9b, since he is in a conversation with Vroni at this moment.

Pass 9b works generally well when there are only two people involved in a conversation; however, there are some places with more subtle second person pronoun addressing where it fails:

EMILIA.
Es ist wahr, mit einer Haarnadel soll ich —*Sie fährt mit der Hand nach dem Haare, eine zu suchen, und bekömmt die Rose zu fassen.* [Du]$_1$ noch hier? — Herunter mit [dir]$_1$! [Du]$_1$ gehörest nicht in das Haar einer, — wie mein Vater will, daß ich werden soll!

ODOARDO.
O, meine Tochter! —

EMILIA.
O, mein Vater, wenn ich [Sie]$_1$ erriete! — Doch nein; das wollen [Sie]$_1$ auch nicht. Warum zauderten [Sie]$_1$ sonst?

Figure 5.11.: Snippet from TextGrid Repository (2012d): *Lessing, Gotthold Ephraim. Emilia Galotti*, extended with markup showing coreference mistakes made by pass 9b. For a translation, see Figure C.12

In this example, Emilia is in a dialogue with her father Odoardo. For the majority of the conversation, second person pronouns are resolved correctly, however, at this point in the conversation, Emilia talks to a hairpin while taking it out of her hair and addresses it with *du* and *dir* (eng. *you*). These pronouns are then falsely made coreferent with the formal pronoun *Sie* which address her father.

Pass 14 mostly suffers from long distances between entities and scenery changes, which leads to previously coreferent lemmas changing their referent.

Erster Akt
Erste Szene

[...]

FRANZ.
Die Post ist angekommen – [ein Brief von unserm Korrespondenten in Leipzig]$_1$ –

[...]

FRANZ *nimmt [den Brief]$_1$ aus der Tasche.*

[...]

Erster Akt
Zweite Szene

[...]

MOOR *fliegt ihm entgegen.*
Bruder! Bruder! [den Brief]$_1$! [den Brief]$_1$!

Figure 5.12.: Snippet from TextGrid Repository (2012g): *Schiller, Friedrich. Die Räuber.* extended with markup showing coreference mistakes made by pass 14. For a translation, see Figure C.13

This example shows how in two adjacent scenes in the same act, the mention *den Brief* (eng. *the letter*) refers to two different letters. The first letter in the first scene of the first act is a letter by Karl to his father, the second letter in the second scene of the first act is a letter of Karl's brother Franz to Karl. However, pass 14 falsely assigns both letters to the same coreference cluster. This mistake is of course only committed in the act-wide setup, but not when applying DramaCoref on single scenes.

Pass 1 suffers from similar mistakes.

PHILIPP.
Ey! mit Ihrer Erlaubniß, [gnädiger Herr]$_1$! das kann nicht seyn.

HERR ORGON.
Das kann nicht seyn! Und warum?

```
[...]

DAMON.
Ich kann unmöglich länger bleiben, ich würde mich zu sehr verrathen.
    Himmel! wie reizend ist sie nicht!

Er will abgehen.
LISETTE.
Pst! Pst! [gnädiger Herr]_1, wo gehen Sie hin?
```

Figure 5.13.: Snippet from TextGrid Repository (2012a): *Cronegk, Johann Friedrich von. Der Mißtrauische*, extended with markup showing coreference mistakes made by pass 1. For a translation, see Figure C.14

The phrase *gnädiger Herr* (eng. my lord) is used by Philipp to address Orgon and later by Lisette to address Damon. However, pass 1 assigns both to the same coreference cluster.

Pass 12a is sometimes too general, since the synsets in GermaNet cover a wide range of possible synonyms. For example in

```
MELLEFONT.
Du störest mich, Norton!

NORTON.
Verzeihen Sie also [mein Herr]_1 — \textit{Indem er wieder zurück gehen
    will.}

[...]

NORTON.
Könnte Sie wohl besorgt, aber nicht niedergeschlagen machen. — Sie
    beunruhiget etwas anders. Und ich will mich gern geirret haben, wenn
    Sie es nicht lieber gesehen hätten, [der Vater]_1 wäre noch nicht
    versöhnt. Die Aussicht in einen Stand, der sich so wenig zu Ihrer
    Denkungsart schickt — —
```

Figure 5.14.: Snippet from TextGrid Repository (2012e): *Lessing, Gotthold Ephraim. Miß Sara Sampson*, extended with markup showing coreference mistakes made by pass 12a. For a translation, see Figure C.15

the phrases *mein Herr* (eng. Sir, Lord) and *der Vater* (eng. the father) are made coreferent, probably because of synonymy in religious contexts.

Pass 12b has similar problems with generality, since often phrases which share a common hyperonym are not coreferent, like in

1 EMILIA.
2 Es ist wahr, mit einer Haarnadel soll ich —*Sie fährt mit der Hand nach dem Haare, eine zu suchen, und bekömmt die Rose zu fassen.* Du noch hier? — Herunter mit dir! Du gehörest nicht in das Haar einer, — wie [mein Vater]$_1$ will, daß ich werden soll!
3
4 ODOARDO.
5 O, [meine Tochter]$_1$! —

Figure 5.15.: Snippet from TextGrid Repository (2012d): *Lessing, Gotthold Ephraim. Emilia Galotti*, extended with markup showing coreference mistakes made by pass 12b. For a translation, see Figure C.16

where *mein Vater* (eng. my father) and *meine Tochter* (eng. my daughter) are made coreferent, since they both have the common hyperonym of family. A similar case occurs at the end of the play, where *Gott* (eng. God) and *Teufel* (eng. devil) are marked as coreferent by pass 12b:

1 DER PRINZ
2 *nach einigem Stillschweigen, unter welchem er den Körper mit Entsetzen und Verzweiflung betrachtet, zu Marinelli.*
3
4 Hier! heb' ihn auf. — Nun? Du bedenkst dich? — Elender! — *Indem er ihn den Dolch aus der Hand reißt.* Nein, dein Blut soll mit diesem Blute sich nicht mischen. — Geh, dich auf ewig zu verbergen! — Geh! sag' ich. — [Gott]$_1$! [Gott]$_1$! — Ist es, zum Unglücke so mancher, nicht genug, daß Fürsten Menschen sind: müssen sich auch noch [Teufel]$_1$ in ihren Freund verstellen?

Figure 5.16.: Snippet from TextGrid Repository (2012d): *Lessing, Gotthold Ephraim. Emilia Galotti*, extended with markup showing coreference mistakes made by pass 12b. For a translation, see Figure C.17

Furthermore, one could argue that the exclamation *Gott!* is not referring here if read as an idiomatic expression.
Many of these examples show that a main pitfall for DramaCoref passes applied on GerDraCor-Coref is long distances between entities, which also often includes a change in scenery and context. As the analyses in Chapter 4 demonstrated, long distance mentions and long spanning coreference chains are very prevalent in dramatic texts compared to other domains. This also means that many passes developed for shorter texts will generalize too much and cannot consider the necessary context. Interestingly, Krug et al. (2015) report no such observations for resolving coreferences on novels, which could mean that plays feature a much higher frequency of scenery changes on average or the type of entities with similar lemmas is higher in plays compared to novels.

5.5. Summary

This chapter presented experiments for coreference resolution on dramatic texts. For mention detection, the performance of two constituency parsers was tested against the performance of different transformer models and the Stanford parser achieved overall the best results. Hence, its output was chosen as being used as the input for the CR system. The rule-based CR system DramaCoref was presented and evaluated on the corpus GerDraCor-Coref, as well as on corpora from other domains. The system achieved an average F1 CoNLL score of 0.31 for the test set of GerDraCor-Coref. The results furthermore showed that DramaCoref is generally able to resolve coreferences on theatre plays in cases in which other systems have issues. When applying DramaCoref on genres other than dramatic texts, it achieved its highest precision on GerDraCor-Coref, but was still able to perform comparably on the other corpora. Applying DramaCoref on single scenes as compared to whole acts achieves higher results. Using a post-processing filter that merged character-based coreference chains did not show a large effect. The results also showed that dramatic texts present a couple of unique challenges with regard to CR. In particular, the system suffers from misclassifying mentions with long distances between them, as well as cases where first person pronouns do not refer to the character speaking, for instance when a letter is read out loud.
The neural mention detection did not work as well as the coreference resolution part,

hence the output of the Stanford constituent parser was used in order to feed DramaCoref with mentions. However, it also became apparent that the mention detection was prone to mistakes in general and DramaCoref could perform much better on gold mentions, on which it also outclassed many other systems.

[M]an muß erst eine Weile unter den Menschen gelebt haben um Charaktere beurteilen zu können. Der Herr Pätus, oder wie er da heißt, hat sich Ihnen bisher immer nur unter der Maske gezeigt; jetzt kommt sein wahres Gesicht erst ans Tageslicht: [. . .]
([O]ne must first have lived among people for a while to be able to judge characters. Mr. Pätus, or whatever his name is, has only ever shown himself to you under a mask; only now is his true face coming to light: [. . .])

Hofmeister in Jakob Michael Reinhold Lenz's "Der Hofmeister oder Vorteile der Privaterziehung"

6 Character Type Detection

Protagonist detection and detecting other forms of character types has been a productive field of research in CLS in the past years (Bamman, Underwood, and Smith 2014; Jannidis et al. 2016; Algee-Hewitt 2017; Fischer et al. 2018; Reiter et al. 2018; Krautter et al. 2018; Jahan et al. 2020; Krautter et al. 2020). This chapter describes ML approaches to identify title characters (Section 6.2), protagonists (Section 6.3) and schemers (Section 6.4) in dramatic texts, as well as methods to interpret the results offered by the ML models, so that literary scholars are able to work with the output and draw conclusions for the literary works in question.

Section 6.1 gives a general overview of the necessary operationalization when annotating and automatically detecting character types and describes how this operationalization was carried out for the three types discussed in this chapter. While the three character types are just examples of many possible types, they were chosen in order to show results for character types of different nature and complexity. While title characters are defined purely structurally and as part of the paratext of a play, protagonists are based on the plot and content of a play and thus more complex in nature. The third type, schemers, are a rather specific, literary-studies driven character type and arguably the most complex and layered of the three types. It is worth pointing out that compared to the previous chapter, which dealt with ML for linguistic categories and a well established field in CL, namely coreference resolution, this chapter covers the detection of literary categories, with a slightly different goal: Detecting character types and uncovering which features were helpful in the classification and can inform further investigations and interpretations

of characters of literary works and be useful as input for downstream tasks such as automatic modelling of narration and plot in literary texts (see for example Jahan, Mittal, and Finlayson 2021).
Section 6.2 describes results reported in Krautter and Pagel (2019), Section 6.3 is based on the work described in Krautter et al. (2018) and Section 6.4 negotiates the findings of Krautter and Pagel (2024, to appear).

6.1. Operationalization of Character Types

When attempting to automatically classify characters in dramatic texts according to their type, several immediate questions may arise:

1. What are possible character types?
2. How to identify which character belongs to which type?
3. How to implement models to automatically assign types to characters?

The answer to the first question will be answered in this chapter in a relatively straightforward way: any type that is interesting to literary studies and that has been discussed in the context of literary studies discourse or any type that is interesting from a modelling point of view. From the many possible types arising from these constraints, three have been chosen for investigation. Title characters, since they are interesting from a structural point of view and easy to operationalize in terms of annotating them, and protagonists and schemers since they are character types frequently discussed in the literary studies' research discourse and have varying degrees of broadness with protagonists being a very widely applicable type and schemers being a more narrow and specialized type. The other two questions concern the topic of *operationalization* which has been touched upon above. In this chapter, I will understand operationalization in the sense that it was used in Pichler and Reiter (2021), namely as "the development of a method for tracing a (theoretical) term back to text-surface phenomena" (Pichler and Reiter (2021, p. 1)).
Therefore, I will use the term *operationalization* as concretizing a theoretical literary concept, for example *protagonist*, by developing features which allow to identify specimen of the group of protagonists in a text.

As outlined above, two aspects of operationalization need to be addressed: (i) operationalizing the concept so that human annotators can assign types to characters and (ii) operationalizing the concept so that ML models can assign types to characters.
The first aspect will be individual to the concept in question and most likely make use

of theoretical deliberations from literary studies. The way in which the three types considered in this chapter are operationalized for human annotators is described in their respective sections (6.2-6.4).
The second aspect is also done on an individual basis, but there are some general considerations that can be made. In general, in order to allow an ML algorithm to perform mathematical operations on the chosen features, the features need to be either numerical or values that can be easily translated into numbers, like boolean values. Secondly, characters can be categorized by a multitude of textual features from which the algorithm can then choose which were most helpful in identifying the specific character type. Choosing a multi-dimensional approach in which the feature set draws from many different aspects of textual properties of characters is beneficial, since one can form strong hypotheses about which features might be helpful for an ML algorithm, but cannot be certain. Furthermore, if character types are multi-faceted, a multi-dimensional approach will handle this multi-facetteness most appropriately.
Figure 6.1 shows a tree representation of a possible set of dimensions and features for operationalizing schemers, taken from Krautter and Pagel (2024, to appear). The hierarchy has been originally developed in Krautter et al. (2020) for a variety of character types and can therefore be considered to be more general than just applying to schemers. It shows six different dimensions to consider when developing features for automatic character type detection: *character speech style*, *sentiment*, *aboutness*, *interaction*, *stage presence* and *action*. For each of these six dimensions, there can be several sub-dimensions, for instance, stage presence can either be active or passive and action can be defined via verbs that occur either in the stage direction or in the character speech. The dimensions follow general intuitions about what literary scholars might find relevant for characterizing literary characters, but can naturally not be complete or authoritative.
The concrete implementation of features for each of these dimensions, as well as intuitions for why a certain feature might capture a certain aspect of a character type, is described in the respective section (6.2-6.4).

6.2. Title Character Detection

Title characters or eponymous characters are characters whose name is included in the title of a drama. In English, plays by Shakespeare whose title is made up of a character occurring in it include *Romeo and Juliet*, *Macbeth*, *Hamlet*, *King Lear*, *Richard III*, *The Two Gentleman of Verona*, *Julius Caesar* or *Antony and Cleopatra*. From this selection,

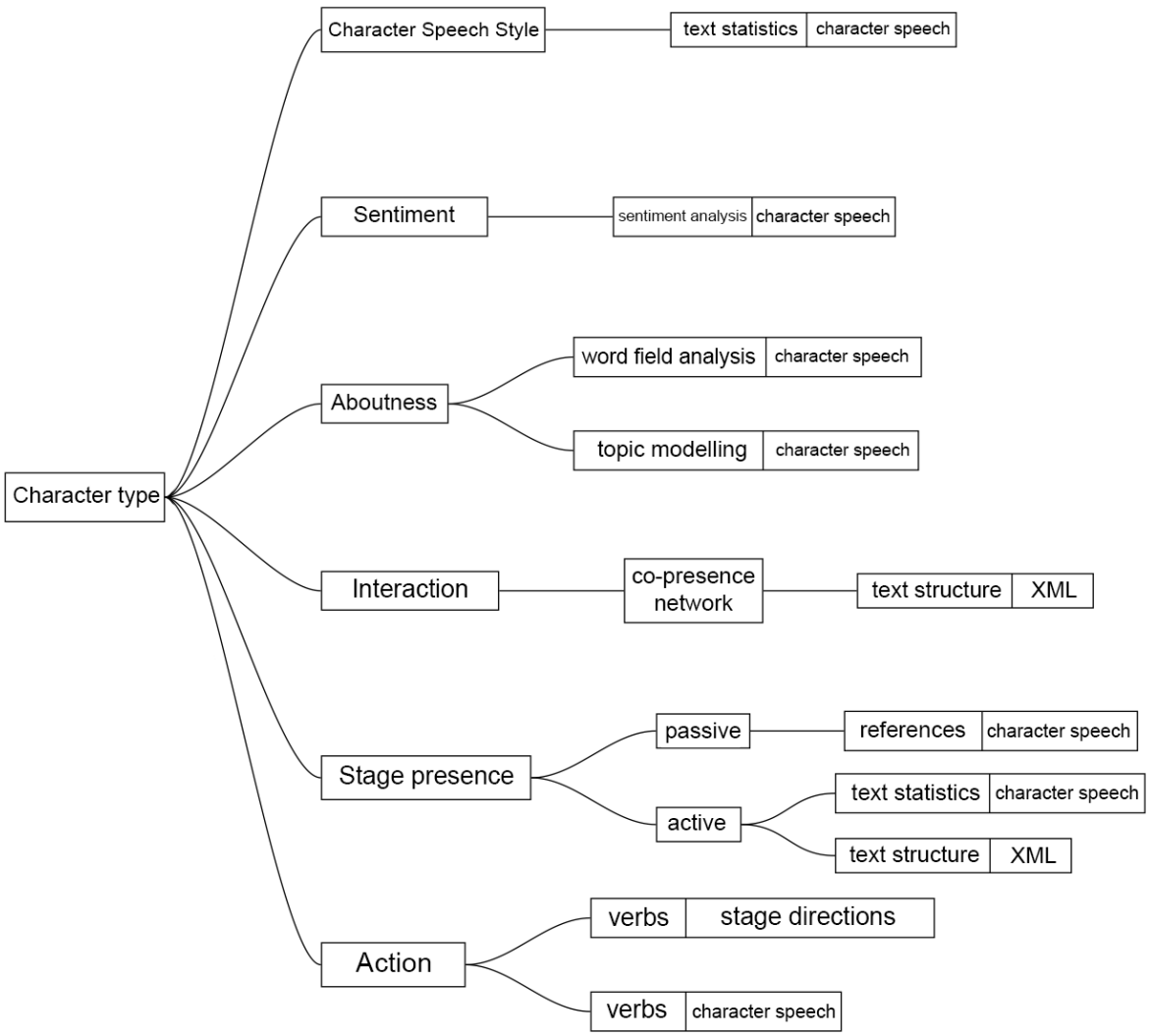


Figure 6.1.: Operationalization hierarchy for detecting schemers from Krautter and Pagel (2024, to appear).

several observations can be made:

1. There can be more than one title character (*Romeo and Juliet*, *Antony and Cleopatra*)
2. Characters can be addressed with several parts of their name, e.g. first name (*Romeo and Juliet*, *Antony and Cleopatra*), last name (*Macbeth*), both (*Julius Caesar*), with or without title (*King Lear*, *Richard III*, but *Macbeth* instead of *Lord Macbeth*, *Hamlet* instead of *Prince Hamlet*).
3. The characters may not be mentioned by name at all (*The Two Gentleman of Verona*)

Title characters will often also be protagonists of a play, however, this does not necessarily need to be the case. One example might be *Emilia Galotti* by Lessing, where one could argue that Emilia plays only a minor active role in the play and is rather subject of intrigue and opportunism of the other characters of the play and therefore not a real protagonist. Furthermore, characters that would be considered to be protagonists are often not title characters, even though they may share many similarities with the title characters who are protagonists.

6.2.1. Annotation and Data

For the data used in the experiments, the dataset by Krautter and Pagel (2019) is used. In Krautter and Pagel (2019), 42 characters from 38 plays were categorized as eponymous characters by the first author. This simply entailed to manually identify characters whose name occurred in the title of a play. In contrast stood 1166 characters that were not eponymous. This makes a total of 1208 characters.

6.2.2. Experimental Setup

The following experiments are labeled as TITLECHARACTER.
For the annotated plays, several features are extracted automatically using custom R[1] scripts and which can be classified into the following major groups:

- Textual
- Character networks
- Utterance content
- Stage presence
- Metadata

[1] `https://www.r-project.org/`

Each group contains several features, which are explained below.

Feature Group	Feature Name	Short Name	Value Range
Textual	Tokens	tokens	$[0,1] \in \mathbb{Q}$
Characters networks	Degree	degree	$[0,1] \in \mathbb{Q}$
	Weighted Degree	wdegree	$[0,+\infty) \in \mathbb{N}$
	Closeness	close	$[0,1] \in \mathbb{Q}$
	Betweenness	between	$[0,1] \in \mathbb{Q}$
	Eigenvector	eigen	$[0,1] \in \mathbb{Q}$
Speech content	Topics	T1-T10	$[0,1] \in \mathbb{Q}$
Stage presence	Active presence	actives	$[0,1] \in \mathbb{Q}$
	Passive presence	passives	$[0,1] \in \mathbb{Q}$
	In final act?	lastAct	$\{0,1\}$, Boolean
Metadata	Epoch, Genre	SD, BT, WK, POP, NAT, WM, ROM, AUF, VM	$\{0,1\}$, Boolean
	Number of characters	nfig	$[1,+\infty) \in \mathbb{N}$

Table 6.1.: Features used in experiment TITLECHARACTER. Given are the broader feature groups, the single features associated to a single group and the possible values a feature can take.

Table 6.1 gives an overview of the feature groups and the features associated with them.

Tokens is the number of tokens a character utters throughout the play. This value is normalized by the total number of tokens of the whole play:

$$\frac{\text{Number of tokens character utters}}{\text{Number of tokens in play}} \tag{6.1}$$

Degree is a centrality measure. Centrality quantifies a certain property of a given character network; in the case of degree the number of edges a node has with other nodes in the network. The character networks are created by first creating a square matrix of characters. Each cell in the matrix contains the number of scenes the character in the row and column appear together in stage. From this matrix, a network can be directly generated by representing each character as a node and drawing edges between characters do appear together.[2] The number of co-occurrences is then set as the weight of the edge.

[2]In fact, the co-occurrence matrix and the character network can be transformed into each other and are simply different representations of the same information.

Figure 6.2 gives an example of the co-occurrence matrix and character network for the play *Miß Sara Sampson* by Lessing. In the matrix (Fig. 6.2b), each character of the play occurs once in a row and once in a column, so every character is paired with each other character and with themselves. The diagonal always contains the pairings of characters with themselves and thus represents the number of scenes a character occurs in in total. The cells between different characters contain the number of scenes that the respective characters co-occur on stage. The matrix is mirrored on the diagonal, so each pairing occurs twice. The character network (Fig. 6.2a) can be derived from one half the matrix, leaving the diagonal and the other half of the matrix out. The weights on the edges correspond to the values in the co-occurrence matrix. Additionally, the thickness of an edge corresponds to its weight, so the edges with high weights can be easily identified visually. Degree corresponds to the number of edges that connect with a node. Degree can also be normalized by dividing the number of edges of a node by the total number of nodes minus one[3] in the network. This way, networks of different sizes can be compared with each other. In the example of Figure 6.2, the character Sara has a degree of 6 and a normalized degree of 0.6 ($\frac{6}{11-1}$) and the character Sir William a degree of 5 and a normalized degree of 0.5 ($\frac{5}{11-1}$). For all experiments, the normalized version of degree is used.

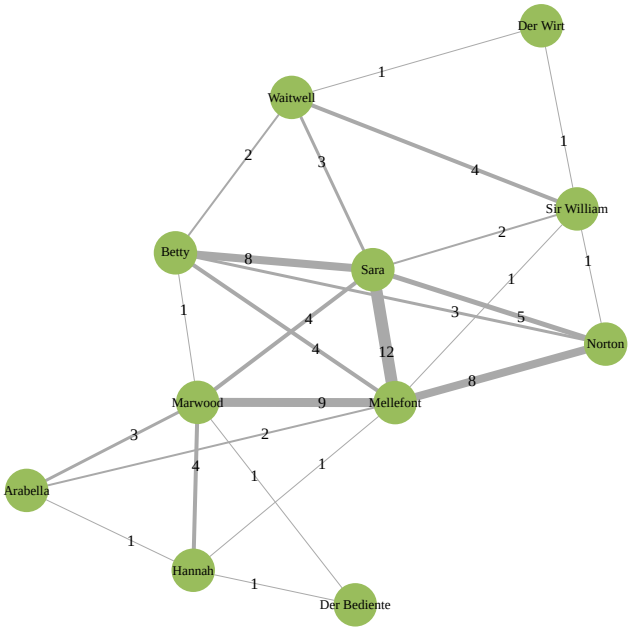
Weighted degree is the same as degree, except that instead of counting each edge once, weighted degree is the sum of all weights of all edges connected to a node. For Figure 6.2, the weighted degree for Sara is 34 ($12+4+8+3+2+5$) and the weighted degree for Sir William is 9 ($1+1+1+2+4$). Weighted degree cannot easily be normalized, hence weighted degree as a feature is used as is.

Closeness centrality (Beauchamp 1965) measures the average distance of a node to all other nodes in the network by taking the inverse of the mean of all shortest paths from this node to all other nodes of the network:

$$C(v_k) = \frac{1}{\sum_{i=1}^{n} d(v_i, v_k)}, \tag{6.2}$$

where d is a distance function returning the number of shortest paths of two nodes and v_k is the node for which closeness centrality should be calculated (Freeman 1978/1979, p. 225). Intuitively, the higher the closeness of a node, the easier (shorter) it is to reach

[3]The total number of nodes is subtracted with one so that the node for which the degree is calculated is taken out of the calculation.



(a) Character network

	Sir William	Waitwell	Der Wirt	Mellefont	Norton	Betty	Sara	Marwood	Hannah	Der Bediente	Arabella
Sir William	7	4	1	1	1	0	2	0	0	0	0
Waitwell	4	7	1	0	0	2	3	0	0	0	0
Der Wirt	1	1	1	0	0	0	0	0	0	0	0
Mellefont	1	0	0	24	8	4	12	9	1	0	2
Norton	1	0	0	8	11	3	5	0	0	0	0
Betty	0	2	0	4	3	9	8	1	0	0	0
Sara	2	3	0	12	5	8	22	4	0	0	0
Marwood	0	0	0	9	0	1	4	15	4	1	3
Hannah	0	0	0	1	0	0	0	4	4	1	1
Der Bediente	0	0	0	0	0	0	0	1	1	1	0
Arabella	0	0	0	2	0	0	0	3	1	0	3

(b) Co-occurrence matrix

Figure 6.2.: Character network (a) and co-occurrence matrix (b) for Lessing's play *Miß Sara Sampson*. Each node represents a character and each edge represent a scenic co-occurrence of two characters. The network can be derived from the co-occurrence matrix.

all other nodes from this node and the node can be said to be more central in the network. Closeness centrality can also be normalized by multiplying the result of Equation 6.2 by the number of nodes in the network minus one, or expressed as a division similar to Equation 6.2:

$$C_{\text{norm}}(v_k) = \frac{N-1}{\sum_{i=1}^{n} d(v_i, v_k)} = C(v_k) \times (N-1), \tag{6.3}$$

where N is the number of nodes in the network (Freeman 1978/1979, p. 226). It is also possible to calculate a variant of closeness centrality where the weights of edges are interpreted as distances between nodes in order to factor weights into the final result.
In Figure 6.2, the closeness centrality for Sara is 0.071 ($\frac{1}{14}$, norm.: 0.71) and the weighted closeness centrality is 0.027 ($\frac{1}{37}$, norm.: 0.27). The closeness centrality for Sir William is 0.0625 ($\frac{1}{16}$, norm.: 0.625) and the weighted closeness centrality is 0.043 ($\frac{1}{23}$, norm.: 0.43). For the experiments, weighted and normalized closeness centrality is always used.

Betweenness centrality (Freeman 1977) measures how often a node lies on the shortest path of two other nodes. This measure models how often a node is a "messenger", i.e. how many other nodes in the network it connects with each other. It can be computed by dividing the number of shortest paths through three nodes v_k, v_i and v_j by the number of shortest paths through the nodes v_i and v_j and summing this up for all possible combinations of nodes v_i and v_j. v_k is the node for which betweenness centrality should be computed and v_i and v_j are two other nodes in the network.

$$B(v_k) = \sum_{i=1, j=1, i\neq j, i\neq k, j\neq k}^{N} \frac{s_{v_i v_k v j}}{s_{v_i v_j}} \tag{6.4}$$

$s_{v_i v_j}$ denotes the shortest path between the two nodes v_i and v_j and $s_{v_i v_k v_j}$ denotes the shortest path between nodes v_i and v_j with node v_k lying on this shortest path. If $s_{v_i v_k v_j}$ is the only existing shortest path between v_i and v_j, $B(v_k)$ increases by one, otherwise it increases by the ratio of the number of shortest paths including v_k and the overall number of shortest paths of v_i and v_j (see also Freeman 1977, p. 37).
Betweenness centrality can also be normalized given the unnormalized betweenness B via the formula

$$B_{\text{norm}} = \frac{2B}{(N-1)(N-2)}, \tag{6.5}$$

where N is the number of nodes in the graph. Like for closeness centrality, it is possible

to interpret weights as distances when counting shortest paths, resulting in a weighted variant of betweenness centrality.
For Figure 6.2, the betweenness centrality for Sara is 5.03 (norm.: 0.11) and the weighted betweenness centrality is 0 as well as the normalized betweenness centrality. For Sir William, betweenness centrality is 6.66 (norm.: 0.15) and the weighted betweenness centrality is 18.5 (norm.: 0.41). Once again, weighted and normalized betweenness is used for all experiments.

Eigenvector centrality is a measure similar to degree, but is higher for nodes which are themselves connected to nodes which are connected to many other nodes. It does so by utilizing the eigenvector of the co-presence matrix on which the network is based (see Newman 2010). When calculating weighted eigenvector centrality, the weights of nodes are used as a measure of the strength (like for weighted degree), instead of just counting if an edge exists between two nodes (like for degree).
In Figure 6.2, the eigenvector centrality for Sara is 0.924 and the weighted eigenvector centrality is 0.942. For Sir William, eigenvector centrality is 0.69 and the weighted eigenvector centrality is 0.17. For all experiments, weighted eigenvector centrality is used.

Topics are ten topics (*T1-T10*) from a topic model which was trained on GerDraCor-Coref using LDA (Blei, Ng, and Jordan 2003). For each character the posterior probability for each topic can be calculated by taking all utterances of a character and calculating the probability of a certain topic being present in the utterances. A higher posterior probability means that a character is more likely to use words present in a certain topic, or in other words, how likely a character is to talk about a certain topic. Doing this results in ten feature values per character, each being the posterior probability for one of the topics *T1–10*.

Active presence and passive presence are measures for the presence of characters on the stage. Active presence is straightforwardly the number of times a character c appears on stage (s_c) divided by the total number of scenes in the play (S).

$$\text{Active presence} = \frac{s_c}{S} \tag{6.6}$$

The division normalizes the result and ensures that comparisons between plays with a differing number of scenes are possible. Passive presence assumes that characters can also have a presence on stage when they are not physically present, but are mentioned

by other characters. Consequently, passive presence (Willand et al. 2020) is defined as the number of times a character c is *not* present on stage but mentioned by another character (m_c), divided by the number of scenes S.

$$\text{Passive presence} = \frac{m_c}{S} \tag{6.7}$$

As before for active presence, the division by the total number of scenes ensures that plays of different lengths can be properly compared with each other.

If a character appears in the final act is a boolean value that is 0 when the character does not appear in the final act and 1 if they do. This features encodes the idea that the final act of a play contains the resolution of the final conflict and that important characters will most likely be part of this resolution and present.

Epoch and genre are also represented as boolean values and are 0 when the play a character appears in is not in a certain epoch or genre and 1 if they are. Note that both features *epoch* and *genre* are used since it is often not possible to distinguish between the two and oftentimes genres denote a certain period of time and vice versa.

The total number of characters is given as a normalizing factor for the machine learning model, so that it is possible to distinguish values for characters in large plays to those of characters in small plays.

For the ML algorithm, *Random Forest* (RF) has been used (Ho 1995; Ho 1998; Breiman 2001). During training, the algorithm takes in all values for all characters and all features described above, as well as the *true* class value for each character, i.e. if the character is a protagonist or not, coming from the annotations. This way, the algorithm learns a mapping between the quantitative representation of characters via the features and its role in the play as protagonist or not-protagonist. Random forest does this by creating decision trees for all possible feature combinations and finding an optimal ensemble of decision trees via regression. This ensemble of decision trees is then able to decide, based on a given set of features, if the character represented by these features would be most likely a protagonist or not.
The RF model is trained on 80% of the characters and its performance evaluated on the remaining 20% of characters. During training, 10-fold cross validation is applied and the best performing model of the ten runs is chosen as the model to be applied on the

test data. The training data was also sampled using the SMOTE algorithm (Chawla et al. 2002), which performs up-sampling based on some statistical properties of the existing data. Sampling is beneficial, since the number of instances for the positive class, i.e. the protagonist class, is rather small when compared to the number of instances for non-protagonist characters. Upsampling ensures that there is a comparable amount of instances for both classes.

6.2.3. Results

	Title character			Not-Title-Character		
	Precision	Recall	F1	Precision	Recall	F1
Majority BL	-	0.00	-	0.97	1.00	0.98
Tokens BL	0.28	0.88	0.42	1.00	0.93	0.96
Without Epochs/Genre	0.32	0.89	0.47	1.00	0.94	0.97
Without tokens	0.24	0.86	0.38	1.00	0.92	0.96
All features	0.32	0.89	0.47	1.00	0.94	0.97

Table 6.2.: Results for the random forest model on predicting title characters.

Table 6.2 shows the results for classifying the title characters. The table consists of the results for a majority baseline, a model only using the tokens feature, a model using all features *but* the *tokens* feature and a full model with all features. Recall is consistently high for all models, but precision does not go over 35%, ranging from 24 to 32%. The table shows that the full model is the best performing model. The performance of the *tokens* baseline model and the model not using the *tokens* feature is almost identical. This means that all features combined contribute as much as the *tokens* feature alone, but using all features together improves the performance further. Therefore, the features seem to complement each other and different features cover aspects of being a title character that other features do not. We observe a significant drop in performance of 9 percentage points with respect to the F1 score (from 47% for the full model to 38% for the model without *tokens*). An evaluation of a model not using any epoch or genre-based features shows that there is no difference in performance, suggesting that this information is not very helpful or relevant for classification.
Figure 6.3 shows the feature importance for TITLECHARACTER. Feature importance for RFs can be calculated by dropping one feature for the classification and measuring the difference in performance (Breiman 2001, pp. 23–25). The feature that leads to the

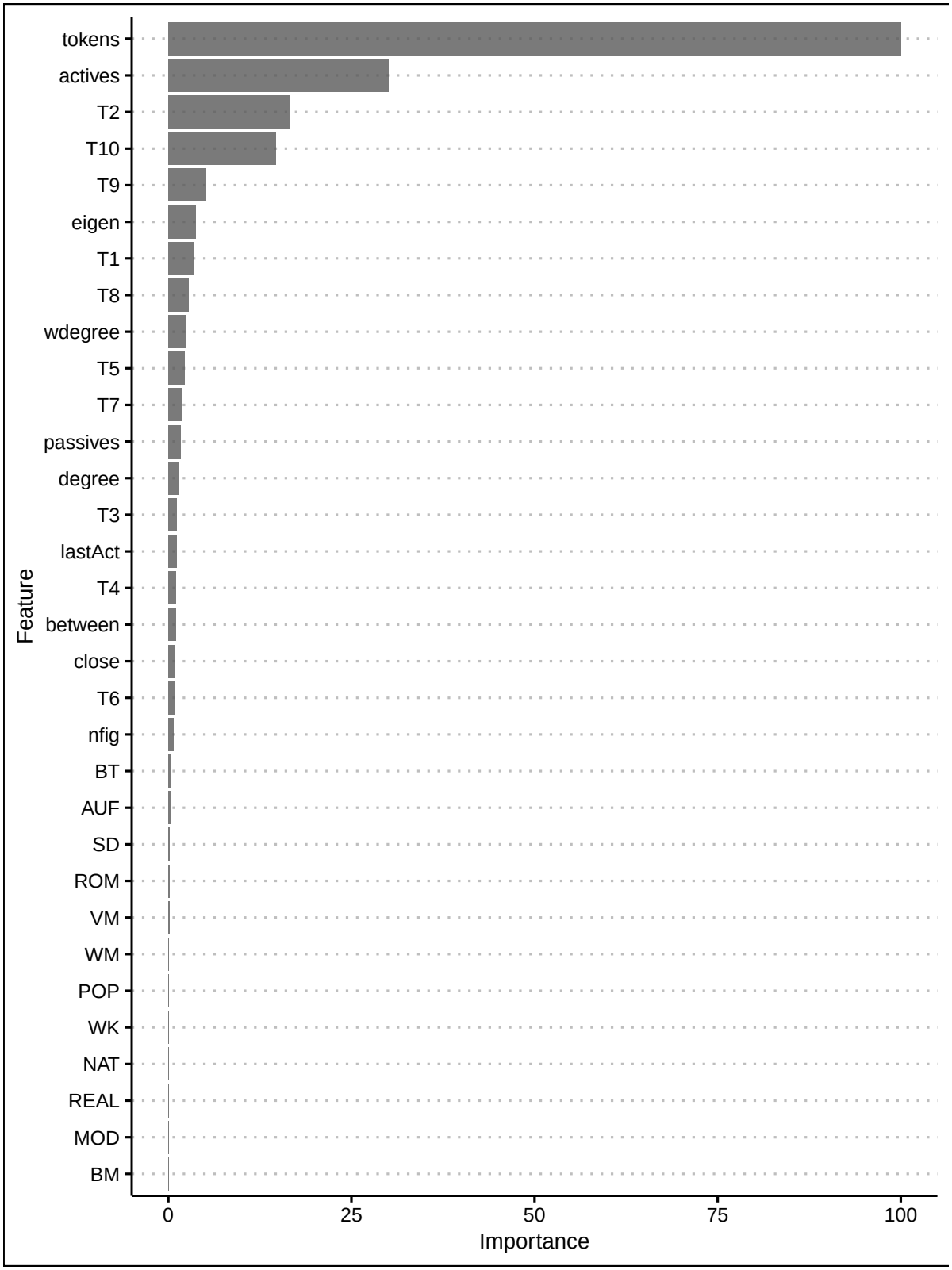


Figure 6.3.: Feature importance for TITLECHARACTER.

highest drop in performance when left out receives the highest feature importance value and all other features are scaled accordingly relative to this top feature. The *tokens* feature is the most important feature, as expected, followed by some topics, the *actives* feature and eigenvector centrality. In general, tokens, topics, stage presence and network features are the best predictors, while the priors and the *lastAct* feature do not contribute much.

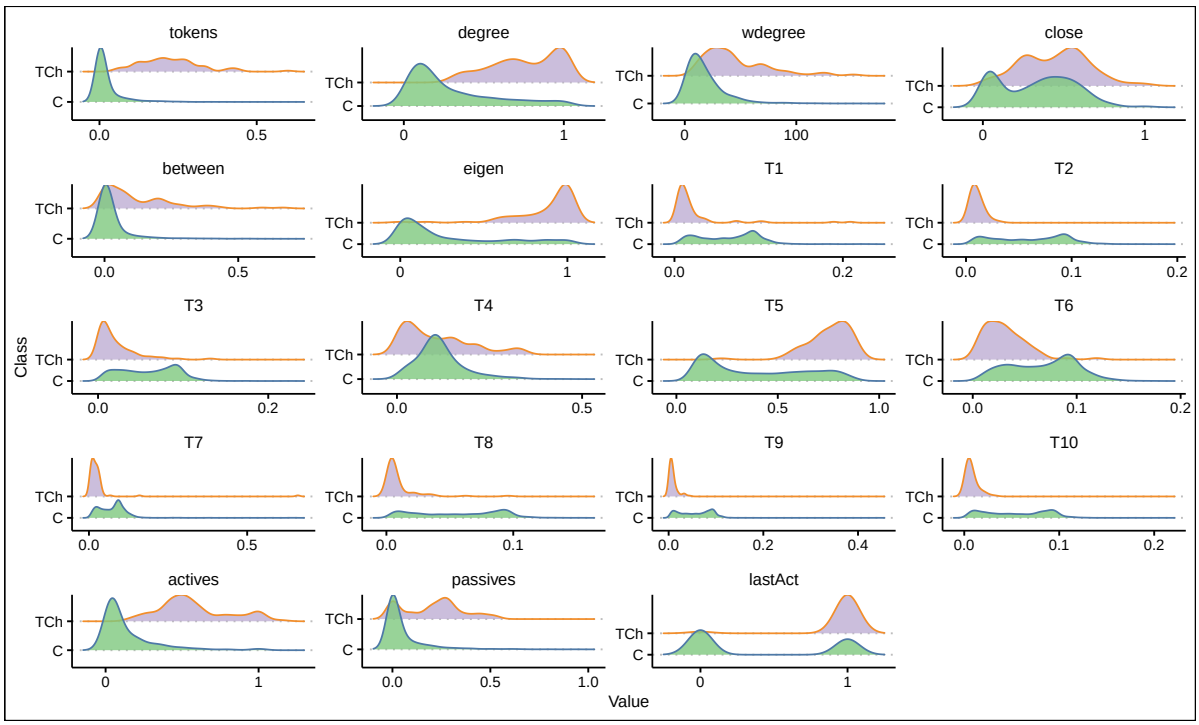


Figure 6.4.: Feature distribution for TITLECHARACTER.

In order to get a better understanding of the overall feature distribution, Figure 6.4 shows the distribution of values per class for each feature. Once again, the *tokens* feature sticks out, as title characters have much higher *tokens* values than other characters. The same is true for the stage presence features and most centrality features except for closeness centrality. Also almost all title characters appear in the last act, hence it does not contribute much as a feature.

6.2.4. Discussion

The F1 score for the best model using all features lies at 0.47. It should also be noted that much less characters were annotated as title characters than as protagonists (43 title characters vs. 106 to 176 protagonists) and therefore less data points available for the

model to generalise. The recall is always high and the precision much lower. This shows that the models assume many characters to be title characters that are actually not. The *tokens* baseline is very strong, but using all other features excluding the *tokens* feature is equally strong and using all features together yields the highest results. This points towards an understanding of title character as a multi-dimensional phenomenon, where features covering different spheres or dimensions are needed in order to correctly classify characters. This is also confirmed when looking at the feature distribution in Figure 6.4. Almost all title characters have a high *tokens* value and almost all non-title-characters have a low *tokens* value, however there is an overlap. The same is true for features that cover other aspects, such as for eigenvector centrality, topic *T5* and *actives* and *passives*.

6.3. Protagonist Detection

Protagonist detection denotes the task of detecting all potential protagonists in a literary text. The term "protagonist" can be (and has been) defined in a variety of ways. Furthermore, there exist related and sometimes interchangeable concepts such as *hero*, *main/principal character* and *antagonist*. For Aristotle, a *hero* is tied to a tragic plot and a character that is, by committing a fatal mistake, thrown from fortune into misfortune (so called *peripeteia*, Aristoteles 1982, pp. 39, 41). A central aspect here is the so called *anagnorisis*, which denotes a moment of recognition that is the trigger for the path to misfortune (Aristoteles 1982, p. 35). An infamous example is the case of the hero Oedipus in Sophocles' play *Oedipus Rex*, who unknowingly kills his own father and marries his mother and realizing this precipitates his ruin. *Oedipus Rex* is also one of the works discussed by Aristotle (Aristoteles 1982, e.g. p. 35).

Pfister (1988) does not give a direct definition of the term *protagonist*, but writes that "[o]ne model for the [...] structure of conflicts is the widespread distinction between the hero and his opposite number, or between the protagonist and the antagonist" (Pfister 1988, p. 170). Shortly after, Pfister identifies the character Dorimant from George Etherege's play *The Man of Mode* to be the protagonist based on his dominance in a character configuration matrix of the play (Pfister 1988, pp. 172–73). This suggests that Pfister understands protagonists to at least be related to some sort of conflict and stage presence. Moretti (2011) also gives no direct definition of protagonists, but suggests to consider the position of a character in a character network as an indicator for the status of being a protagonist (Moretti 2011, p. 4).

In contrast to the title character detection experiments presented previously, protagonists

require a more elaborate operationalization in order to be annotatable. Therefore, I present a definition of *protagonist* that was developed in publications I have contributed to, namely Reiter et al. (2018) and Krautter et al. (2018).
The following definition of *protagonist* is the basis for all following annotations and experiments involving protagonists:

> [...] [W]e [...] define protagonist as *characters that have a central scope of action either by acting themselves or by triggering the action.* (Reiter et al. 2018, p. 1)

This also implies that there can be more than one protagonist as long as the characters proposed as protagonists adhere to the properties described in the definition, and that there is no distinction between "hero" and "opponent" or protagonist and antagonist; both these archetypes may be counted as protagonists (Reiter et al. 2018, p. 1).
Central scope of action can be understood as moving the plot forward and being involved in the main conflict(s) of the play, either causing them or being affected by them. This also involves more passive characters, such as Emilia in Lessing's play *Emilia Galotti*, who has "actions done to her" and suffers from the intrigues by others, but is not actively moving the plot forward. Still, by the above definition, she can be counted as a protagonist, since she triggers the actions of Marinelli and the Prince (by simply existing).
The series of experiments described in this section is called PROTAGONIST and comprises extended experiments, which are based on previous experiments of Reiter et al. (2018) and Krautter et al. (2018), but present new results using extended features. The experiments attempt to show where the differences between classifying protagonists and classifying title characters lie exactly and if it is possible for machine learning models to distinguish protagonists from non-protagonists in the first place.

6.3.1. Annotation and Data

For Reiter et al. (2018), we compiled a list of plays from several epochs and genres, namely *Sturm und Drang* (SuD), *bourgeois tragedy* (BT) and *weimar classicism* (WC). The annotations were carried out by two of the authors of Reiter et al. (2018).
For Krautter et al. (2018), the annotation process was similar to the one in Reiter et al. (2018), but this time three annotators, who were Master's students in German studies annotated the plays by following the definition of *protagonist* from the beginning of Section 6.3. In order to speed up the annotation process, the annotators were using encyclopædias and literary lexicons, in which they read the summaries of the plot and descriptions of the characters involved in the plot. Based on these entries, they decided

to assign the label *protagonist* to a character.

Annotation	Epochs/Genres	Plays	Protagonists (%)	Not-Protagonists (%)	Characters
A1	NAT, SD, WK, WM	34	171 (16)	910 (84)	1081
A2	AUF, BT, ROM, SD	37	176 (16)	928 (84)	1104
A3	BT, POP, VM, WK	36	106 (8)	1296 (92)	1402

Table 6.3.: Statistics concerning the annotations for PROTAGONIST. Shown are the epochs/genres annotated by an annotator, the number of plays, the number of characters annotated as either *protagonist* or *not-protagonist* and the total number of characters.

Table 6.3 shows statistics about the annotations, in particular how many plays were annotated per annotator, which epochs/genres these plays belonged to, and the number of characters classified as protagonist or not as well as the total number of characters. The epochs and genres that are used are *Sturm und Drang* (SD), *Weimarer Klassik* (WK, *weimar classicism*), *Bürgerliches Trauerspiel* (BT, *bourgeois tragedy*), *Aufklärung* (AUF, *enlightenment*), *Romantik* (ROM, *romantic era*), *Naturalismus* (NAT, *naturalism*), *Populäre Stücke* (POP, *popular plays*), *Wiener Moderne* (WM, *Vienna Moderne*) and *Vormärz* (VM, *pre-March era*). *BT*, *SD* and *WK* have both been annotated by two annotators. In these cases, the same plays were annotated by the two annotators for the respective epoch/genre.

Combination	Epoch/Genre	$\lvert\cap$Plays$\rvert$	$\lvert\cap$Characters$\rvert$	Cohen's κ
A1+A2	SD	6	157	0.83***
A1+A3	WK	6	238	0.46***
A2+A3	BT	7	110	0.43***

Table 6.4.: Inter-Annotator Agreement for the plays from the same epoch/genre. Shown are the epoch/genre, the number of plays that overlap between the annotations, the number of characters that overlap, Cohen's κ as well as the p-values in star notation ($*$ with $p < 0.05$, $**$ with $p < 0.01$ and $***$ with $p < 0.001$).

Table 6.4 shows the IAA for these overlapping plays using Cohen's κ (Cohen 1960). The agreement for annotations A1 and A2 is high, while the agreement for A3 with these

two annotations is much lower. This can also be explained by the fact that A3 chose to annotate only half of the characters as protagonists on average when compared to annotations A1 and A2 (8% for A3 compared to 16% for A1 and A2, see Table 6.3). The p-values for all annotation pairings are very low, showing that the κ values are significant.

6.3.2. Experimental Setup

The features used are identical to the ones used in TITLECHARACTER (see Table 6.1). As with title character prediction, *random forest* was used, with 10-fold cross validation during training and splitting the data into 80% train and 20% test set. Additionally, the data was up-sampled using the SMOTE algorithm.
In addition to the RF model, two baselines are applied to get a better idea of how well or not well the model is doing. The first baseline is the majority baseline, which assigns each character the majority class, i.e. not being a protagonist. The *tokens* baseline is a RF model that was only trained using the *tokens* feature, since it turned out to be a very strong feature in pre-experiments.

6.3.3. Results

		Protagonist			Not-Protagonist		
		Precision	Recall	F1	Precision	Recall	F1
Majority Baseline	A1	-	0.00	-	0.84	1.00	0.91
	A2	-	0.00	-	0.84	1.00	0.91
	A3	-	0.00	-	0.92	1.00	0.96
Tokens Baseline	A1	0.72	1.00	0.84	1.00	0.93	0.96
	A2	0.70	0.99	0.82	1.00	0.92	0.96
	A3	0.44	1.00	0.61	1.00	0.90	0.95
Random Forest	A1	0.84	1.00	0.91	1.00	0.96	0.98
	A2	0.80	1.00	0.89	1.00	0.95	0.98
	A3	0.51	1.00	0.68	1.00	0.92	0.96

Table 6.5.: Classification results for the two baselines and the random forest model for all three annotations A1–3. Classification results are additionally divided into choosing (i) the protagonist class and (ii) the not-protagonist class as the positive classes when calculating precision and recall.

The results can be seen in Table 6.5. A1-A3 are the three different annotations. The

majority baseline gives accuracy scores of 84% to 92%, which are improved by 7 to 2 percentage points for using only the *tokens* feature and 9 to 5 percentage points for using all features. Comparing the F1-scores for correctly classifying the protagonist class for the *tokens* baseline and the full model, it can be seen that the tokens-only model achieves acceptable results with 61 to 84% F1, but the full model is able to outperform the tokens-only model with 68 to 91% F1 scores. Furthermore, recall values are consistently high, while precision scores range from 51 to 84% for the full model.

	Protagonist			Not-Protagonist		
	Precision	Recall	F1	Precision	Recall	F1
A1 without Tokens	0.82	0.98	0.89	1.00	0.96	0.98
A2 without Tokens	0.78	1.00	0.88	1.00	0.95	0.97
A3 without Tokens	0.51	1.00	0.67	1.00	0.92	0.96

Table 6.6.: Classification results for random forest models without using the tokens feature for all three annotations A1–3. Classification results are additionally divided into choosing (i) the protagonist class and (ii) the not-protagonist class as the positive classes when calculating precision and recall.

Since the *tokens* feature offers such a strong baseline, the question arises how well the full model would perform without using the tokens feature. Table 6.6 gives results for classifying using all features, *except for* the *tokens* feature. It can be seen that the full model is still able to perform well, also without the *tokens* feature, dropping 1 to 3 percentage points in the F1 score compared to the full model using all features including the *tokens* feature. Compared to the experiments classifying title characters, the drop in performance when leaving out the *tokens* feature is also much lower (13 percentage points dropped in F1 score for title character classification compared to 1 to 3 percentage points dropped in F1 score for protagonist classification).

Figure 6.5 shows feature importance values for the full models on the three annotations A1–3. The *tokens* feature is always the best feature, followed by some topics and the *SD* feature for the annotations A1 and A2 and the lastAct feature for annotation A3. All other features contribute relatively little to the overall performance.

In order to check which features are important when the *tokens* feature is not dominating the classification, feature importance analysis is also performed for the models that do not use the *tokens* feature (*A1woTokens*, *A2woTokens* and *A3woTokens*).

Figure 6.6 shows the results. For A1 and A3, the features are ordered similarly to Figure 6.5 with some minor switches. For A2, the *actives* feature now takes over the

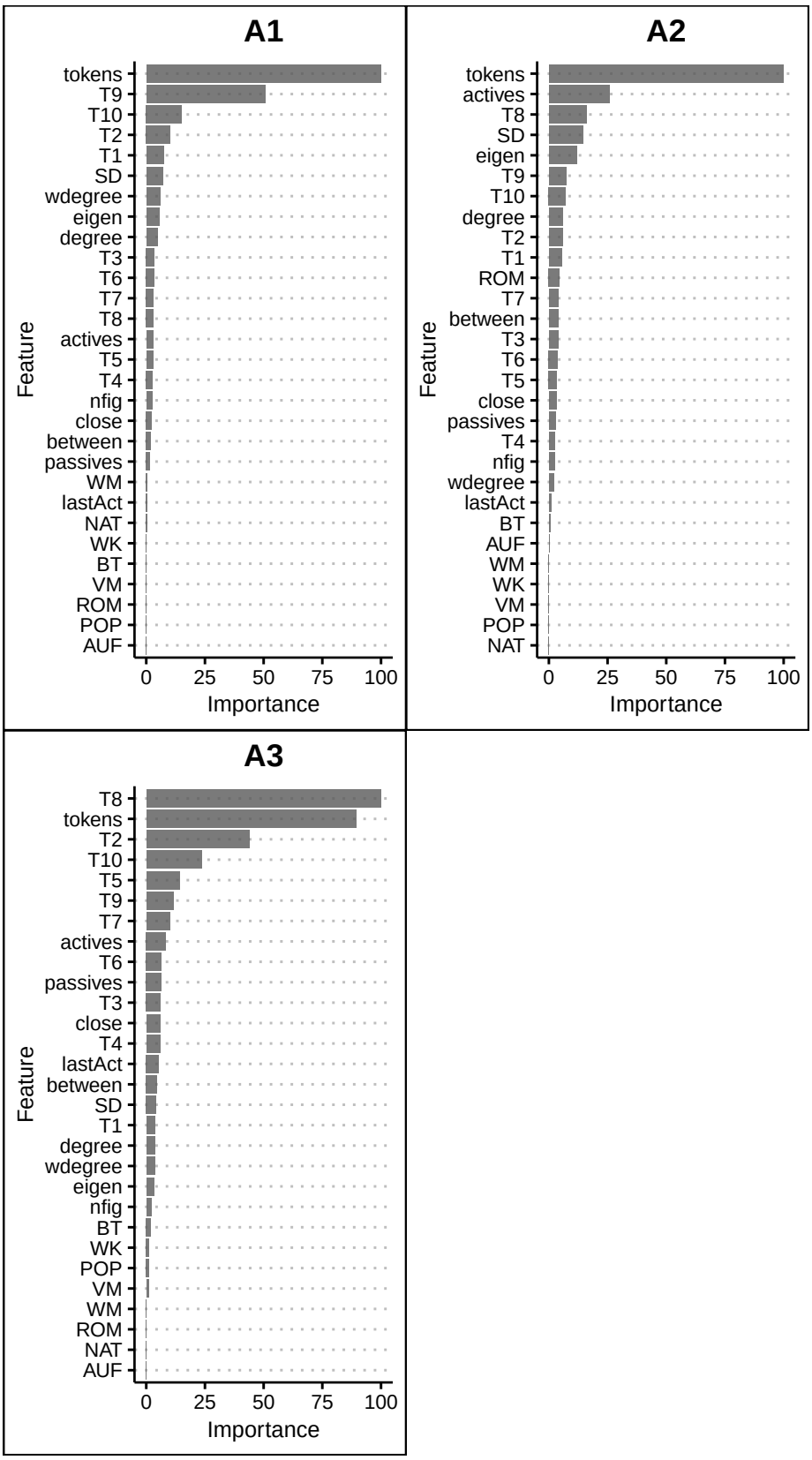


Figure 6.5.: Feature importance analysis for PROTAGONIST and the different annotations A1–3.

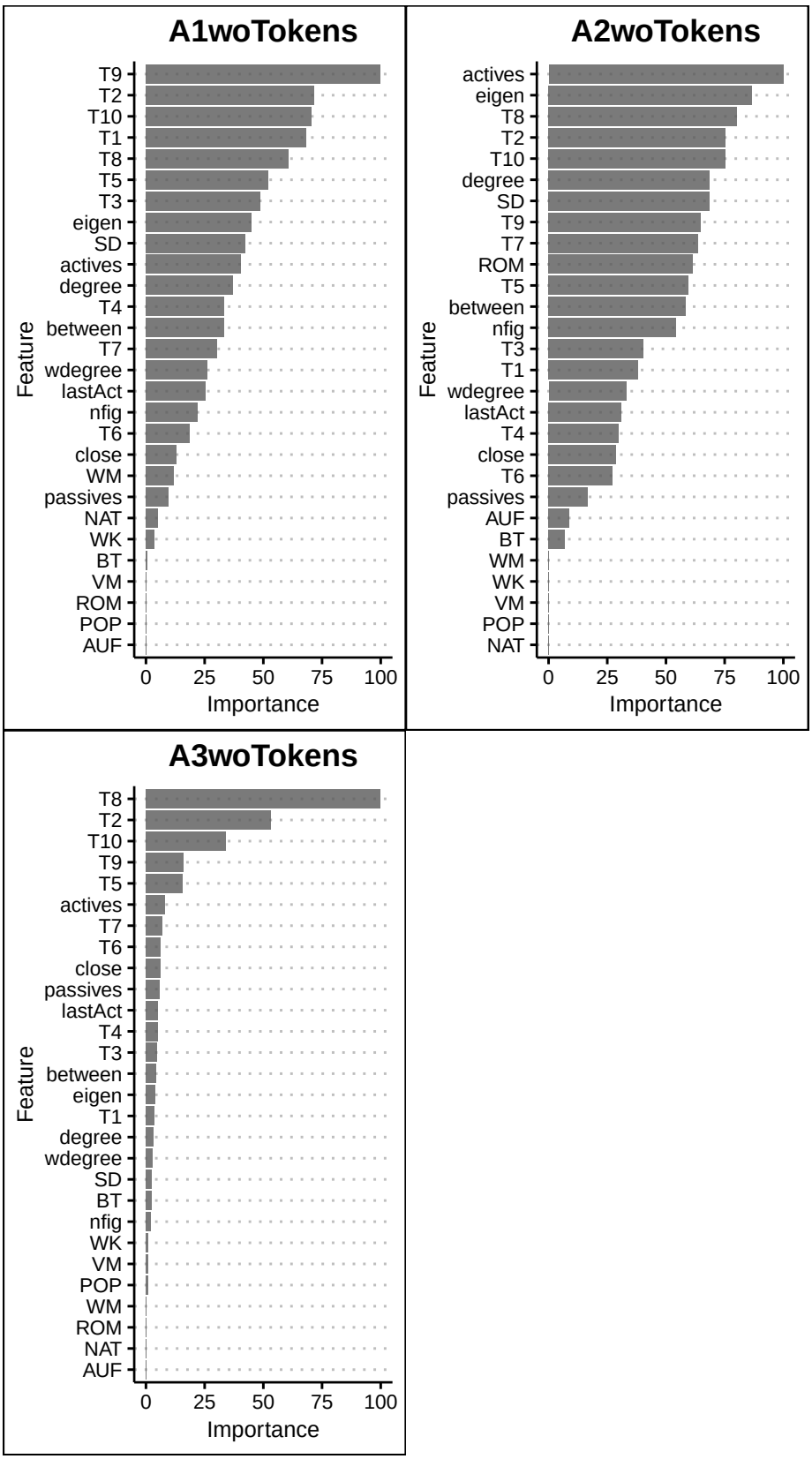


Figure 6.6.: Feature importance analysis for PROTAGONIST and the different annotations A1-3 for the model not using the *tokens* feature.

place of the *tokens* feature, while it did not play a role in Figure 6.5. Also the *ROM* feature is now much more important and many topic features less important than before. For all annotations, many features now play a more important role than before, as their importance values increased overall.
Figure 6.7 shows the feature distribution for the different annotations. It can be seen that the *tokens* feature is very distinctive for all annotations. Overall, the annotations do not differ much in terms of their feature distributions, with some exceptions in the topics, closeness centrality and *actives* and *passives* features. As for TITLECHARACTER, the *tokens*, *degree*, *eigen*, *actives* and topic *T5* are rather distinctive, with many protagonists having a high value in these features while the not-protagonist characters have mostly low features. For all other features, there is often a large overlap in the lower value range.

6.3.4. Discussion

The most conspicuous observation is that the *tokens* feature is a very strong predictor for protagonist detection. There might be two possible explanations for this: first, this is due to a structural property of the plays, because characters that have an important role in the plot will always need to speak more in order to solidify their importance; second, it has been speculated that the centrality of a character in a co-presence network should be a very strong predictor for protagonism (Pfister 1988; Moretti 2011; Algee-Hewitt 2017; Fischer et al. 2018), hence there might be a correlation between centrality in a character network and the number of tokens a character utters, since both metrics measure stage presence to a certain extend.[4] However, since centrality measures alone are not as strong as a predictor as the *tokens* feature, it is clear that centrality alone cannot be the only answer to detect protagonists. Similarly, the *tokens* feature alone is also not the only answer, as it does not yield a close to one-hundred percent F1-score and adding the other types of features further improves the score when used together with the tokens feature. Therefore, we can conclude that protagonist detection is also a multi-dimensional task which requires features that cover different aspects of a characters role and profile. These features can be, but are not necessarily limited to, the features shown above: stage presence, content of speech (topics), and co-presence.
The second pressing observation regards the fact that the recall is always much higher than the precision. The conclusion from this is that the model over-generalizes protagonists and is oftentimes predicting characters to be protagonists that actually are not. On the

[4] In the full model, this effect will not occur, since RF models eliminate highly correlating features internally.

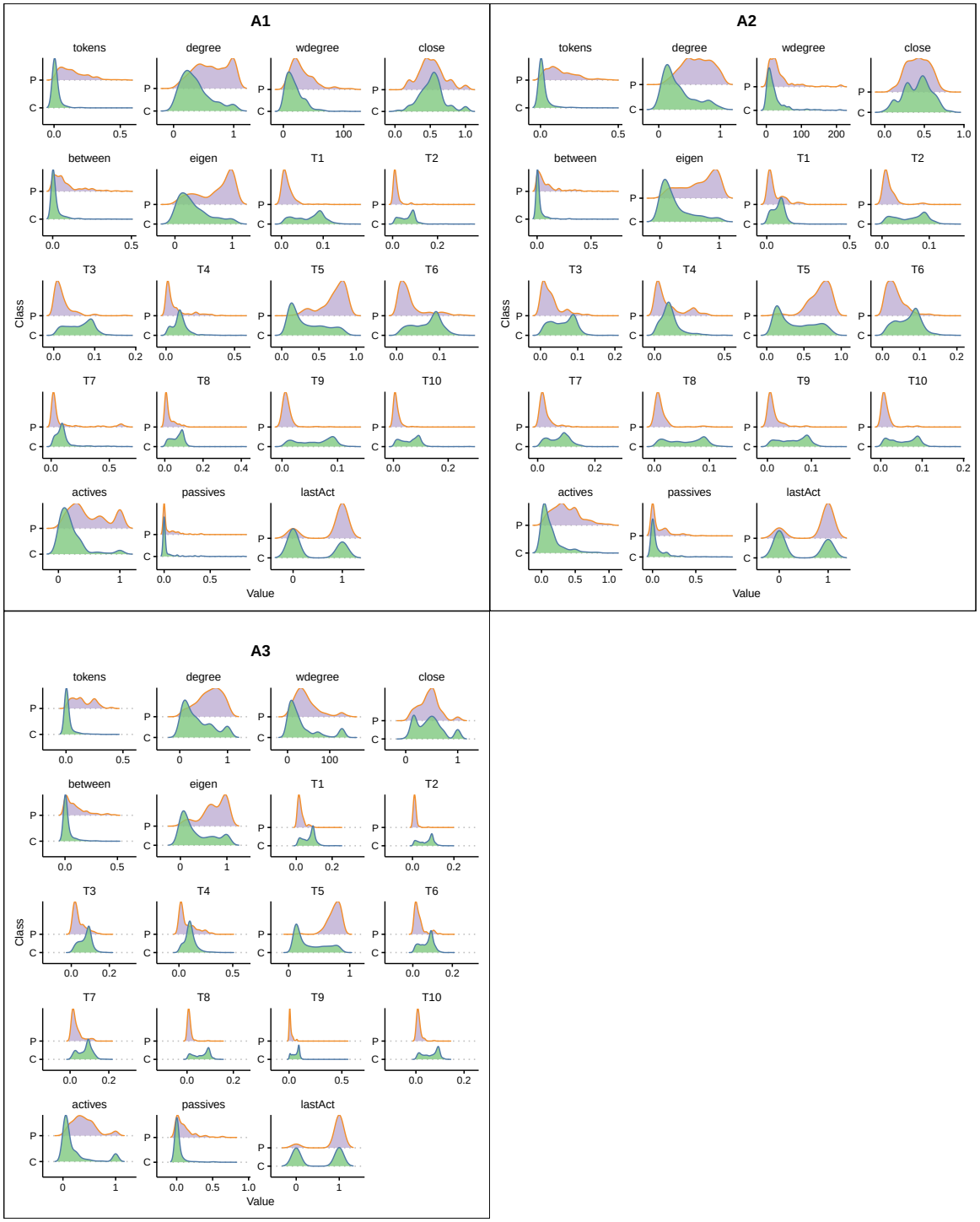


Figure 6.7.: Feature distribution for PROTAGONIST and the different annotations A1–3.

other hand, the model is often able to find almost all protagonists in a play. A possible explanation is that many characters behave similarly to protagonists, but would not be classified as protagonists according to the operationalization given to the annotators. In this view, protagonism lies on a spectrum with prototypical traits in the centre and characters that somewhat behave like typical protagonists but can be considered borderline.
As a third observation, topics are always strong predictors when tokens are not involved, according to the feature importance analysis. This indicates that next to structural features, content-based features are important and the content of a character's speech contributes to protagonist detection in a unique way. However, topics are notoriously difficult to interpret and an analysis of the topics' content in Krautter et al. (2018, p. 35) showed that the topics are not very interpretable from a human perspective.
As a last observation, there are several differences between the three annotations, but also similarities. For two out of the three annotations, the *tokens* feature is by far the strongest predictor, according to feature importance. Should the *tokens* feature be removed, different features become important, but as mentioned before, certain topics are important for all models. In terms of pure evaluation results, models trained on annotation *A1* and *A2* yield almost identical results, while the annotations of *A3* seem much harder to predict for the models. This shows two things: the intersubjectivity of the annotation guidelines is principally given, but the operationalization might still lead different annotators to annotate different phenomena, showing that the operationalization of the concept *protagonist* remains a difficult challenge. Also note that it is not entirely clear how high the quality of the annotations are in regards to the operationalization of the definition of protagonism from the beginning of this Section 6.3. While all annotators have used encyclopediæ to base their annotations on, the knowledge encoded in these works might be different and the interpretation of the annotator of the interpretation of an encyclopedic article about a certain literary work might also differ among annotators. Also, for A3, the genres/epochs might have been more difficult to annotate than for A1 and A2.

6.4. Schemer Detection

Schemers are one of the character types studied in German literary studies. One prominent example is Alt (2004), who describes schemers as typically servants to aristocrats, who manipulate social interactions and relationships in order to gain certain benefits and

autonomy from their superior (Alt 2004, p. 1). Schonlau (2017) works out different particularities about the typical communication style of schemers: typically the audience is aware of the intrigue of the schemer, while the victims of the intrigue are clueless (Schonlau 2017, p. 160) and schemers often eavesdrop in order to get information which they can later use in their intrigue (Schonlau 2017, p. 178).
Being able to automatically detect schemers can help to study a wider range of plays and characters. Furthermore, classifying schemers using a certain set of features can help to detect prototypical properties of schemers and might lead to discovering characters that could be considered schemers, but were previously not in the range of vision of common literary studies.

6.4.1. Annotation and Data

The annotation has been carried out by the first author of Krautter and Pagel (2024, to appear). To this end, several literary encyclopediæ were consulted and if an entry mentions a character of a play to be a schemer, this character was annotated as being a schemer. All other features used for the experiments are extracted automatically from the texts and existing TEI annotations. In total, 50 (5.9%) characters were annotated as schemers, leaving 798 (94.1%) characters annotated as non-schemer. These 848 characters come from a total of 38 plays.

6.4.2. Experimental Setup

The experiments described in the following will be referred to as Schemer. Schemer uses some additional features compared to Protagonist and Titlecharacter, all other features are as described in Section 6.3.
Schemer divides features into different groups that slightly digress from the ones used in Protagonist and Titlecharacter. The groups are

- Action
- CharacterStyle
- Aboutness
- Interaction
- Stage Presence
- Sentiment
- Priors

That the groups are different to the ones used in Protagonist and Titlecharacter

is owed to the fact that detecting schemers is a more complex task than detecting protagonists, since the group of schemers is much more diverse than the group of protagonists and the data is much less in the former case. Hence it seemed necessary to device new features which are able to cover the different aspects that make a character a schemer and explore properties that were not covered by previous features. These new features are described in the following:

Type-Token-Ratio is the number of types, i.e. the unique tokens a character utters, divided by the total number of tokens a character utters.

utteranceLengthMean and utteranceLengthSd are the average length of all utterances a character utters and the standard deviation of all the utterance lengths' of a character.

Word fields contains seven lists of words with lemmas of a certain topic, namely *family*, *love*, *war*, *reason*, *religion*, *politics* and *economy*. The word fields were created by two literary scholars. For each character's utterances, the number of tokens contained within a certain word field are counted and divided by the total number of tokens of this character.

posRatio and negRatio are the number of tokens a character utters that either appear as positive or negative in the SentiWS corpus (Remus, Quasthoff, and Heyer 2010), divided by the total number of tokens of this character.

firstBegin and firstEnd are the offsets in the plays at which a character makes a first or a last utterance, respectively.

decade is a group of boolean feature and encodes in which decade a play has been written by indicating for every character if the play was written in a certain decade or not (one-hot encoding).

prose is a boolean feature which encodes if a play is written in prose or verse.

Table 6.7 gives an overview of all the features used in the experiments of SCHEMER.

Feature Group	Feature Name	Short Name	Value Range
Stage presence	Active presence	actives	$[0,1] \in \mathbb{Q}$
	Passive presence	passives	$[0,1] \in \mathbb{Q}$
	First utterance	firstBegin	$[0,+\infty] \in \mathbb{N}$
	Last utterance	lastEnd	$[0,+\infty] \in \mathbb{N}$
	Tokens	tokens	$[0,1] \in \mathbb{Q}$
	Utterances	utterances	$[0,+\infty] \in \mathbb{N}$
Action	Action Verbs (Utterances)	utt.geben, utt.gehen, utt.hören, utt.kommen, utt.lassen, utt.machen, utt.sagen, utt.sehen, utt.tun, utt.wissen	$[0,1] \in \mathbb{Q}$
	Action Verbs (Stage Directions)	sd.fallen, sd.geben, sd.gehen, sd.kommen, sd.nehmen, sd.sehen, sd.setzen, sd.stehen, sd.treten, sd.werfen	$[0,1] \in \mathbb{Q}$
CharacterStyle	Type-Token-Ratio	TTR	$[0,1] \in \mathbb{Q}$
	Mean Length of Utterances	utteranceLengthMean	$[0,+\infty] \in \mathbb{N}$
	Standard Deviation of Length of Utterances	utteranceLengthSd	$[0,+\infty] \in \mathbb{N}$
Aboutness	Topics	T1-T20	$[0,1] \in \mathbb{Q}$
	Word Fields	love, family, war, reason, religion, economy, politics	$[0,1] \in \mathbb{Q}$
Interaction	Degree	degree	$[0,1] \in \mathbb{Q}$
	Weighted Degree	wdegree	$[0,+\infty) \in \mathbb{N}$
	Closeness	close	$[0,1] \in \mathbb{Q}$
	Betweenness	between	$[0,1] \in \mathbb{Q}$
	Eigenvector	eigen	$[0,1] \in \mathbb{Q}$
Sentiment	Positive	posRatio	$[0,1] \in \mathbb{Q}$
	Negative	negRatio	$[0,1] \in \mathbb{Q}$
Priors	Decade	decade	$\{0,1\}$, Boolean
	Prose/lines	prose	$\{0,1\}$, Boolean

Table 6.7.: Features used in experiment SCHEMER. Given are the broader feature groups, the single features associated to a single group and the possible values a feature can take.

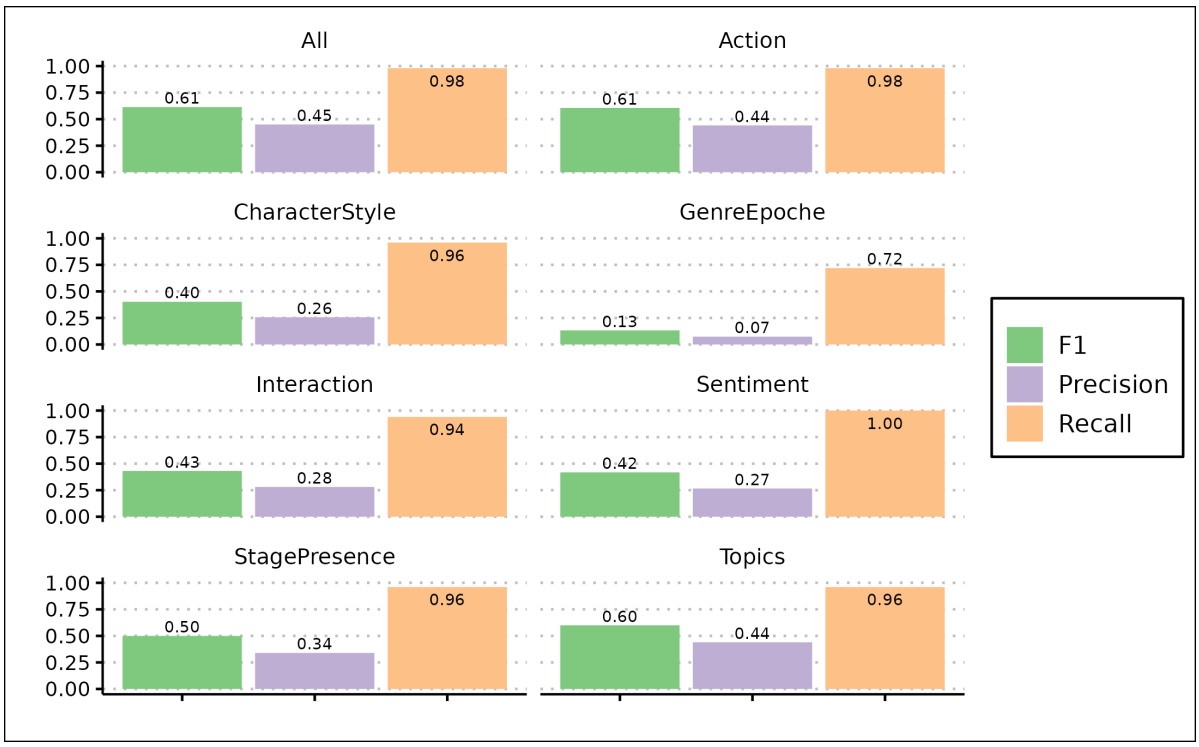


Figure 6.8.: Results for classifying schemers.

6.4.3. Results

Figure 6.8 shows the results of SCHEMER as a barplot. Different feature combinations are shown that correspond to the groups in Table 6.7. Additionally, the results for a model using all features are shown. This model also performs best with an F1 score of 61%, followed by a model using the action features (also 61% F1 score, but one percentage point less in precision) and a model using the topic features (60% F1). As with the other experiments involving protagonists and title characters, recall is consistently high, while the models tend to identify characters as schemers that are actually not, lowering precision.

Figure 6.9 shows the feature importance analysis for SCHEMER and the model using all features. As before, the tokens feature performs best, followed by topic features, number of utterances and word fields (religion, family and politics). Compared to the feature importance distribution from Figures 6.5 and 6.3, it can be seen that the *tokens* feature is not as dominant and that overall many features contribute a good amount towards the performance of the model.

Since the concept of schemer can also be thought of as a prototype classification, with the

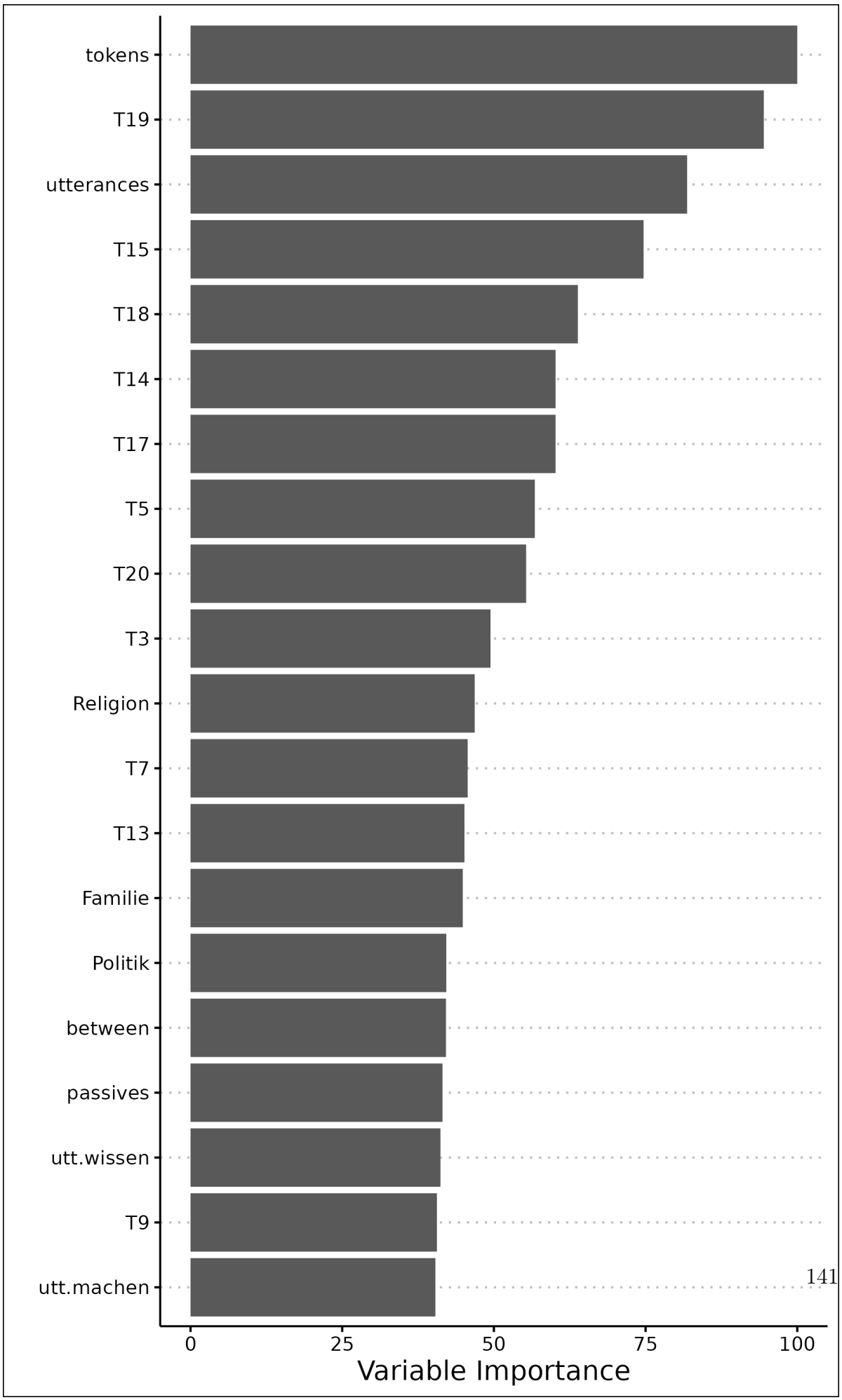


Figure 6.9.: Feature importance for classifying schemers.

most prototypical schemers at the center and outliers further from the center, rather than a strict yes-no classification, Principal Component Analysis (PCA) has been performed. PCA is a dimension reduction technique, which rearranges and re-calculates features from a given dataset and reorganizes them into principal components (PCs), where the first PC is supposed to represent the strongest correlation between the underlying feature distribution. Furthermore, PCA allows to reduce a high-dimensional space to be reduced to two dimension via the first two PCs and thus allows for a visualization of the approximated underlying feature space.
Figure 6.10 shows the PCA for the first two PCs. Data points in the vector space are color-coded by class. It can be seen that for the first two PCs, there is a group of non-schemers and a group of schemers which lie orthogonally to each other. However, the variance explained by both PCs is rather low (10.7% for PC1 and 5.8% for PC2), meaning that the PCs are not able to capture all the particularities of the features in the underlying feature distribution. This also means that this non-linearity of the data once again calls for an approach that utilizes features from multiple independent sources. Furthermore, there are still many non-schemers inside the group that the PCA identified as the schemer group.
These characters might behave more like prototypical schemers and lead to potential classification errors. To verify this, Figure 6.11 shows the same PCA, but characters that were miss-classified by the full model (false positives) are shown in yet another color. It can be seen that indeed the model got thrown off by characters that lie in a similar area to the actual schemers.

6.4.4. Discussion

The best results, coming from the model using all features, are 3 to 30 percentage points below the best results for Protagonist and Titlecharacter, respectively. This is not surprising, since the concept of schemer is conceptually much more complex than the concepts of protagonist or title character. On the other hand, with 3 percentage points of difference to the best results of Titlecharacter, the results for Schemer are not that far of from the results of Titlecharacter. This might also point towards that the concept of title character cannot easily be operationalized using text based features, i.e. the reasons for why a character is chosen as title character might often lie outside of the textual surface. However, more investigation would be needed to further support this claim. Coming back to schemers, the topic and action based features performed best on their own, which is similar to Protagonist and Titlecharacter for the case of

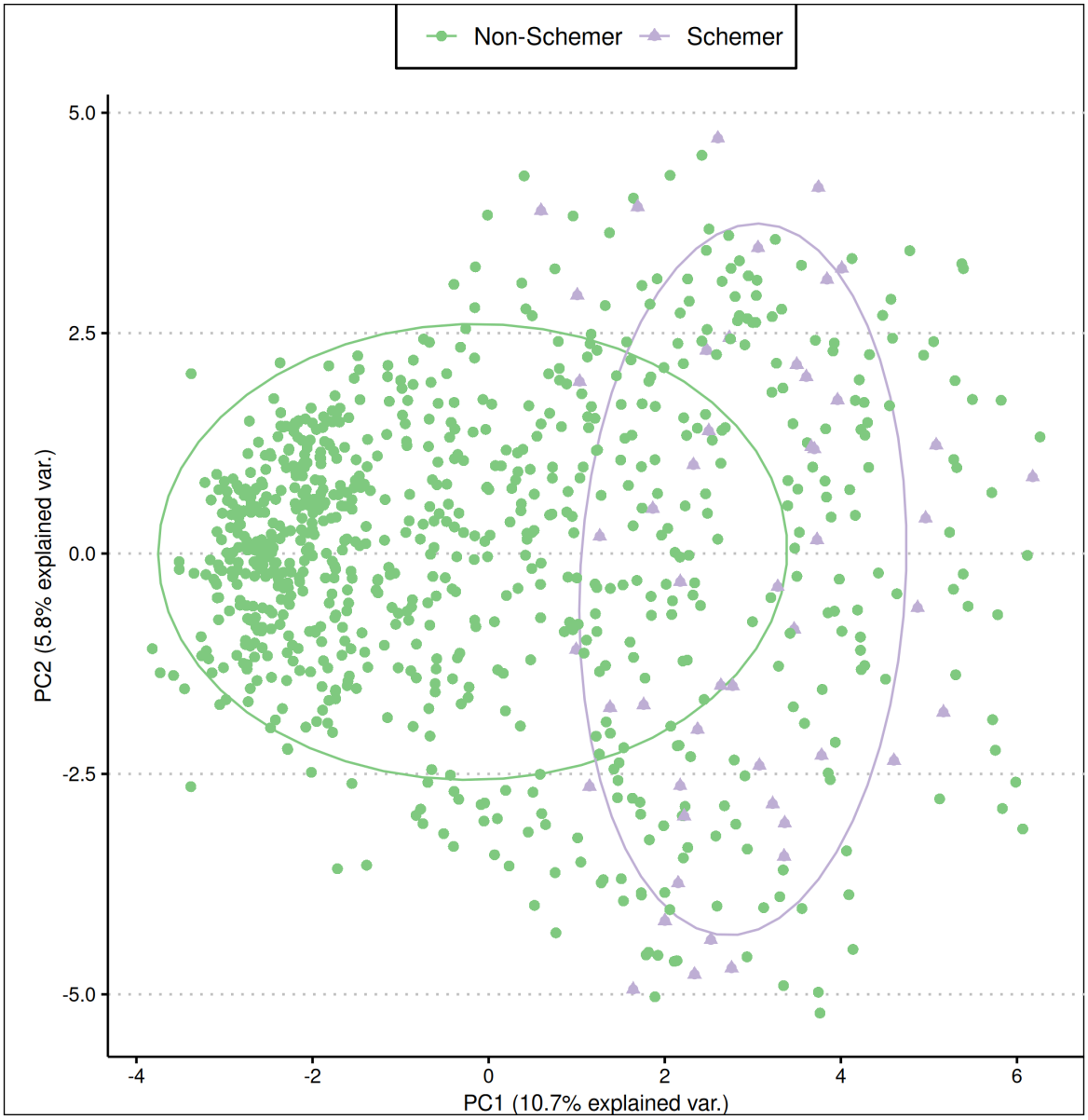


Figure 6.10.: PCA for classifying schemers.

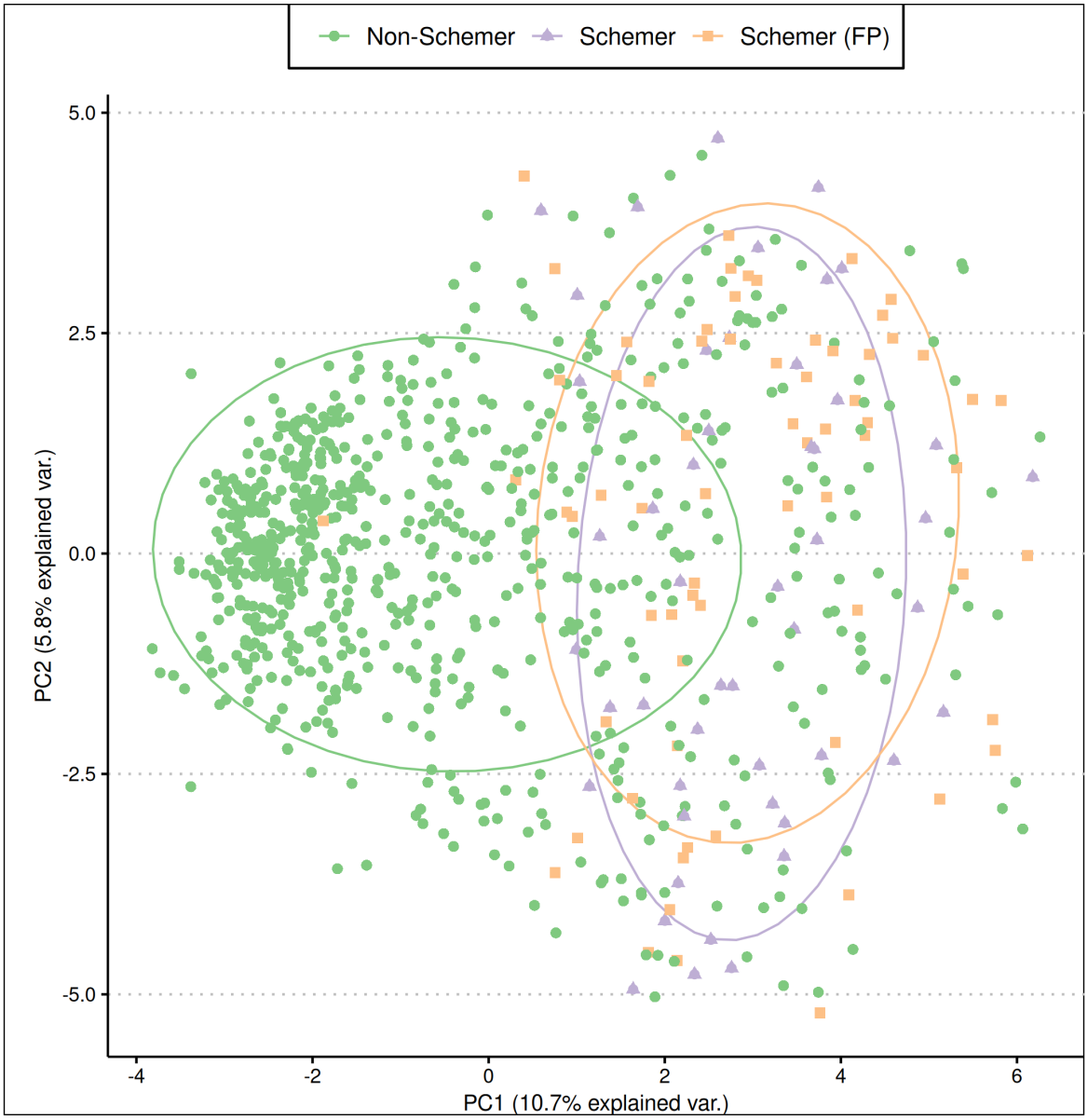


Figure 6.11.: PCA with false positives for classifying schemers.

topics, and interesting for the case of action verbs, since such features were not used for PROTAGONIST and TITLECHARACTER. This might point towards that next to token frequency, the semantic component of a character's speech is the biggest contributing factor for classifying character type. Also note that, different from the ten topics trained for PROTAGONIST, the twenty topics trained for SCHEMER are much more interpretable, as described in Krautter and Pagel (2024, to appear).

When looking at the feature importance of the action verbs, two specific words were most useful for the model, *wissen* (*to know*) and *machen* (*to do/to make*), both use in character speech. This is in line with literary research, were the knowledge of the schemer over other characters (eavesdropping) and knowledge dynamics in the sense of the schemer's victims not knowing of an intrigue that the schemer and the audience are well aware of, are defining characteristics of schemers. Action verbs occurring in stage directions are not important for the model, however this might be due to the lower number of occurrences rather than an actual reflection of the usefulness of these features. The *tokens* feature is once again the single most important feature, much in line with the PROTAGONIST and TITLECHARACTER experiments.

Lastly, the PCA analysis shows that the group of schemers and the group of other characters lie perpendicular to each other, at least in the reduction to a two dimensional feature space. It should however also be noted that the explained variance for the PCs 1 and 2 is rather low, hence the true representation of the full feature space is limited. More interestingly, highlighting the false positives in this feature space shows that they overlap with the groups of true schemers. Therefore, the model's mistakes are not that severe since based on the feature space, the false positives appear like plausible schemers. This also shows that the model is able to generalise to a certain extend and is not overfitting on the data.

6.5. Summary

This chapter presented results on different types of character detection, in particular for title character, protagonist and schemer detection. Protagonist detection proved to be the best performing task with an F1 score of 0.84 for the best performing model using all features. Single well performing features were the *tokens* feature, which also performed very strong on its own, and different topic-based features, as well as stage presence features. The models generally struggle to produce high precision values, but perform very strongly regarding recall.

The results were similar for detecting title characters, however the results were overall lower. This can be explained by the smaller amount of training data and that the task is potentially harder to operationalize and a phenomenon that can only properly decided with features from outside the text surface.
Once again, for detecting schemers, the results are in line with the previous two experiments, with the scores being once again slightly lower than for title character detection. The explanation for this is once again an even smaller dataset and a difficult task. A PC analysis was able to show that the group of schemers lies perpendicular to the group of other characters in a two-dimensional feature space.

Und außerdem, bedenken Sie, was können Sie als Fremde, ohne Schutz, ohne Verbindungen gegen mich durchsetzen, wenn ich als Ihr Feind auftrete? Deshalb schlage ich Ihnen eine Vereinigung vor.
(And besides, think about it, what can you do against me as a stranger, without protection, without connections, if I act as your enemy? That is why I am proposing a union.)

Udaschkin in Gustav Freytag's "Graf Waldemar"

7

Enhancement of Character Type Detection using Coreference Information

It stands to reason that coreference information is helpful for the automatic detection of literary character types. On the one hand, counting the mentions of characters can be seen as complementary to existing features used for character type detection such as counting the tokens a character utters and the amount of times a character is present on stage. On the other hand, coreference information offers the possibility to exploit previously unattainable sources of information, such as considering what a character is saying about another character and how often a character is mentioned when themselves not present on stage.
Consequently, in this chapter, the work on coreference annotation and resolution presented previously in Chapters 4 and 5 is combined with the approaches of character type detection of Chapter 6. New features based on the coreference information are used, namely features based on character mentions and character networks based on mentions, as well as a modified version of the *passives* feature. All experiments of Chapter 6 are re-run using the new features and re-evaluated. It can be shown that coreference information greatly helps to improve and inform character type detection. This shows the importance of coreference annotation and resolution for literary texts and its usefulness for downstream tasks commonly performed in CLS. As coreference information for character type detection

is rarely used (see Jahan, Mittal, and Finlayson 2021, as a notable exception), this chapter hopes to shed new light on the potential of coreference information for relevant CLS tasks such as character type detection and offers new experimental evaluation on the topic.

7.1. Enhancing Character Type Prediction using Coreference Information

The section describes experiments under the collective name CHARACTERTYPE-COREF, with sub-experiments PROTAGONIST-COREF, TITLECHARACTER-COREF and SCHEMER-COREF. The goal is to use the coreference informations gathered in Chapters 4 and 5 in order to improve the performance of the experiments PROTAGONIST, TITLECHARACTER and SCHEMER of Chapter 6.

7.1.1. Annotation and Data

The annotations are same as for the PROTAGONIST, TITLECHARACTER and SCHEMER experiments of Chapter 6. The only new information is the addition of coreference annotations from Chapter 4. These annotations are used to inform new coreference-based features described in the following section.

7.1.2. Experimental Setup

The features used are the same as for PROTAGONIST, TITLECHARACTER and SCHEMER in Chapter 6, except for the modified and added features listed below.

Mentions is the number of mentions of a character. Mentions can either be by the same character (e.g. first person pronouns) or mentions by other characters. This value is normalized by the total number of mentions in a play.

Passive-coref is a modification of the *passives* feature. Instead of counting scenes where a character is mentioned by name, the mentions from the coreference annotations are used.

Degree-coref is related to the *degree* feature (see Section 6.2.2), but is computed on the coreference network of a play. Coreference networks are networks that are build on

the coreference information instead of on the co-presence information of a play. See Pagel (2022a) for details. **Wdegree-coref**, **between-coref**, **close-coref** and **eigen-coref** are, like *degree-coref*, based on coreference networks and otherwise computed like their co-presence counterparts from Section 6.2.2.

Table 7.1 shows the features used for CHARACTERTYPE-COREF.

Feature Group	Feature Name	Short Name	Value Range
Stage presence	Active presence	actives	$[0,1] \in \mathbb{Q}$
	Passive presence	passives	$[0,1] \in \mathbb{Q}$
	Passive presence on coreference networks	passive-coref	$[0,1] \in \mathbb{Q}$
	First utterance	firstBegin	$[0,+\infty] \in \mathbb{N}$
	Last utterance	lastEnd	$[0,+\infty] \in \mathbb{N}$
	Tokens	tokens	$[0,1] \in \mathbb{Q}$
	Mentions	mentions	$[0,1] \in \mathbb{Q}$
	Utterances	utterances	$[0,+\infty] \in \mathbb{N}$
Action	Action Verbs (Utterances)	utt.geben, utt.gehen, utt.hören, utt.kommen, utt.lassen, utt.machen, utt.sagen, utt.sehen, utt.tun, utt.wissen	$[0,1] \in \mathbb{Q}$
	Action Verbs (Stage Directions)	sd.fallen, sd.geben, sd.gehen, sd.kommen, sd.nehmen, sd.sehen, sd.setzen, sd.stehen, sd.treten, sd.werfen	$[0,1] \in \mathbb{Q}$
CharacterStyle	Type-Token-Ratio	TTR	$[0,1] \in \mathbb{Q}$
	Mean Length of Utterances	utteranceLengthMean	$[0,+\infty] \in \mathbb{N}$
	Standard Deviation of Length of Utterances	utteranceLengthSd	$[0,+\infty] \in \mathbb{N}$
Aboutness	Topics	T1-T10	$[0,1] \in \mathbb{Q}$
	Word Fields	love, family, war, reason, religion, economy, politics	$[0,1] \in \mathbb{Q}$
Interaction	Degree	degree	$[0,1] \in \mathbb{Q}$
	Weighted Degree	wdegree	$[0,+\infty) \in \mathbb{N}$
	Closeness	close	$[0,1] \in \mathbb{Q}$
	Betweenness	between	$[0,1] \in \mathbb{Q}$
	Eigenvector	eigen	$[0,1] \in \mathbb{Q}$
	Degree on coreference networks	degree-coref	$[0,+\infty) \in \mathbb{N}$
	Weighted Degree on coreference networks	wdegree-coref	$[0,+\infty) \in \mathbb{N}$
	Closeness centrality on coreference networks	close-coref	$[0,1] \in \mathbb{Q}$
	Betweenness centrality on coreference networks	between-coref	$[0,1] \in \mathbb{Q}$
	Eigenvector centrality on coreference networks	eigen-coref	$[0,1] \in \mathbb{Q}$
Sentiment	Positive	posRatio	$[0,1] \in \mathbb{Q}$
	Negative	negRatio	$[0,1] \in \mathbb{Q}$
Priors	Decade	decade	$\{0,1\}$, Boolean
	Prose/lines	prose	$\{0,1\}$, Boolean

Table 7.1.: Features used in the set of experiments CHARACTERTYPE-COREF. Given are the broader feature groups, the single features associated to a single group and the possible values a feature can take.

	Title character			Not-Title-Character		
	Precision	Recall	F1	Precision	Recall	F1
Without Epochs/Genre	0.40	0.88	0.55	0.99	0.94	0.97
Without tokens	0.38	0.88	0.53	0.99	0.94	0.97
All features	0.39	0.88	0.54	0.99	0.94	0.97

Table 7.2.: Classification results for TITLECHARACTER-COREF using coreference information for the random forest model. Classification results are additionally divided into choosing (i) the protagonist class and (ii) the not-protagonist class as the positive classes when calculating precision and recall. For a comparison with the results without coreference information, see Table 6.2.

	Protagonist			Not-Protagonist			
	Precision	Recall	F1	Precision	Recall	F1	Accuracy
A1	0.89	1.00	0.94	1.00	0.96	0.98	0.97
A2	0.86	1.00	0.93	1.00	0.96	0.98	0.97
A3	0.66	1.00	0.80	1.00	0.93	0.96	0.94

Table 7.3.: Classification results for PROTAGONIST-COREF using coreference information for the random forest model for all three annotations A1–3. Classification results are additionally divided into choosing (i) the protagonist class and (ii) the not-protagonist class as the positive classes when calculating precision and recall. For a comparison with the results without coreference information, see Table 6.5.

	Protagonist			Not-Protagonist			
	Precision	Recall	F1	Precision	Recall	F1	Accuracy
A1oTokens	0.90	1.00	0.94	1.00	0.97	0.98	0.97
A2oTokens	0.86	1.00	0.92	1.00	0.95	0.98	0.96
A3oTokens	0.67	1.00	0.80	1.00	0.93	0.97	0.94

Table 7.4.: Classification results for PROTAGONIST-COREF for random forest models without using the tokens feature for all three annotations A1–3. Classification results are additionally divided into choosing (i) the protagonist class and (ii) the not-protagonist class as the positive classes when calculating precision and recall. For a comparison with the results without coreference information, see Table 6.6.

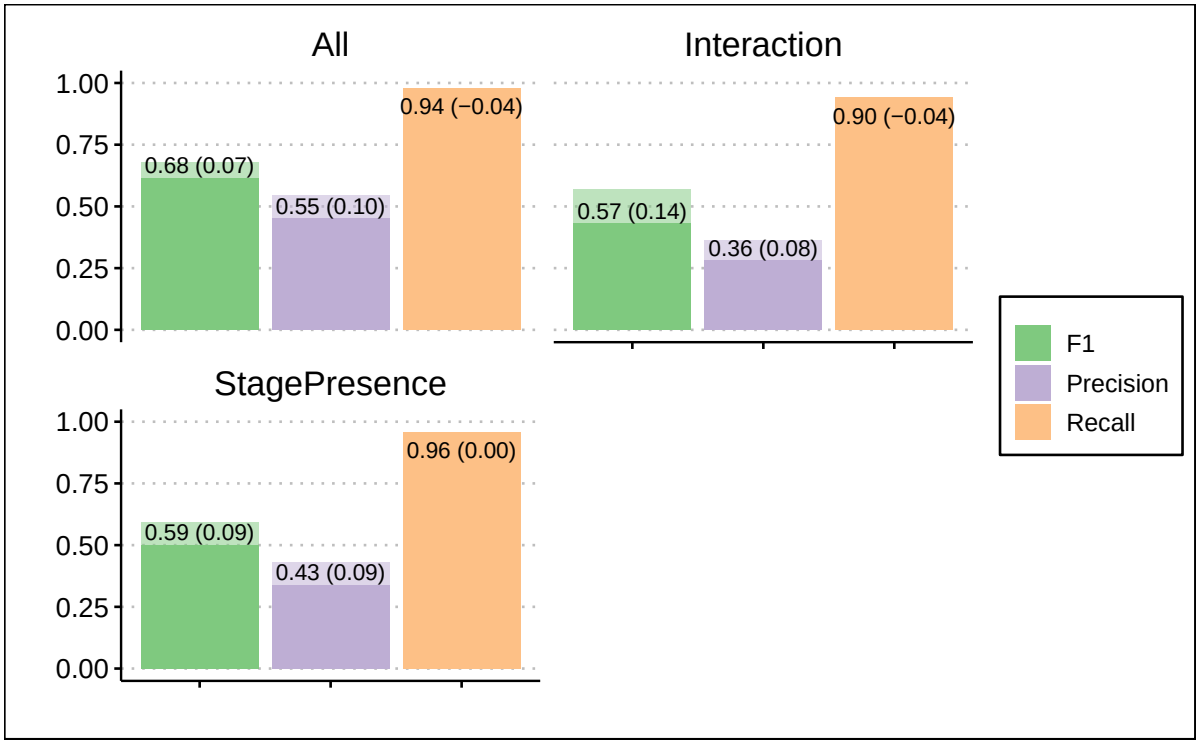


Figure 7.1.: Results for SCHEMER-COREF. The improvement over the results in Figure 6.8 is indicated by a lighter colouring and the value of the difference is given in brackets. Only the feature groups for which coreference information was used are shown.

7.1.3. Results

Tables 7.2, 7.3, 7.4 and Figure 7.1 show the updated results for TITLECHARACTER-COREF, PROTAGONIST-COREF and SCHEMER-COREF, respectively.
For all three experiments, the results are higher than for the experiments from Chapter 6. Table 7.2 shows the results for TITLECHARACTER-COREF. The improvements in F1 score for classifying title characters compared to the results in Table 6.2 ranges from 0.1 to 15 percentage points. Notably, the improvements for the model using everything but the *tokens* model are highest, presumably since the *mentions* feature is also very strong. Leaving out the epoch/genre features further improves the results of the model with all features by one percentage point. There is not much difference between TITLECHARACTER and TITLECHARACTER-COREF results regarding classifying not-title-characters, but the classification result were already very high to begin with.
As for PROTAGONIST-COREF, the results can be seen in Table 7.3. Using coreference-based features has a huge impact, improving the F1 score by up to 12 percentage points, compared to Table 6.5. Annotation A3, which had much lower results in PROTAGONIST compared to A1 and A2, profits especially well. The same can be seen for the results not using the *tokens* feature in Table 7.4, improving the F1 scores by 4 to 13 percentage points over the scores in PROTAGONIST Table 6.6, with A3 once again profiting the most.
As can be seen in Figure 7.1, the results for SCHEMER-COREF are overall higher as for the SCHEMER experiments. For the model using all features, the improvement amounts to +7 percentage points in F1 score, +10 percentage points in precision, but −4 percentage points in recall. Given that the recall is very strong for all models and experiments, these are promising results and show that coreference information is able to make significant contributions towards schemer detection. Naturally, the highest improvements could be made for the feature groups utilizing the coreference information.
As can be seen from Figure 7.2, the distribution of the *mentions* feature is spread out similarly to the distribution of the *tokens* feature. The *eigenvector* feature distribution is much more clearly separated between title characters and not-title-characters than in TITLECHARACTER. Interestingly, the distribution of the *passives* feature did not change much, suggesting that using mentions by name already approximated the use of coreferntial mentions. The situation is similar for the feature distributions of PROTAGONIST-COREF in Figure 7.3, although the *passives* feature is now distributed much more distinctly between protagonists and not-protagonists.
The feature importance for TITLECHARACTER-COREF is similar to the one in TITLECHARACTER, as seen in Figure 7.4. Two notable differences are that the *mentions*

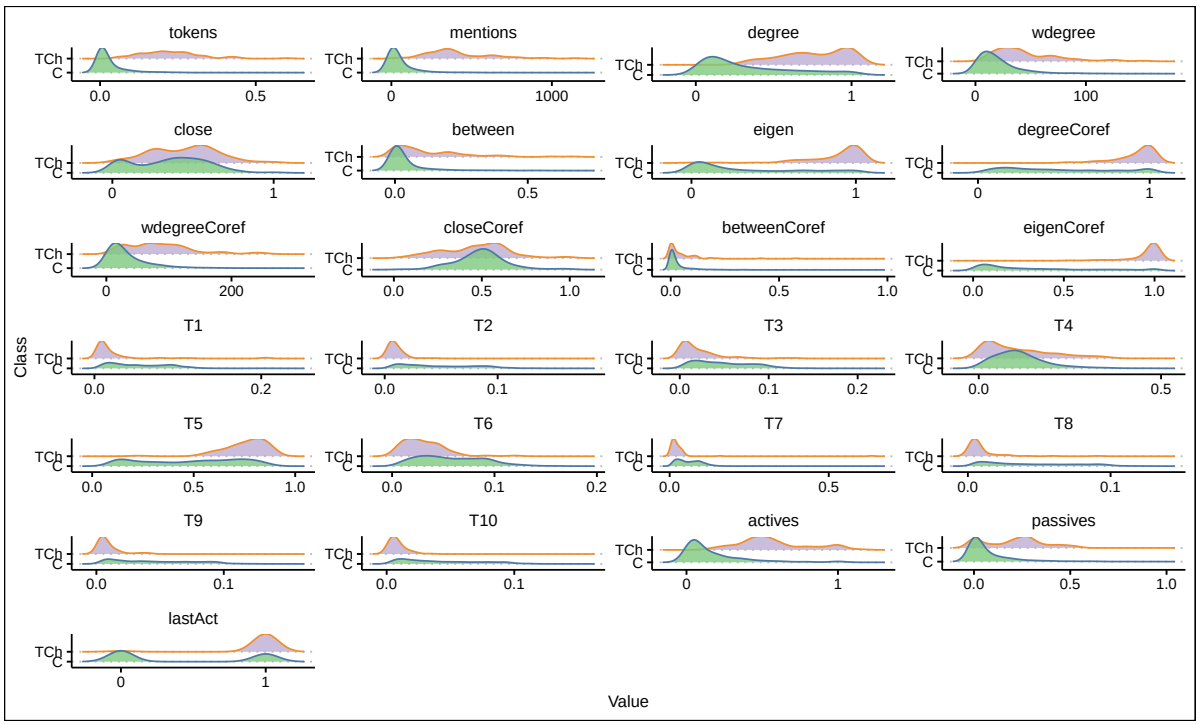


Figure 7.2.: Feature distribution for TitleCharacter-Coref.

feature is more important than the *tokens* feature and the *eigenCoref* feature is relatively important, as could also already be seen from the feature distribution in Figure 7.2.
For Protagonist-Coref's feature importance in Figure 7.5, the situation is similar. The *mentions* feature is either most important (A1 and A3) or the second most important feature (A2). The *eigenCoref* feature is also the most important feature after some topics features, except for annotation A3. For A1, it can also be seen that many more features have gained relative importance when compared to Protagonist.
When looking at the feature importance for Schemer-Coref in Figure 7.6, it becomes clear that the *tokens* feature and topic-based features are still the best performing features, but especially the *passives* feature, the *mentions* feature, the *eigenCoref* and *wdegreeCoref* features are able to make great contributions towards the results.

7.1.4. Discussion

These results show that coreference is generally a useful feature for character type detection and helps improving precision scores with which the models previously struggled. On the other hand, in a feature importance analysis, not all added coreference features were evaluated to be useful for the system, however there might also be hidden correlations

Figure 7.3.: Feature distribution for PROTAGONIST-COREF and the different annotations A1–3.

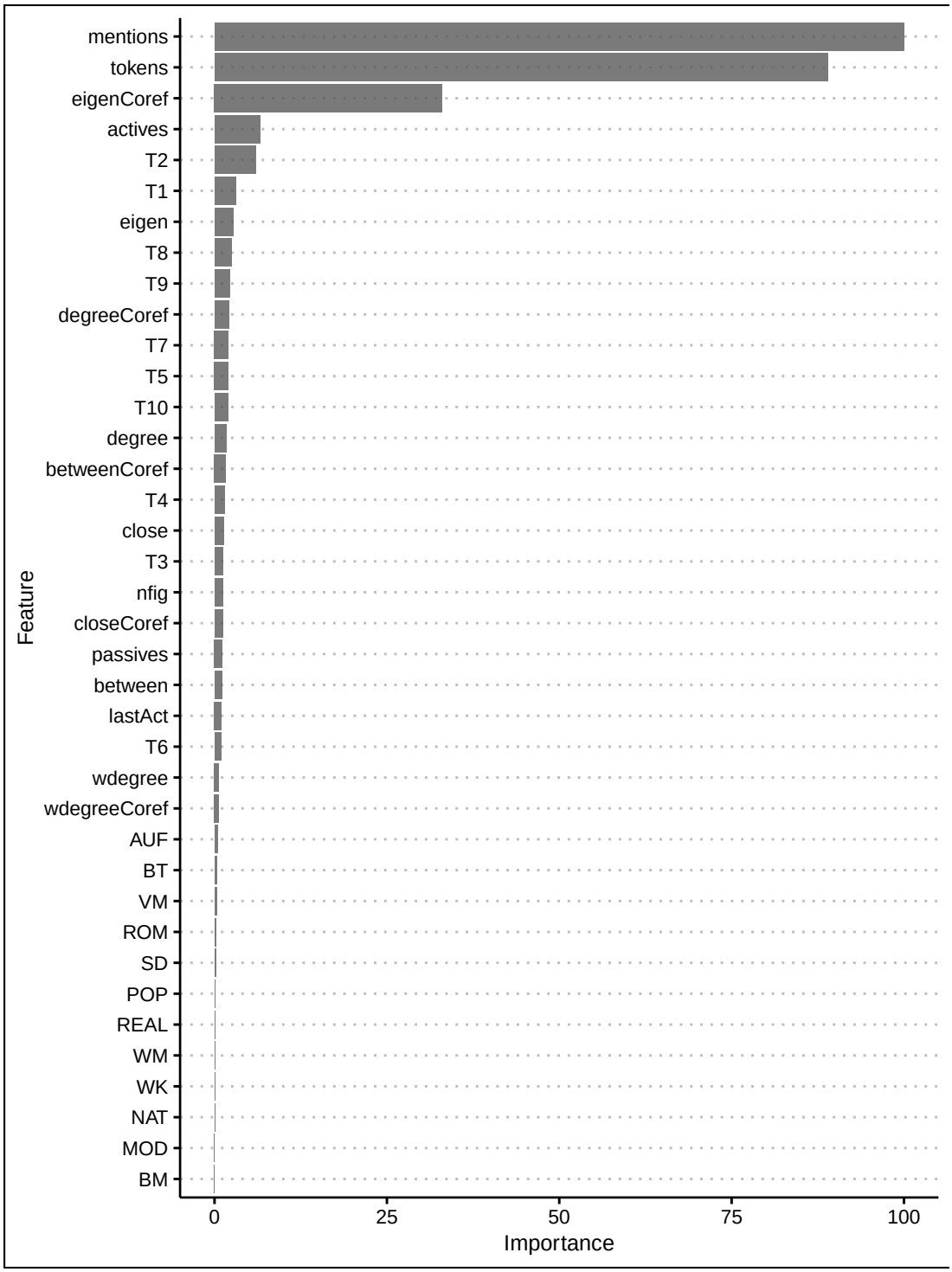


Figure 7.4.: Feature importance analysis for TitleCharacter-Coref.

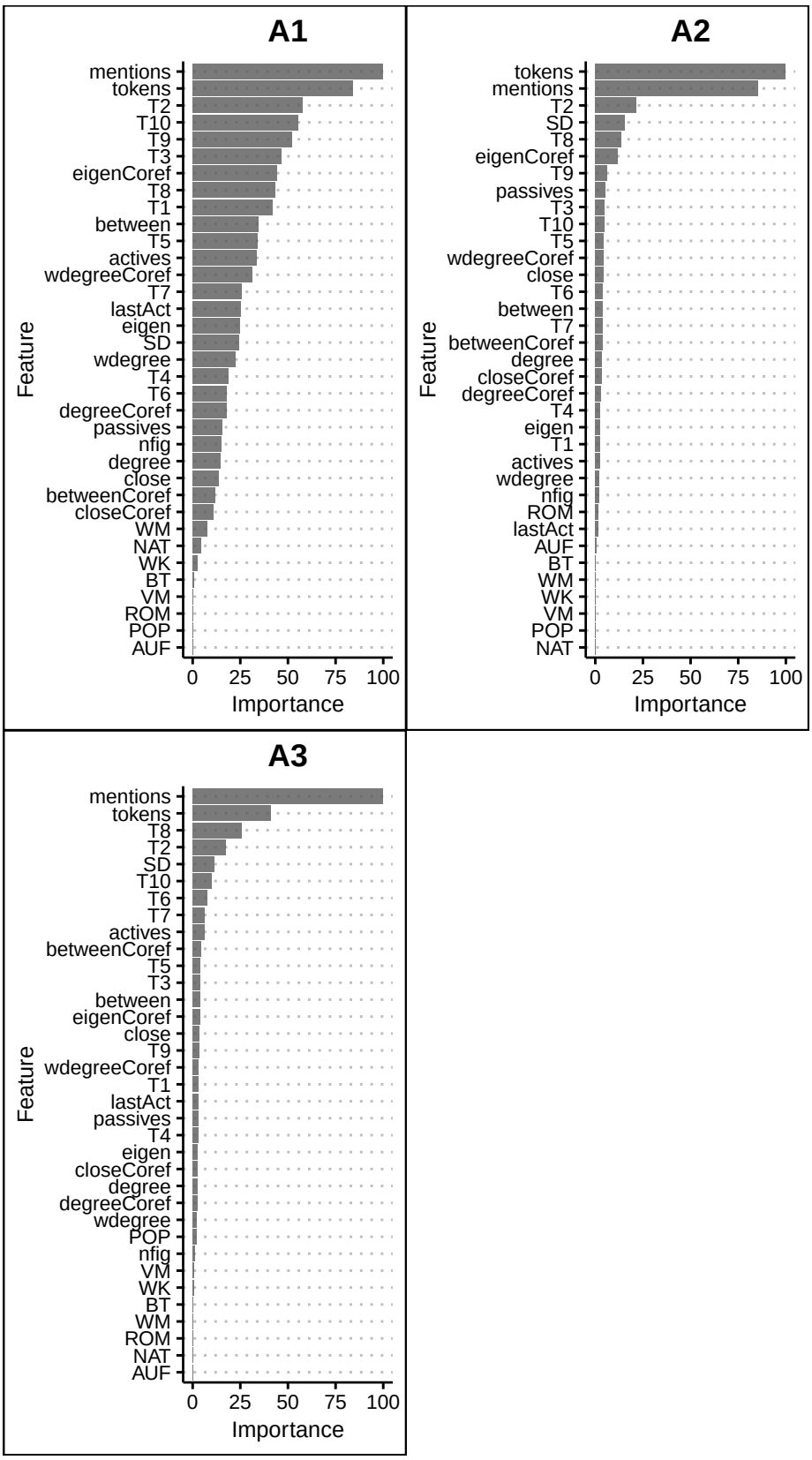


Figure 7.5.: Feature importance analysis for Protagonist-Coref and the different annotations A1–3.

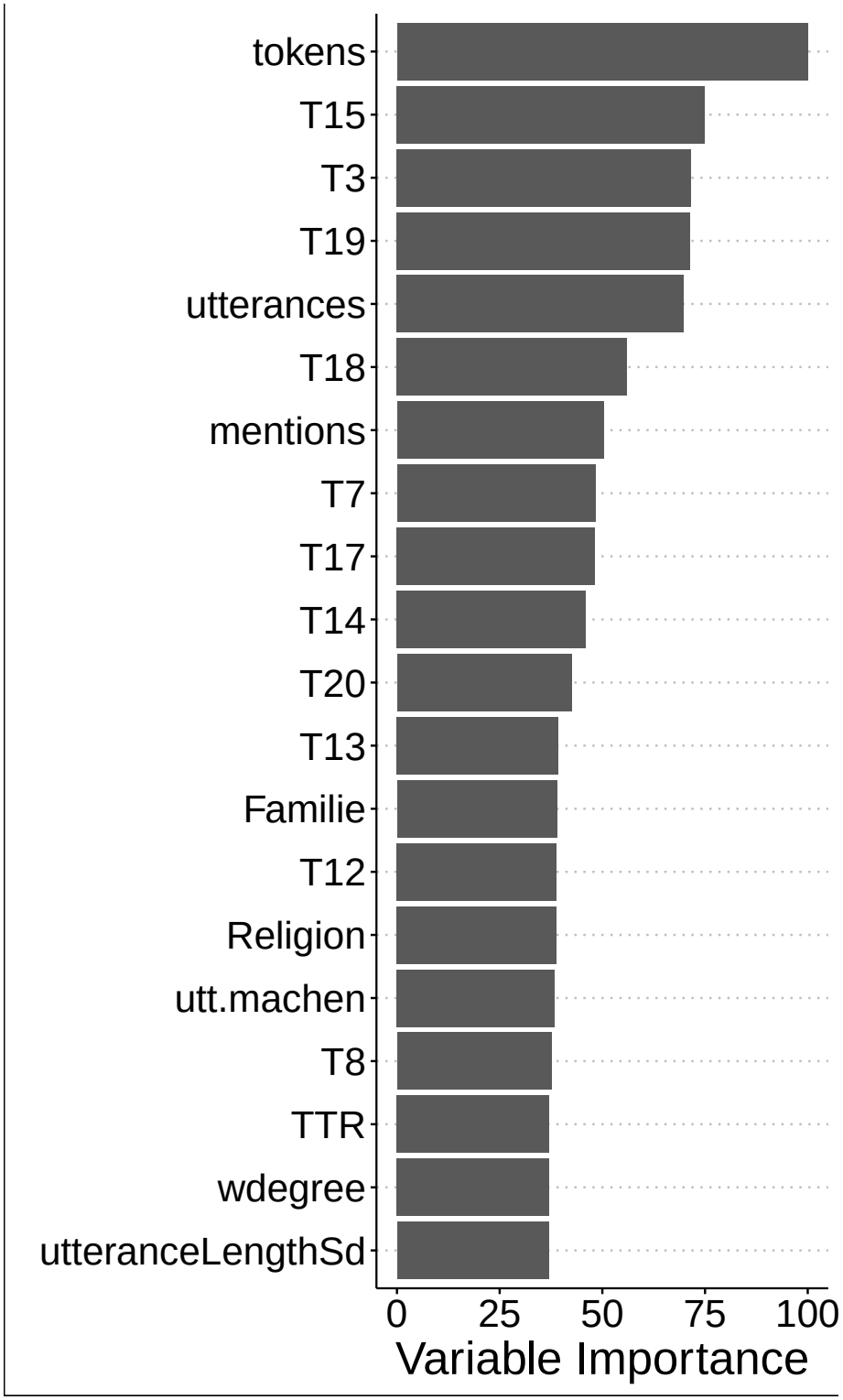


Figure 7.6.: Feature importance for the best model for SCHEMER-COREF.

between features that make other feature equally useful and therefore skew the feature importance results. Overall, it could be shown that coreference information helps to improve character type detection significantly. Hatzel and Biemann (2021) found similar results for scene segmentation of novels and could show that using coreference information induced into a transformer architecture helps to improve scene segmentation.

7.2. Summary

By enriching some of the existing features of Chapter 6 with coreference information and adding new features utilizing coreference information, it could be shown that adding such information helps to improve detection rates of character types even further. Critically, it helps to improve precision, with which all models across experiment struggled the most with. Therefore, the results show the necessity for including coreference information when performing character type detection. It could also be shown that the *mentions* feature outperforms the *tokens* feature, suggesting that to detect character types it is more important how often a character is talked about rather than how often a character talks.

Wir sind zu Ende, mein Herr!
(We are done, sir!)

Madame Wölbing in Johanna von Weißenthurn's "Das Manuscript"

8 Conclusions and Outlook

This chapter provides an overview of the outcomes of Chapters 4, 5, 6 and 7 as well as an outlook on tangible next steps and possible future research.

Chapter 4 presented a new resource for coreference resolution on German theatre plays, called GerDraCor-Coref. In several analyses, interesting differences between theatre plays to other types of text emerged. In an inter-annotator agreement study of four plays which were annotated by more than one annotator, the agreement scores were lower than for other agreement studies on coreference annotation. One could conclude from this that coreference annotation on plays is more difficult than for other genres. Statistical analysis on the annotations showed that GerDraCor-Coref contains on average more mentions but less entities than other types of text. Furthermore, GerDraCor-Coref contains more pronouns, but much less adjectives than other types of text. The low number of adjectives can possibly be traced back to the fact that for GerDraCor-Coref, adjectives are mainly used for character descriptions, while other types of texts additionally use adjectives for locations and organizations, which are not so prevalent in plays. An analysis of long coreference chains and chains with long distances between mentions showed that both phenomena occur more frequently in GerDraCor-Coref compared to other types of text. Lastly, an analysis of topics and characters involved in coreferent mentions showed that in coreference chains generic topics are most prevalent, and that on average, female characters mention male characters more than male characters mention female characters and that male characters mention other male characters more often than

female characters mention other female characters.

Chapter 5 presented experiments on GerDraCor-Coref regarding coreference resolution. To this end, a hybrid neural-rule-based system, DramaCoref, was introduced which is especially tailored towards resolving coreferences for plays. The neural component of the system was responsible for extracting mentions and performed slightly better than making use of a constituent parser. However, the fact that both approaches did not perform very well showed the difficulty of resolving mentions on plays. The rule-based component which resolved the coreferences between the extracted mentions performed comparable to other available systems on other types of texts and outperformed the systems on plays. An error analysis showed that especially in terms of recall, the system has further room for improvements.

Chapter 6 utilized the results of the two previous chapters by using the coreference informations for enhancing character type detection. To this end, experiments for carried out in order to automatically classify different forms of character types, namely protagonists, title characters and schemers. For this purpose, different random forest models were trained and evaluated. The models performed best on the task of protagonist detection and worst for the task of schemer detection, which can be labeled as the hardest of the three tasks. In all cases, *tokens*, topic model and centrality features were the most significant contributors for the models, suggesting the need for a multi-dimensional approach towards character type detection.

Chapter 7 presented experiments on character type detection with the addition of utilizing coreference information. Using the coreference information further pushed the results by up to fifteen percentage points, showing the need and usefulness of coreference information for the task.

Overall, it became apparent that

1. Theatre plays behave differently in terms of coreference compared to other types of texts, such as newspaper articles, radio news or radio interviews
2. Coreference information is helpful when automatically detecting different types of characters in theatre plays
3. Character type detection is a multi-dimensional endeavour that profits from features that cover very different properties of the character representation

This suggests a number of consequences for the field of character analysis as well as for coreference resolution in the context of CLS.

On the one hand, given the beneficial effect of coreference information on character type detection, strengthen efforts in annotating literary texts with coreference information seems expedient. On the other hand, NLP and CLS research can profit from working on dramatic texts, as the unique properties of these texts can help in developing new and customized methods not otherwise needed for newspaper data.
Future research might furthermore look into ways to utilize recent end-to-end neural network approaches for coreference resolution on dramatic texts, for which currently the available data size seems not to be sufficient.

A

Supplementary material: Inter-Annotator Agreement

The supplementary material presented in the following gives additional information on the IAA study carried out in Section 4.3.
Table A.1 gives an overview of the concrete distribution of plays and acts for the IAA study of Section 4.3.

Play	Act	Annotator ID
Friedrich Schiller: Die Räuber	I	1
		2
		3
	II	3
	III	1
		2
	IV	1
		2
	V	2
		3
Gotthold Ephraim Lessing: Emilia Galotti	I	4
		5
		6
		7
Gotthold Ephraim Lessing: Miß Sara Sampson	I	1
		8
		2
		3
	II	7
		1
		8
	III	7
		1
		3
	IV	1
		8
		2
	V	1
		3
Johann Christoph Gottsched: Der sterbende Cato	II	9
		2
		6

Table A.1.: Overview of plays that were used for computing IAA scores in Section 4.3, as well as the acts and annotation versions that were compared.

B

Supplementary material: Plays in GerDraCor-Coref

This section gives an overview of all plays present in GerDraCor-Coref, specifying author, title, years in which a play was written, firstly printed and/or premiered as well as the act or acts that have been annotated for coreference, given in Table B.1.

	Author	Title	Year Written	Year Printed	Year Premiered	Annotated Act(s)
1	Anzengruber, Ludwig	Der Meineidbauer	NA	1871	1871	I
2	Braun von Braunthal, Karl Johann	Faust	NA	1835	NA	IV
3	Cronegk, Johann Friedrich von	Der Mißtrauische	NA	1760	1766	II
4	Essig, Hermann	Überteufel	1906	1912	1923	I
5	Freytag, Gustav	Graf Waldemar	1847	1850	1848	IV
6	Gessner, Salomon	Evander und Alcimna	1762	1762	NA	I
7	Goethe, Johann Wolfgang	Egmont	1787	1788	1789	II
8	Goethe, Johann Wolfgang	Die natürliche Tochter	1803	1803	1803	I-V
9	Gottsched, Johann Christoph	Der sterbende Cato	1730	1732	1731	II
10	Grillparzer, Franz	Ein treuer Diener seines Herrn	1827	NA	NA	IV
11	Grillparzer, Franz	Sappho	1817	1819	1818	V
12	Hauptmann, Carl	Ephraims Breite	1899	1900	1900	V
13	Hauptmann, Carl	Musik	NA	1918	NA	IV
14	Hebbel, Friedrich	Gyges und sein Ring	NA	1856	1889	IV
15	Hensler, Karl Friedrich	Die Teufelsmühle am Wienerberg	NA	1799	NA	IV
16	Heyse, Paul	Don Juan's Ende	NA	1883	1884	IV
17	Hofmannsthal, Hugo von	Der Rosenkavalier	1910	1911	1911	I-III
18	Kleist, Heinrich von	Die Familie Schroffenstein	1802	1803	1804	I-V
19	Klinger, Friedrich Maximilian	Sturm und Drang	1776	1777	1777	V
20	Körner, Theodor	Zriny	1812	1814	1812	III
21	Kotzebue, August von	Die Indianer in England	1788	1790	1789	III
22	Krüger, Johann Christian	Die Candidaten oder Die Mittel zu einem Amte zu gelangen	NA	1748	1747	V
23	Lenz, Jakob Michael Reinhold	Der Hofmeister oder Vorteile der Privaterziehung	1772	1774	1778	I-V
24	Lessing, Gotthold Ephraim	Emilia Galotti	NA	1772	1772	I-V
25	Lessing, Gotthold Ephraim	Der Freigeist	1749	1755	NA	III
26	Lessing, Gotthold Ephraim	Nathan der Weise	NA	1779	1783	I-V
27	Lessing, Gotthold Ephraim	Miß Sara Sampson	NA	1755	1755	I-V
28	Ludwig, Otto	Der Erbförster	1849	1853	1850	V
29	Moser, Gustav von	Das Stiftungsfest	NA	1862	NA	I
30	Mühsam, Ludwig	Judas. Ein Arbeiterdrama	1920	1921	NA	I
31	Mylius, Christlob	Die Schäferinsel	NA	1749	NA	III
32	Nestroy, Johann	Einen Jux will er sich machen	1842	1844	1842	II

33	Pfeil, Johann Gottlob Benjamin	Lucie Woodvil	NA	1756	1756	I-V
34	Platen, August von	Die verhängnisvolle Gabel	1826	1826	NA	V
35	Quistorp, Theodor Johann	Der Hypochrondrist	NA	1745	NA	III
36	Rosenow, Emil	Kater Lampe	1900	1906	1902	IV
37	Rubiner, Ludwig	Die Gewaltlosen	NA	1919	1920	IV
38	Schiller, Friedrich	Die Braut von Messina oder Die feindlichen Brüder	1803	1803	1803	I-V
39	Schiller, Friedrich	Die Piccolimini	1798	NA	1799	II
40	Schiller, Friedrich	Die Räuber	1780	1781	1882	I-V
41	Schink, Johann Friedrich	Hanswurst von Salzburg mit dem hölzernen Gat	NA	1778	NA	II
42	Schlegel, August Wilhelm	Jon	NA	1803	1802	III
43	Wagner, Heinrich Leopold	Die Reue nach der That	NA	1775	1775	V
44	Wedekind, Frank	König Nicolo oder So ist das Leben	1901	1902	1902	I
45	Wedekind, Frank	Die Büchse der Pandora	NA	1902	1904	I
46	Weißenthurn, Johanna von	Das Manuscript	NA	1817	NA	V
47	Wildenbruch, Ernst von	Die Quitzows	NA	1888	1888	IV

Table B.1.: All plays included in GerDraCor-Coref. If a year in which a play was either written, printed or premiered is not known, *NA* is given. Annotated acts are given in roman numerals and ranges of annotated acts are indicated by a hyphen.

C

Supplementary material: Translations

This section provides English translations for the German language snippets of several plays. All translations were carried out by the author of this thesis.

First Scene
Franconia. Hall in the Moor's castle.
Franz. The old Moor.

Franz.
But are you well, father? You look rather pale.

The Old Moor.
Quite well, my son — what did you have to tell me?

Franz.
The mail has arrived — a letter from our correpsondent in Leipzig —

The Old Moor
eagerly.
Any news of my son Karl?

Franz.
Hm! Hm! — So it is. But I'm afraid — I don't know — whether I — your health? — Are you really quite well, my father?

The Old Moor.
Like a fish in water! He writes about my son? — Why are you so worried? You have asked me twice.

FRANZ.
If you are ill — if you have the slightest suspicion of becoming ill, let me — let me speak to you at a more appropriate time. *Half to himself.} This news is not meant for a fragile body.*

Figure C.1.: English Translation of Figure 1.1.

First Act

First Scene

The setting shows a hall in [the inn]$_1$.

[Sir William Sampson]$_2$ and Waitwell enter, dressed in traveling clothes.

[SIR WILLIAM]$_2$.
[My]$_2$ daughter, here? Here in [this wretched inn]$_1$?

Figure C.2.: English Translation of Figure 1.2.

[OCTAVIAN]$_1$.
[The field marshal]$_2$?

[MARSCHALLIN]$_3$.
There was [a noise] in the [courtyard]$_4$ of [[horses] and [people]] and [he]$_2$ was there.
[I]$_3$ was suddenly awake with [fright], no, just look,
look how childish [I]$_3$ am: [I]$_3$ can still hear [the rumble in the [courtyard]$_4$]$_5$.
[I]$_3$ can't get [it]$_5$ out of [the ear]. Can [you]$_1$ hear [anything]?

[OCTAVIAN]$_1$.
Yes, of course [I]$_1$ hear [something]$_6$, but does [it]$_6$ have to be [[your]$_3$ husband]$_2$!
Just [you]$_3$ think, where [he]$_2$ is: in [Raitzenland]$_7$, still beyond [Esseg].

[MARSCHALLIN]$_3$.
Is [that]$_7$ really far for sure?

16 Well, then [it]$_5$ will be [something else]. Then it's alright.

Figure C.3.: English Translation of Figure 4.2.

1 Sylvius.
2 Agnes, [where is Philipp]$_1$?
3
4 Agnes.
5 Dear God, I tell [it]$_1$ to you every day,
6 And write [it]$_1$ to you on a sheet, were you not blind.
7 Come hither, I'll [write [it]$_1$ in your hand]$_2$.
8
9 Sylvius.
10 Does [that]$_2$ help?
11
12 Agnes.
13 [[It]$_2$ helps]$_3$, believe me [that]$_3$.
14
15 Sylvius.
16 Oh, [it]$_2$ doesn't help.
17
18 Agnes.
19 I mean,
20 From forgetting.

Figure C.4.: English Translation of Figure 4.3.

1 Albertine.
2 Mother! did you not notice in his countenance any of the fine mockery with
which [the superior man]$_1$ is so fond of ridiculing the attempts of
women – even if [he]$_1$ cannot quite condemn them?
3
4 Madame Wölbing.
5 No, the recognition of your talent came from the heart.
6
7 Albertine.
8 *quickly.*

I do not fear his rebuke, only his ridicule! He shall be strict, but he shall be honest. His rebuke shall instruct me, but his wit shall not make me bitter. If you can expect that from him, bring him.

Figure C.5.: English Translation of Figure 4.4.

[Sara]$_1$.

[...]

So [I]$_1$ should leave [my]$_1$ homeland as [a criminal]$_{2,generic,predicate}$? And as [such]$_{2,generic,predicate}$, you think, [I]$_1$ would have enough courage to confide [myself]$_1$ to the sea?

Figure C.6.: English Translation of Figure 4.5.

Waitwell.

[...]

[Evil people]$_1$ always seek the dark because [they]$_1$ are [evil people]$_{predicate}$. But what good is it to [them]$_1$ if [they]$_1$ hide from the whole world?

Figure C.7.: English Translation of Figure 4.6.

[Saladin]$_1$.
Look at that! so [you]$_2$ would have lost with diligence, if [you]$_2$
Had lost, [little sister]$_2$?

[Sittah]$_2$.
At least it can't be that [your]$_1$
Generosity, [[my]$_2$ dear brother]$_1$,
Is to blame for [my]$_2$ not learning to play better.

[Saladin]$_1$.
[We]$_3$ are deviating from the game. Make an end!

[...]

[Sittah]$_2$.
Ah, so
[You]$_1$ wish but to blunt the sting of loss.
Enough, [you]$_1$ have been scattered; and more than [I]$_2$.

[Saladin]$_1$.
Than [you]$_2$? What would have distracted [you]$_2$?

[Sittah]$_2$.
[Your]$_1$
Distraction certainly not! – [O Saladin]$_1$,
When will [we]$_3$ play so diligently again!

Figure C.8.: English Translation of Figure 4.7.

Second act

Seventh scene

Marwood.
You remind me not to race against the right person. The father must go ahead! He must already be in that world, when his daughter's spirit, with a thousand sighs, follows him. *She goes at him with a dagger, which she snatches from her bosom.* Therefore die, traitor!

[...]

Fourth act

Third scene

Mellefont.
Look, this murderer's iron I tore from her hand, *He shows him the dagger he took from Marwood.* when she wanted to pierce my heart with it in the most terrible rage.

[...]

Fifth act

Tenth scene

Mellefont.
Not like this, sir! This saint commanded more than the human nature is capable of! You cannot be my father. – See, sir, *By pulling the dagger from his bosom.* this is the dagger Marwood jerked at me today. To my misfortune, I had to disarm her. If I had fallen as the guilty victim of her jealousy, Sara would still be alive. You would still have your daughter, and you would have her without Mellefont. It is not up to me to undo what has happened; but to punish me for what has happened – that is up to me! *He stabs himself and falls down at Sara's chair.*

Figure C.9.: English Translation of Figure 4.11.

Seventh scene
Emilia. Odoardo.

Emilia.
How? You here, [my father]$_0$? – And only you? – And [my mother]$_1$? not here? – And [the Count]$_2$? not here? – And you so restless, [my father]$_0$?

Odoardo.
And you so calm, [my daughter]$_3$?

Emilia.
Why not, [my father]$_0$? – Either nothing is lost: or all. To be able to be calm, and to have to be calm: does it not come down to the same thing?

Odoardo.
But what do you think is the case?

Emilia.
That all is lost; – and that we must be calm, [my father]$_0$.

ODOARDO.
And you would be calm because you must be calm? — Who are you? A girl? and [my daughter]$_3$? So the man and [the father]$_4$ should be ashamed before you? — But let me hear: what do you call all lost? — that [the Count]$_2$ is dead?

EMILIA.
And why he is dead! Why! — Ha, is it true then, [my father]$_0$? Is the whole dreadful story I read in [my mother's]$_5$ wet and wild eye true? — Where is [my mother]$_1$? Where has she gone, [my father]$_0$?

Figure C.10.: English Translation of Figure 5.1.

VRONI.
The seal is all crumbled anyway, [I]$_1$ will open it!

JAKOB.
Do it, it's up to you now!

VRONI *opens the letter.*
It's from father's brother, the Kreuzweghofbauer! — Dear Lord!

JAKOB.
You're scaring me!

VRONI.
For God's sake, brother, go ahead, just go ahead, what he wrote to father: » Dear Jakob! [I]$_1$ have received your will, in which you appoint the citizens Vroni and her two children as heirs to all your possessions. It is not nice that you're giving [me]$_1$ and [my]$_1$ children such a raw deal ...«

Figure C.11.: English Translation of Figure 5.10.

EMILIA.
It is true, with a hairpin I should —*She runs her hand through her hair, looking for one, and gets hold of the rose.* [You]$_1$ are still here? — Get down, [you]$_1$! [You]$_1$ do not belong in the hair of someone, — as my father wants me to be!

ODOARDO.
O, my daughter! –

EMILIA.
O, my father, if I could guess [you]$_1$! – But no; [you]$_1$ don't want that either. Why else did [you]$_1$ hesitate?

Figure C.12.: English Translation of Figure 5.11.

First act
First scene

[...]

FRANZ.
The mail has arrived – [a letter from our correspondent in Leipzig]$_1$ –

[...]

FRANZ *takes [the letter]$_1$ out of his pocket.*

[...]

First act
Second scene

[...]

MOOR *flies toward him.*
Brother! Brother! [the letter]$_1$! [the letter]$_1$!

Figure C.13.: English Translation of Figure 5.12.

PHILIPP.
With your permission, [my lord]$_1$! that can't be.

HERR ORGON.

It can't be! And why?

[...]

Damon.
I can't possbibly stay longer, I would betray myself too much. Heavens! how charming she is!

He wants to leave.
Lisette.
Shh! Shh! [my lord]$_1$, where are you going?

Figure C.14.: English Translation of Figure 5.13.

Mellefont.
You're disturbing me, Norton!

Norton.
Pardon me [my lord]$_1$ – *By wanting to go back.*

[...]

Norton.
Might well make you anxious, but not downcast. – Something else troubles you. And I will gladly have erred, if you would not rather [the father]$_1$ were not yet reconciled. The prospect of a position that is so ill–suited to your way of thinking – –

Figure C.15.: English Translation of Figure 5.14.

Emilia.
It is true, with a hairpin I should –*She runs her hand through her hair, looking for one, and gets hold of the rose.* You are still here? – Get down! You do not beling in the hair of someone, – as [my father]$_1$ wants me to be!

Odoardo.
O, [my daughter]$_1$! –

Figure C.16.: English Translation of Figure 5.15.

THE PRINCE
after some silence, under which he looks at the body with horror and despair, to Marinelli.

Here! pick it up. — Well? You are thinking of yourself? — Wretch! — *By snatching the dagger from his hand.* No, your blood shall not mix with this blood. — Go, hide yourself forever! — Go! I say. — $[\text{God}]_1$! $[\text{God}]_1$! — Is it not enough, to the misfortune of many, that princes are human: must $[\text{devils}]_1$ also disguise themselves as friends?

Figure C.17.: English Translation of Figure 5.16.

Bibliography

Algee-Hewitt, Mark (2017). "Distributed Character: Quantitative Models of the English Stage, 1550–1900". In: *New Literary History* 48.4, pp. 751–782. DOI: 10.1353/nlh.2017.0038.

Alt, Peter-André (Sept. 2004). "Dramaturgie des Störfalls. Zur Typologie des Intriganten im Trauerspiel des 18. Jahrhunderts". In: *Internationales Archiv für Sozialgeschichte der deutschen Literatur* 29.1, pp. 1–28. DOI: 10.1515/IASL.2004.1.1.

Andresen, Melanie, Katharina Krüger, Michael Vauth, and Heike Zinsmeister (Dec. 2018). "Can we describe a literary character by its explicit attributions based on syntactic annotation?" In: *Book of Abstracts of the European Association for Digital Humanities (EADH)*. Galway, Ireland. URL: https://eadh2018.exordo.com/files/papers/108/final_draft/Abstract_EADH.pdf.

Andresen, Melanie and Michael Vauth (Aug. 2018a). "Added Value of Coreference Annotation for Character Analysis in Narratives". In: *Proceedings of the Workshop on Annotation in Digital Humanities (annDH)*. Sofia, Bulgaria, pp. 1–6. URL: http://ceur-ws.org/Vol-2155/andresen.pdf.

Aralikatte, Rahul and Anders Søgaard (May 2020). "Model-based annotation of coreference". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*. Marseille, France, pp. 74–79. URL: https://aclanthology.org/2020.lrec-1.9.

Aristoteles (1982). *Poetik*. Trans. by Manfred Fuhrmann. Stuttgart, Germany: Reclam.

Artstein, Ron and Massimo Poesio (Dec. 2008). "Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4, pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004.

Bagga, Amit and Breck Baldwin (May 1998). "Algorithms for Scoring Co-Reference Chains". In: *The Linguistic Co-Reference Workshop at The First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain, pp. 563–566.

Baldwin, Breck (July 1997). "CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources". In: *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*. Madrid, Spain, pp. 38–45. URL: https://aclanthology.org/W97-1306.

Baldwin, Breck, Tom Morton, Amit Bagga, Jason Baldridge, Raman Chandraseker, Alexis Dimitriadis, Kieran Snyder, and Magdalena Wolska (Apr. 1998). "Description of the UPENN CAMP System as Used for Coreference". In: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, VA, USA. URL: https://aclanthology.org/M98-1022.

Bamman, David, Olivia Lewke, and Anya Mansoor (May 2020). "An Annotated Dataset of Coreference in English Literature". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 44–54. URL: https://www.aclweb.org/anthology/2020.lrec-1.6.

Bamman, David, Brendan O'Connor, and Noah A. Smith (Aug. 2013). "Learning Latent Personas of Film Characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sofia, Bulgaria, pp. 352–361. URL: https://aclanthology.org/P13-1035.

Bamman, David, Sejal Popat, and Sheng Shen (June 2019). "An Annotated Dataset of Literary Entities". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, MN, USA, pp. 2138–2144. DOI: 10.18653/v1/N19-1220. URL: https://www.aclweb.org/anthology/N19-1220.

Bamman, David, Ted Underwood, and Noah A. Smith (June 2014). "A Bayesian Mixed Effects Model of Literary Character". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD, USA, pp. 370–379. DOI: 10.3115/v1/P14-1035. URL: https://aclanthology.org/P14-1035.

Beauchamp, Murray (1965). "An improved index of centrality". In: *Behavioral Science* 10.2, pp. 161–163. DOI: 10.1002/bs.3830100205.

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (2020). *Longformer: The Long-Document Transformer*. Version 2. DOI: 10.48550/arXiv.2004.05150. arXiv: arXiv:2004.05150v2 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.

Björkelund, Anders, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer (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)*. Reykjavík, Iceland, pp. 3222–3228.

Björkelund, Anders and Jonas Kuhn (June 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). Baltimore, MD, USA, pp. 47–57. URL: https://aclweb.org/anthology/P14-1005.

Blei, David, Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Blessing, Andre, Nora Echelmeyer, Markus John, and Nils Reiter (Aug. 2017). "An End-to-end Environment for Research Question-Driven Entity Extraction and Network Analysis". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH)*. Vancouver, Canada, pp. 57–67. DOI: 10.18653/v1/W17-2208. URL: https://aclanthology.org/W17-2208.

Bohnet, Bernd and Joakim Nivre (July 2012). "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*. Jeju Island, Korea, pp. 1455–1465. URL: https://aclanthology.org/D12-1133.

Borthen, Kaja (Aug. 2004). "Predicative NPs and the annotation of reference chains". In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. Geneva, Switzerland, pp. 1175–1178. DOI: 10.3115/1220355.1220524. URL: https://aclanthology.org/C04-1169.

Breiman, Leo (Oct. 2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf.

Bußmann, Hadumod, ed. (2008). *Lexikon der Sprachwissenschaft*. 4th ed. Stuttgart: Alfred Kröner Verlag.

Cai, Jie and Michael Strube (Sept. 2010). "Evaluation Metrics For End-to-End Coreference Resolution Systems". In: *Proceedings of the SIGDIAL 2010 Conference*. Tokyo, Japan, pp. 28–36. URL: https://aclanthology.org/W10-4305.

Chamberlain, Jon, Massimo Poesio, and Udo Kruschwitz (May 2016). "Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, pp. 2039–2046. URL: https://www.aclweb.org/anthology/L16-1323.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Chen, Bin, Jian Su, and Chew Lim Tan (Aug. 2010). "A Twin-Candidate Based Approach for Event Pronoun Resolution using Composite Kernel". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling)*. Beijing, China, pp. 188–196. URL: https://aclanthology.org/C10-1022.

Chen, Wei-Te and Will Styler (June 2013). "Anafora: A Web-based General Purpose Annotation Tool". In: *Proceedings of the 2013 NAACL HLT Demonstration Session*. Atlanta, GA, USA, pp. 14–19. URL: https://aclanthology.org/N13-3004.

Chomsky, Noam (1981). *Lectures on Government and Binding*. 1st ed. Vol. 9. Studies in Generative Grammar. Dordrecht: Foris Publications Holland.

– (1993). *Lectures on Government and Binding. The Pisa Lectures*. Ed. by Jan Koster and Henk van Riemsdijk. 7th ed. Vol. 9. Studies in Generative Grammar. Berlin/New York: Mouton de Gruyter. ISBN: 9783110141313. DOI: 10.1515/9783110884166.

Clark, Herbert H. (1975). "Bridging". In: *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing (TINLAP)*. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 169–174.

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning (Apr. 2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia (Online). URL: https://openreview.net/forum?id=r1xMH1BtvB.

Clark, Kevin and Christopher D. Manning (Aug. 2016). "Improving coreference resolution by learning entity-level distributed representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany, pp. 643–653. URL: https://www.aclweb.org/anthology/P16-1061.

Cohan, Steven (1983). "Figures beyond the Text: A Theory of Readable Character in the Novel". In: *NOVEL: A forum on Fiction* 17.1, pp. 5–27. DOI: 10.2307/1344821.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Crystal, David (2008). *A Dictionary of Linguistics and Phonetics*. Ed. by David Crystal. 6th ed. The Language Library. Malden, Oxford, Carlton: Blackwell Publishing.

Da, Nan Z. (2019). "The Computational Case against Computational Literary Studies". In: *Crititcal Inquiry* 45.3, pp. 601–639. DOI: 10.1086/702594.

Daumé III, Hal and Daniel Marcu (Oct. 2005). "A Large-Scale Exploration of Effective Global Features for a Joint Entity Detection and Tracking Model". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in*

Natural Language Processing (EMNLP-HLT). Vancouver, British Columbia, Canada, pp. 97–104. URL: https://aclanthology.org/H05-1013.

Denis, Pascal and Jason Baldridge (Mar. 2009). "Global joint models for coreference resolution and named entity classification". In: *Procesamiento del Lenguaje Natural* 42, pp. 87–96. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2806/1305.

Dennerlein, Katrin, Thomas Schmidt, and Christian Wolff (July 2023). "Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century". In: *Digital Scholarship in the Humanities* 38.4, pp. 1466–1481. DOI: 10.1093/llc/fqad046.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, MN, USA, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Dipper, Stefanie, Anke Lüdeling, and Marc Reznicek (Nov. 2013). "NoSta-D: A Corpus of German Non-Standard Varieties". In: *Non-Standard Data Sources in Corpus-Based Research*. Ed. by Marcos Zampieri and Sascha Diwersy. Vol. 5. ZSM-Studien. Düren, Maastricht: Shaker, pp. 69–76. URL: https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/nosdac13.pdf.

Dönicke, Tillmann, Florian Barth, Hanna Varachkina, and Caroline Sporleder (Sept. 2022). "MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy". In: *Proceedings of the 18th Conference on Natural Language Processing (KONVENS)*. Potsdam, Germany, pp. 8–15. URL: https://aclanthology.org/2022.konvens-1.2.

Dudenredaktion (n.d.). *„Zeitung" on Duden online*. Last access: 2021-01-25T23:20:00. URL: https://www.duden.de/node/209268/revision/209304.

Eckart de Castilho, Richard and Iryna Gurevych (Aug. 2014). "A broad-coverage collection of portable NLP components for building shareable analysis pipelines". In: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*. Dublin, Ireland, pp. 1–11. DOI: 10.3115/v1/W14-5201. URL: http://www.aclweb.org/anthology/W14-5201.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Annette Frank, and Chris Biemann (Dec. 2016). "A Web-based

Tool for the Integrated Annotation of Semantic and Syntactic Structures". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan, pp. 76–84. URL: https://aclanthology.org/W16-4011.

Eschenbach, Carola, Christopher Habel, Michael Herweg, and Klaus Rehkämper (Apr. 1989). "Remarks on Plural Anaphora". In: *Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Manchester, England, pp. 161–167. URL: https://aclanthology.org/E89-1022.

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 of the ACL*. Association for Computational Linguistics, pp. 363–370. URL: https://www.aclweb.org/anthology/P05-1045.

Finlayson, Mark A. (Feb. 2012). "Learning Narrative Structure from Annotated Folktales". PhD thesis. Cambridge, MA, USA: Massachusetts Institute of Technology. URL: http://hdl.handle.net/1721.1/71284.

– (June 2017). "ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory". In: *Digital Scholarship in the Humanities* 32.2, pp. 284–300. DOI: 10.1093/llc/fqv067.

Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke (Aug. 2017). "Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts". In: *Book of Abstracts of the DH2017 conference*. Montréal, Canada. URL: https://dh2017.adho.org/abstracts/071/071.pdf.

Fischer, Frank, Peer Trilcke, Christopher Kittel, Carsten Milling, and Daniil Skorinkin (June 2018). "To Catch a Protagonist: Quantitative Dominance Relations in German-Language Drama (1730–1930)". In: *Book of Abstracts of DH 2018*. Mexico City, Mexico, pp. 193–201. URL: https://dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf#page=193.

Fleiss, Joseph L. (1971). "Measuring Nominal Scale Agreement Among Many Raters". In: *Psychological Bulletin* 76.5, pp. 378–382.

Freeman, Linton C. (Mar. 1977). "A Set of Measures of Centrality Based on Betweenness". In: *Sociometry* 40.1, pp. 35–41. DOI: 10.2307/3033543.

– (1978/1979). "Centrality in social networks. Conceptual clarification". In: *Social Networks* 1.3, pp. 215–239. DOI: 10.1016/0378-8733(78)90021-7.

Gaizauskas, R., T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks (Nov. 1995). "University of Sheffield: Description of the LaSIE System as Used for MUC-6". In:

Proceddings of the Sixth Message Understanding Conference (MUC-6). Columbia, MD, USA, pp. 207–220. URL: https://aclanthology.org/M95-1017.

Halliday, M. A. K. and Ruqaiya Hasan (2013). *Cohesion in English*. Ed. by Randolph Quirk. Vol. 9. English Language Series. First published 1976 by Pearson Education Limited. London, New York: Routledge. DOI: https://doi.org/10.4324/9781315836010.

Hamp, Birgit and Helmut Feldweg (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*. Madrid, Spain, pp. 9–15. URL: https://www.aclweb.org/anthology/W97-0802.

Han, Sooyoun, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi (Nov. 2021). "FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View". In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*. Punta Cana, Dominican Republic, pp. 24–35. DOI: 10.18653/v1/2021.crac-1.3. URL: https://aclanthology.org/2021.crac-1.3.

Hatzel, Hans Ole and Chris Biemann (Sept. 2021). "LTUHH@STSS: Applying Coreference to Literary Scene Segmentation". In: *Proceedings of the Shared Task on Scene Segmentation co-located with the 17th Conference on Natural Language Processing*. Düsseldorf, Germany, pp. 29–34. URL: https://ceur-ws.org/Vol-3001/paper3.pdf.

Hicke, Rebecca M. M. and David Mimno (2024). *[Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs*. Version 1. DOI: 10.48550/arXiv.2401.17922. arXiv: arxiv:2401.17922 [cs.CL]. URL: https://arxiv.org/abs/2401.17922.

Ho, Tin Kam (Aug. 1995). "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Montreal, QC, Canada, pp. 278–282. DOI: 10.1109/ICDAR.1995.598994.

– (Aug. 1998). "The random subspace method for constructing decision forests". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8, pp. 832–844. DOI: 10.1109/34.709601.

Hobbs, Jerry R. (1978). "Resolving Pronoun References". In: *Readings in Natural Language Processing*. Ed. by Barbara J. Grosz, Karen Sparck Jones, and Bonnie Lynn Webber. Los Altos, CA, USA: Morgan Kaufmann Publishers, Inc., pp. 339–352.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hou, Yufang (2016). "Unrestricted Bridging Resolution". PhD thesis. Heidelberg University.

Hubert, Lawrence and Phipps Arabie (Dec. 1985). "Comparing partitions". In: *Journal of Classification* 2, pp. 193–218. DOI: 10.1007/BF01908075.

Hudson, George and Noura Al Moubayed (June 2022). "MuLD: The Multitask Long Document Benchmark". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 3675–3685. URL: https://aclanthology.org/2022.lrec-1.392.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III (June 2016). "Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. San Diego, CA, USA, pp. 1534–1544. DOI: 10.18653/v1/N16-1180. URL: https://www.aclweb.org/anthology/N16-1180.

Jahan, Labiba, Geeticka Chauhan, and Mark A. Finlayson (Oct. 2017). "Building on Word Animacy to Determine Coreference Chain Animacy in Cultural Narratives". In: *The Workshops of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*. Snowbird, UT, USA, pp. 198–203. URL: https://aaai.org/ocs/index.php/AIIDE/AIIDE17/paper/view/15867.

Jahan, Labiba and Mark A. Finlayson (June 2019). "Character Identification Refined: A Proposal". In: *Proceedings of the First Workshop on Narrative Understanding*. Minneapolis, MN, USA, pp. 12–18. DOI: 10.18653/v1/W19-2402. URL: https://aclanthology.org/W19-2402.

Jahan, Labiba, Rahul Mittal, and Mark A. Finlayson (Nov. 2021). "Inducing Stereotypical Character Roles from Plot Structure". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic, pp. 492–497. DOI: 10.18653/v1/2021.emnlp-main.39. URL: https://aclanthology.org/2021.emnlp-main.39.

Jahan, Labiba, Rahul Mittal, W. Victor H. Yarlott, and Mark A. Finlayson (Dec. 2020). "A Straightforward Approach to Narratologically Grounded Character Identification". In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain (Online), pp. 6089–6100. DOI: 10.18653/v1/2020.coling-main.536. URL: https://aclanthology.org/2020.coling-main.536.

Jannidis, Fotis (Jan. 2020). "On the perceived complexity of literature. A response to Nan Z. Da". In: *Journal of Cultural Analytics* 5.1. DOI: 10.22148/001c.11829.

Jannidis, Fotis, Isabella Reger, Markus Krug, Lukas Weimer, Luisa Macharowsky, and Frank Puppe (July 2016). "Comparison of Methods for the Identification of Main Characters in German Novels". In: *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland, pp. 578–582. URL: https://dh2016.adho.org/abstracts/297.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans". In: *Transactions of the Association for Computational Linguistics (TACL)* 8, pp. 64–77. DOI: 10.1162/tacl_a_00300. URL: https://aclanthology.org/2020.tacl-1.5.

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (Nov. 2019). "BERT for Coreference Resolution: Baselines and Analysis". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 5803–5808. DOI: 10.18653/v1/D19-1588. URL: https://aclanthology.org/D19-1588.

Kallenbach, Ulla and Annelis Kuhlmann (2018). "Towards a Spectatorial Approach to Drama Analysis". In: *Nordic Theatre Studies* 30.2, pp. 22–39. DOI: 10.7146/nts.v30i2.112950. URL: https://tidsskrift.dk/nts/article/view/112950/161733.

Kamp, Hans (Oct. 2021). "Sharing real and fictional reference". In: *The Language of Fiction*. Ed. by Emar Maier and Andreas Stokke. Oxford: Oxford University Press. Chap. 3, pp. 37–87. DOI: 10.1093/oso/9780198846376.003.0003.

Kamp, Hans and Uwe Reyle (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Part 2*. 1st ed. Vol. 42. Studies in Linguistics and Philosophy. Dordrecht: Kluwer Academic Publishers. Chap. The Plural, pp. 305–482.

Kaplan, Dain, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga (2011). "Slate – A tool for creating and maintaining annotated corpora". In: *Journal for Language Technology and Computational Linguistics* 26.2, pp. 89–101. DOI: 10.21248/jlcl.26.2011.149.

Klein, Dan and Christopher D. Manning (2002). "Fast exact inference with a factored model for natural language parsing". In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 15, pp. 3–10. URL: https://papers.nips.cc/paper/2325-fast-exact-inference-with-a-factored-model-for-natural-language-parsing.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (June 2018). "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: *Proceedings of the 27th International*

Conference on Computational Linguistics: System Demonstrations (COLING). Santa Fe, NM, USA, pp. 5–9. URL: https://aclanthology.org/C18-2002.

Kolhatkar, Varada, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister (2018). "Anaphora with Non-nominal Antecedents in Computational Linguistics: A Survey". In: *Computational Linguistics* 44.3, pp. 547–612. DOI: 10.1162/coli_a_00327.

Köppe, Tilmann (2020). "Reference in Literature/Literary Studies". In: *Narrative Factuality: A Handbook*. Ed. by Monika Fludernik and Marie-Laure Ryan. Berlin, Boston: De Gruyter, pp. 259–266. DOI: 10.1515/9783110486278-016.

Krautter, Benjamin (June 2018). "Quantitative microanalysis? Different methods of digital drama analysis in comparison". In: *Book of Abstracts of DH 2018*. Mexico City, Mexico, pp. 225–228. URL: https://dh2018.adho.org/en/quantitative-microanalysis-different-methods-of-digital-drama-analysis-in-comparison/.

– (Aug. 2023). "Kopräsenz-, Koreferenz- und Wissens-Netzwerke. Kantenkriterien in dramatischen Figurennetzwerken am Beispiel von Kleists Die Familie Schroffenstein (1803)". In: *Journal of Literary Theory* 17.2, pp. 261–289. DOI: 10.1515/jlt-2023-2012.

Krautter, Benjamin and Janis Pagel (Mar. 2019). "Klassifikation von Titelfiguren in deutschsprachigen Dramen und Evaluation am Beispiel von Lessings "Emilia Galotti"". In: *Book of Abstracts of DHd*. Frankfurt am Main, Germany, pp. 160–164. DOI: 10.5281/zenodo.4622195. URL: https://elib.uni-stuttgart.de/bitstream/11682/10382/1/KRAUTTER_Benjamin_Klassifikation_von_Titelfiguren_in_deutsch.pdf.

– (2024, to appear). "The Schemer in German Drama. Identification and Quantitative Characterization". In: *Computational Drama Analysis. Reflecting Methods and Interpretations*. Ed. by Melanie Andresen and Nils Reiter. Berlin/Boston: De Gruyter.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand (Nov. 2018). "Titelhelden und Protagonisten — Interpretierbare Figurenklassifikation in deutschsprachigen Dramen". In: *LitLab Pamphlets* 7. URL: https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf.

– (2020). ""[E]in Vater, dächte ich, ist doch immer ein Vater". Figurentypen und ihre Operationalisierung". In: *ZfdG* 5.7. DOI: 10.17175/2020_007. URL: http://www.zfdg.de/2020_007.

– (Dec. 2022). "Properties of Dramatic Characters: Automatic Detection of Gender, Age, and Social Status". In: *Computational Stylistics in Poetry, Prose, and Drama*. Ed. by Anne-Sophie Bories, Petr Plecháč, and Pablo Ruiz Fabo. Berlin/Boston: De Gruyter, pp. 179–202. DOI: 10.1515/9783110781502-010.

Krug, Markus, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimer (June 2015). "Rule-based Coreference Resolution in German Historic Novels". In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, CO, USA, pp. 98–104. DOI: 10.3115/v1/W15-0711. URL: https://aclweb.org/anthology/W15-0711.

Krug, Markus, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus, and Fotis Jannidis (2018). "Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]". In: *DARIAH-DE Working Papers* 27. URL: http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2018-27.pdf.

Kübler, Sandra and Heike Zinsmeister (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Academic. DOI: 10.5040/9781472593573.

Lasnik, Howard (1989). "Remarks on Coreference". In: *Essays on Anaphora*. Vol. 16. Studies in Natural Language and Linguistic Theory. First published 1976 in Linguistic Analysis 2. Dordrecht: Springer, pp. 90–109. DOI: 10.1007/978-94-009-2542-7_4.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2013). "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules". In: *Computational Linguistics* 39.4, pp. 885–916. DOI: 10.1162/COLI_a_00152. URL: https://www.aclweb.org/anthology/J13-4004.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (June 2011). "Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*. Portland, OR, USA, pp. 28–34. URL: https://www.aclweb.org/anthology/W11-1902.

Lee, James and Jason Lee (2017). "Shakespeare's Tragic Social Network; or Why All the World's a Stage". In: *Digital Humanities Quarterly* 11.2. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000289/000289.html.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark, pp. 188–197.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 2018). "Higher-order Coreference Resolution with Coarse-to-fine Inference". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, LA, USA, pp. 687–692.

Luo, Xiaoqiang (Oct. 2005). "On Coreference Resolution Performance Metrics". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*. Vancouver, British Columbia, Canada, pp. 25–32. URL: https://aclanthology.org/H05-1004.

Luo, Xiaoqiang, Sameer S. Pradhan, Marta Recasens, and Eduard Hovy (June 2014). "An Extension of BLANC to System Mentions". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD, USA, pp. 24–29. DOI: 10.3115/v1/P14-2005. URL: https://aclanthology.org/P14-2005.

Lyons, John (1977). *Semantics*. 1st ed. Vol. 2. London, New York, Melbourne: Cambridge University Press.

Maier, Emar (2017). "Fictional Names in Psychologistic Semantics". In: *Theoretical Linguistics* 43.1-2, pp. 1–45. DOI: 10.1515/tl-2017-0001.

Maximova, Daria and Frank Fischer (Dec. 2018). "A Quantitative Study of Stage Directions in Russian Drama". In: *Book of Abstracts of the European Association for Digital Humanities (EADH)*. Galway, Ireland. URL: https://eadh2018.exordo.com/files/papers/79/final_draft/Stage_Directions_for_EADH_Conference.pdf.

McIntyre, Dan (Nov. 2008). "Integrating multimodal analysis and the stylistics of drama: a multimodal perspective on Ian McKellen's Richard III". In: *Language and Literature* 17.4, pp. 309–334. DOI: 10.1177/0963947008095961.

Mead, Gerald (1990). "The Representation of Fictional Character". In: *Style* 24.3, pp. 440–452. URL: https://www.jstor.org/stable/42945872.

Mihalcea, Rada and Andras Csomai (Nov. 2007). "Wikify!: linking documents to encyclopedic knowledge". In: *Proceedings of the sixteenth ACM Conference on Information and Knowledge Management (CIKM)*. Lisboa, Portugal, pp. 233–242. DOI: 10.1145/1321440.1321475.

Moosavi, Nafise Sadat and Michael Strube (Aug. 2016). "Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 632–642. DOI: 10.18653/v1/P16-1060. URL: https://aclanthology.org/P16-1060.

Moretti, Franco (2011). "Network Theory, Plot Analysis". In: *Pamphlets of the Stanford Literary Lab* 2, pp. 2–11. URL: https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

Müller, Christoph and Michael Strube (July 2003). "Multi-Level Annotation in MMAX". In: *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*. Sapporo, Japan, pp. 198–207. URL: https://www.aclweb.org/anthology/W03-2117.

– (2006). "Multi-level annotation of linguistic data with MMAX2". In: *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Ed. by Sabine Braun, Kurt Kohn, and Joybrato Mukherjee. Frankfurt a.M., Germany: Peter Lang, pp. 197–214.

Nalisnick, Eric T. and Henry S. Baird (Aug. 2013). "Character-to-Character Sentiment Analysis in Shakespeare's Plays". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sofia, Bulgaria, pp. 479–483. URL: https://aclanthology.org/P13-2085.

Naumann, Karin (May 2007). *Manual for the Annotation of in-document Referential Relations*. Abt. Computerlinguistik Universität Tübingen. URL: http://www.sfs.uni-tuebingen.de/resources/tuebadz-coreference-manual-2007.pdf.

Newman, Mark (2010). *Networks: An Introduction*. 1st ed. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199206650.001.0001. URL: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199206650.001.0001/acprof-9780199206650.

Pagel, Janis (July 2022a). "Co-reference networks for dramatic texts: Network analysis of German dramas based on co-referential information". In: *Book of Abstracts of DH2022*. Tokyo, Japan (Online), pp. 326–329. URL: https://dh2022.dhii.asia/dh2022bookofabsts.pdf.

Pagel, Janis and Nils Reiter (May 2020). "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German". In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 55–64. URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.7.pdf.

– (Nov. 2021). "DramaCoref: A Hybrid Coreference Resolution System for German Theater Plays". In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021)*. Punta Cana, Dominican Republic, pp. 36–46. URL: https://aclanthology.org/2021.crac-1.4.

Pagel, Janis, Nidhi Sihag, and Nils Reiter (Nov. 2021). "Predicting Structural Elements in German Drama". In: *Proceedings of the Second Conference on Computational Humanities Research (CHR2021)*. Amsterdam, The Netherlands (Online), pp. 217–227. URL: http://ceur-ws.org/Vol-2989/short_paper34.pdf.

Passonneau, Rebecca J. (May 2004). "Computing Reliability for Coreference Annotation". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal, pp. 1503–1506. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/752.pdf.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein (July 2006). "Learning Accurate, Compact, and Interpretable Tree Annotation". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*. Sydney, Australia, pp. 433–440. DOI: 10.3115/1220175.1220230. URL: https://aclanthology.org/P06-1055.

Petrov, Slav and Dan Klein (Apr. 2007). "Improved Inference for Unlexicalized Parsing". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies (NAACL-HLT)*. Rochester, NY, USA: Association for Computational Linguistics, pp. 404–411. URL: https://www.aclweb.org/anthology/N07-1051.

Pfister, Manfred (1988). *The Theory and Analysis of Drama*. Trans. by John Halliday. European Studies in English Literature. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511553998.

Pichler, Axel and Nils Reiter (Nov. 2021). "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili". In: *Journal of Literary Theory* 15.1–2, pp. 1–29. DOI: 10.1515/jlt-2021-2008. URL: https://www.degruyter.com/document/doi/10.1515/jlt-2021-2008/html.

Piper, Andrew (Jan. 2020). "Do we know what we are doing?" In: *Journal of Cultural Analytics* 5.1. DOI: 10.22148/001c.11826.

Poot, Corbèn and Andreas van Cranenburgh (Dec. 2020). "A Benchmark of Rule-Based and Neural Coreference Resolution in Dutch Novels and News". In: *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*. Barcelona, Spain (online), pp. 79–90. URL: https://aclanthology.org/2020.crac-1.9.

Pradhan, Sameer S., Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube (June 2014). "Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD, USA, pp. 30–35. URL: https://aclweb.org/anthology/P14-2006.

Pradhan, Sameer S., Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. CoNLL 2012. Jeju, Republic of Korea: Association for Computational Linguistics, pp. 1–40.

Pradhan, Sameer S., Martha Palmer, Eduard Hovy, Lance Ramshaw, Mitch Marcus, and Ralph Weischedel (Dec. 2007). "OntoNotes: A unified relational semantic representation". In: *International Journal of Semantic Computing* 1.4, pp. 405–419.

Pradhan, Sameer S., Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (June 2011). "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Portland, OR, USA, pp. 1–27. URL: https://aclanthology.org/W11-1901.

Propp, Vladimir (1968). *Morphology of the Folktale*. Trans. by Laurence Scott. 2nd ed. Austin: University of Texas Press. ISBN: 978-0-292-78376-8.

Pustejovsky, James and Amber Stubbs (2012). *Natural language annotation for machine learning*. 1st ed. Sebastopol, CA, USA: O'Reilly Media.

Rafferty, Anna and Christopher D. Manning (June 2008). "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines". In: *Proceedings of the Workshop on Parsing German*. Columbus, OH, USA, pp. 40–46. URL: https://www.aclweb.org/anthology/W08-1006.

Raghunathan, Karthik, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning (Oct. 2010). "A Multi-Pass Sieve for Coreference Resolution". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. MIT, MA, USA, pp. 492–501. URL: https://www.aclweb.org/anthology/D10-1048.

Ramponi, Alan and Barbara Plank (Dec. 2020). "Neural Unsupervised Domain Adaptation in NLP—A Survey". In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain (Online), pp. 6838–6855. DOI: 10.18653/v1/2020.coling-main.603. URL: https://www.aclweb.org/anthology/2020.coling-main.603.

Rand, William M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336, pp. 846–850.

Recasens, Marta and Eduard Hovy (Oct. 2011). "BLANC: Implementing the Rand index for coreference evaluation". In: *Natural Language Engineering* 17.4, pp. 485–510. DOI: 10.1017/S135132491000029X.

Reinhart, Tanya (Feb. 1983). "Coreference and Bound Anaphora: A Restatement of the Anaphora Questions". In: *Linguistics and Philosophy* 6.1, pp. 47–88. DOI: 10.1007/BF00868090.

Reiter, Nils (Dec. 2018). "CorefAnnotator - A New Annotation Tool for Entity References". In: *Book of Abstracts of the European Association for Digital Humanities (EADH)*. Galway, Ireland. DOI: 10.18419/opus-10144. URL: https://elib.uni-stuttgart.de/bitstream/11682/10161/1/Abstract.pdf.

Reiter, Nils and Anette Frank (July 2010). "Identifying Generic Noun Phrases". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden, pp. 40–49. URL: https://aclanthology.org/P10-1005.

Reiter, Nils, Benjamin Krautter, Janis Pagel, and Marcus Willand (Dec. 2018). "Detecting Protagonists in German Plays around 1800 as a Classification Task". In: *Book of Abstracts of the European Association for Digital Humanities (EADH)*. Galway, Ireland. DOI: 10.18419/opus-10162. URL: https://elib.uni-stuttgart.de/bitstream/11682/10179/1/article.pdf.

Remus, Robert, Uwe Quasthoff, and Gerhard Heyer (May 2010). "SentiWS - A Publicly Available German-language Resource for Sentiment Analysis". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta, pp. 1168–1171. URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf.

Riester, Arndt and Stefan Baumann (2017). *The RefLex Scheme – Annotation Guidelines*. SinSpeC. Working papers of the SFB 732 Vol. 14. University of Stuttgart. URL: https://elib.uni-stuttgart.de/bitstream/11682/9028/1/RefLex-SinSpec14.pdf.

Rösiger, Ina (2019). "Computational modelling of coreference and bridging resolution". PhD thesis. University of Stuttgart. DOI: 10.18419/opus-10346. URL: https://elib.uni-stuttgart.de/bitstream/11682/10363/1/arbeit.pdf.

Rösiger, Ina and Jonas Kuhn (May 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)*. Portorož, Slovenia, pp. 155–160. URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/633_Paper.pdf.

Rösiger, Ina, Sarah Schulz, and Nils Reiter (Aug. 2018). "Towards coreference for literary text: Analyzing domain-specific phenomena". In: *Proceedings of the Second*

Joint SIGHUM Workshop on Computational Lingusitics for Cultural Heritage, Social Sciences, Humanities and Literature. Santa Fe, New Mexico, pp. 129–138. URL: https://aclweb.org/anthology/W18-4515.

Sahle, Patrick (Feb. 2015). "Digital Humanities? Gibt's doch gar nicht!" In: *Grenzen und Möglichkeiten der Digital Humanities. Sonderband der Zeitschrift für digitale Geisteswissenschaften* 1. Ed. by Constanze Baum and Thomas Stäcker. DOI: 10.17175/sb001_004. URL: https://zfdg.de/sb001_004.

Saif, Mohammad (June 2011). "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales". In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Portland, OR, USA, pp. 105–114. URL: https://aclanthology.org/W11-1514.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (June 2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *CoRR* abs/1910.01108. URL: http://arxiv.org/abs/1910.01108.

Schiller, Friedrich (1781/2017). *Die Räuber. Ein Schauspiel*. Ed. by Christian Grawe. 2nd ed. Ditzingen: Reclam.

Schmid, Hans-Jörg (2000). *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Vol. 34. Topics in English Linguistics. Berlin: De Gruyter Mouton. ISBN: 9783110167672. DOI: 10.1515/9783110808704.

Schmidt, David, Markus Krug, and Frank Puppe (Mar. 2022). "Adapting Coreference Algorithms to German Fairy Tales". In: *Book of Abstracts of DHd 2022*. Potsdam, Germany, pp. 34–37. DOI: 10.5281/zenodo.6328165.

Schmidt, Thomas and Manuel Burghardt (Aug. 2018). "An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing". In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH)*. Santa Fe, New Mexico, pp. 139–149. URL: https://aclanthology.org/W18-4516.

Schmidt, Thomas, Manuel Burghardt, Katrin Dennerlein, and Christian Wolff (May 2019). "Sentiment Annotation for Lessing's Plays: Towards a Language Resource for Sentiment Analysis on German Literary Texts". In: *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LDK-PS)*. Leipzig, Germany, pp. 45–50. URL: https://ceur-ws.org/Vol-2402/paper9.pdf.

Schmidt, Thomas, Katrin Dennerlein, and Christian Wolff (Nov. 2021a). "Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language". In: *Proceedings of the 5th Joint SIGHUM*

Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Ed. by Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, pp. 67–79. DOI: 10.18653/v1/2021.latechclfl-1.8. URL: https://aclanthology.org/2021.latechclfl-1.8.

Schmidt, Thomas, Katrin Dennerlein, and Christian Wolff (2021b). "Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays". In: *Fabrikation von Erkenntnis: Experimente in den Digital Humanities*. Ed. by Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, and Ulrike Wuttke. Vol. <++>. Esch-sur-Alzette: Melusina Press. DOI: 10.26298/melusina.8f8w-y749-udlf.

Schonlau, Anja (2017). *Emotionen im Dramentext. Eine methodische Grundlegung mit exemplarischer Analyse zu Neid und Intrige 1750-1800*. Ed. by Beate Kellner and Claudia Stockinger. 1st ed. Vol. 25. Deutsche Literatur. Studien und Quellen. Berlin/Boston: De Gruyter. DOI: 10.1515/9783110538120.

Schröder, Fynn, Hans Ole Hatzel, and Chris Biemann (Sept. 2021). "Neural End-to-end Coreference Resolution for German in Different Domains". In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS)*. Düsseldorf, Germany. URL: https://konvens2021.phil.hhu.de/wp-content/uploads/2021/09/2021.KONVENS-1.15.pdf.

Schweitzer, Katrin, Kerstin Eckart, Markus Gärtner, Agnieszka Faleńska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn (May 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)*. Miyazaki, Japan, pp. 2887–2895. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/320.pdf.

Skowron, Marcin, Martin Trapp, Sabine Payr, and Robert Trappl (2016). "Automatic Identification of Character Types from Film Dialogs". In: *Applied Artificial Intelligence* 30.10, pp. 942–973. DOI: 10.1080/08839514.2017.1289311. URL: https://www.tandfonline.com/doi/full/10.1080/08839514.2017.1289311.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". In: *Computational Linguistics* 27.4, pp. 521–544.

Sukthanker, Rhea, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu (July 2020). "Anaphora and coreference resolution: A review". In: *Information Fusion* 59,

pp. 139–162. DOI: 10.1016/j.inffus.2020.01.010. URL: https://www.sciencedirect.com/science/article/pii/S1566253519303677.

Telljohann, Heike, Erhard Hinrichs, and Sandra Kübler (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)*. Lisbon, Portugal, pp. 2229–2232. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/135.pdf.

Toshniwal, Shubham, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel (Nov. 2020). "Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 8519–8526. DOI: 10.18653/v1/2020.emnlp-main.685. URL: https://aclanthology.org/2020.emnlp-main.685.

Trilcke, Peer, Frank Fischer, and Dario Kampkaspar (2015). "Digital network analysis of dramatic texts". In: *DH2015 Conference Abstracts*. Sydney, Australia.

Trilcke, Peer, Christopher Kittel, Nils Reiter, Daria Maximova, and Frank Fischer (July 2020). "Opening the Stage: A Quantitative Look at Stage Directions in German Drama". In: *Book of Abstracts of the DH2020 conference*. Ottawa, Canada. URL: https://dh2020.adho.org/wp-content/uploads/2020/07/337_OpeningtheStageAQuantitativeLookatStageDirectionsinGermanDrama.html.

Tuggener, Don (2016). "Incremental Coreference Resolution for German". PhD thesis. University of Zürich. DOI: 10.5167/uzh-124915. URL: https://www.zora.uzh.ch/id/eprint/124915/1/tuggener_diss.pdf.

Vala, Hardik, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths (May 2016). "Annotating Characters in Literary Corpora: A Scheme, the CHARLES Tool, and an Annotated Novel". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, pp. 184–189. URL: https://www.aclweb.org/anthology/L16-1028.

Vala, Hardik, David Jurgens, Andrew Piper, and Derek Ruths (Sept. 2015). "Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal, pp. 769–774. DOI: 10.18653/v1/D15-1088. URL: https://www.aclweb.org/anthology/D15-1088.

Valls-Vargas, Josep, Jichen Zhu, and Santiago Ontañón (Oct. 2014). "Toward Automatic Role Identification in Unannotated Folk Tales". In: *Proceedings of the Tenth Annual*

AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE). Raleigh, NC, USA, pp. 188–194. DOI: 10.1609/aiide.v10i1.12732.

Van Cranenburgh, Andreas (Dec. 2019a). "A Dutch coreference resolution system with an evaluation on literary fiction". In: *Computational Linguistics in the Netherlands Journal* 9, pp. 27–54. URL: https://www.clinjournal.org/clinj/article/view/91.

Van Cranenburgh, Andreas, Esther Ploeger, Frank van den Berg, and Remi Thüss (Nov. 2021). "A Hybrid Rule-Based and Neural Coreference Resolution System with an Evaluation on Dutch Literature". In: *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*. Punta Cana, Dominican Republic, pp. 47–56. DOI: 10.18653/v1/2021.crac-1.5. URL: https://aclanthology.org/2021.crac-1.5.

Van Cranenburgh, Andreas and Gertjan van Noord (Dec. 2022). "OpenBoek: A Corpus of Literary Coreference and Entities with an Exploration of Historical Spelling Normalization". In: *Computational Linguistics in the Netherlands Journal* 12, pp. 235–251. URL: https://www.clinjournal.org/clinj/article/view/157.

Van Deemter, Kees and Kibble Rodger (2000). "On Coreferring: Coreference in MUC and Related Annotation Schemes". In: *Computational Linguistics* 26.4, pp. 629–637. URL: https://aclanthology.org/J00-4005.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). *Attention Is All You Need*. Version 5. DOI: 10.48550/arXiv.1706.03762. arXiv: arXiv:1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Versley, Yannick (2006). "Disagreement Dissected: Vagueness as a Source of Ambiguity in Nominal (Co-)Reference". In: *ESSLLI 2006 Workshop on Ambiguity in Anaphora*. Málaga, Spain, pp. 83–89.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (Nov. 1995). "A Model-Theoretic Coreference Scoring Scheme". In: *Proceedings of the 6th conference on Message understanding (MUC6)*. Columbia, MD, USA, pp. 45–52. URL: https://www.aclweb.org/anthology/M95-1005.

Webber, Bonnie Lynn (June 1988). "Discourse Deixis: Reference to Discourse Segments". In: *26th Annual Meeting of the Association for Computational Linguistics (ACL)*. Buffalo, NY, USA, pp. 113–122. DOI: 10.3115/982023.982037. URL: https://aclanthology.org/P88-1014.

Willand, Marcus, Benjamin Krautter, Janis Pagel, and Nils Reiter (Mar. 2020). "Passive Präsenz tragischer Hauptfiguren im Drama". In: *Book of Abstracts of DHd*. Paderborn, Germany, pp. 177–181. DOI: 10.5281/zenodo.4621812.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, and Jason Weston (July 2015). "Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. Beijing, China, pp. 1416–1426. DOI: 10.3115/v1/P15-1137. URL: https://aclanthology.org/P15-1137.

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (June 2016). "Learning Global Features for Coreference Resolution". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, CA, USA, pp. 994–1004. DOI: 10.18653/v1/N16-1114. URL: https://aclanthology.org/N16-1114.

Worthen, William B. (Oct. 1998). "Drama, Performativity, and Performance". In: *PMLA* 113.5, pp. 1093–1107. DOI: 10.2307/463244.

Xia, Patrick, João Sedoc, and Benjamin Van Durme (Nov. 2020). "Incremental Neural Coreference Resolution in Constant Memory". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 8617–8624. DOI: 10.18653/v1/2020.emnlp-main.695. URL: https://aclanthology.org/2020.emnlp-main.695.

Yavuz, Mehmet Can (Mar. 2020). "Analyses of Character Emotions in Dramatic Works by Using EmoLex Unigrams". In: *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it)*. Bologna, Italy, pp. 471–476. DOI: 10.4000/books.aaccademia.9004.

Yu, Juntao, Bernd Bohnet, and Massimo Poesio (May 2020). "Neural Mention Detection". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 1–10. URL: https://www.aclweb.org/anthology/2020.lrec-1.1.

Yu, Juntao, Alexandra Uma, and Massimo Poesio (May 2020). "A Cluster Ranking Model for Full Anaphora Resolution". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. Marseille, France, pp. 11–20. URL: https://aclanthology.org/2020.lrec-1.2.

Zaporojets, Klim, Johannes Deleu, Yiwei Jiang, Thomas Demeester, and Chris Develder (May 2022). "Towards Consistent Document-level Entity Linking: Joint Models for

Entity Linking and Coreference Resolution". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Dublin, Ireland, pp. 778–784. DOI: 10.18653/v1/2022.acl-short.88. URL: https://aclanthology.org/2022.acl-short.88.

Primary Sources of Plays

DraCor (2020). *Shakespeare, William. Romeo and Juliet.* Potsdam, Germany. URL: https://dracor.org/id/shake000028.

TextGrid Repository (2011). *Anzengruber, Ludwig. Der Meineidbauer*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0001-DD80-2.

– (2012a). *Cronegk, Johann Friedrich von. Der Mißtrauische*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0002-590A-A.

– (2012b). *Hofmannsthal, Hugo von. Der Rosenkavalier*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0003-7902-4.

– (2012c). *Kleist, Heinrich von. Die Familie Schroffenstein*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0003-B161-C.

– (2012d). *Lessing, Gotthold Ephraim. Emilia Galotti*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0003-E98E-1.

– (2012e). *Lessing, Gotthold Ephraim. Miß Sara Sampson*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0003-E5B7-5.

– (2012f). *Lessing, Gotthold Ephraim. Nathan der Weise*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0003-E93B-E.

– (2012g). *Schiller, Friedrich. Die Räuber*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0004-CE8E-8.

– (2012h). *Weißenthurn, Johanna von. Das Manuscript*. Göttingen. URL: https://hdl.handle.net/11858/00-1734-0000-0005-9A9C-9.

Data Resources / Corpora

Andresen, Melanie and Michael Vauth (May 2018b). *Character mentions in the German novel "Corpus Delicti" by Juli Zeh and annotations*. Version 1.0. DOI: 10.5281/zenodo.1239702. URL: https://doi.org/10.5281/zenodo.1239702.

Eckart, Kerstin, Arndt Riester, and Katrin Schweitzer (2012). *DIRNDL: Discourse Information Radio News Database for Linguistic Analysis*. Version 1.1. URL: http://hdl.handle.net/11022/1007-0000-0000-8E31-9.

Hinrichs, Erhard, Heike Zinsmeister, Marie Hinrichs, Yannick Versley, and Heike Telljohann (Nov. 2009). *TüBa-D/Z: Tübingen Treebank of Written German / Newspaper Corpus*. Version 11. URL: https://hdl.handle.net/11858/00-1778-0000-0005-896C-F.

Krug, Markus and Albin Zehe (2018). *Deutscher Romankorpus (DROC)*. URL: https://gitlab2.informatik.uni-wuerzburg.de/kallimachos/DROC-Release.

Mitchell, Alexis, Stephanie Strassel, Mark Przybocki, JK Davis, George R. Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim (Sept. 2003). *ACE-2*. Version 1.0. DOI: 10.35111/kcqk-v224.

Pagel, Janis (May 2022b). *quadrama/gerdracor-coref*. Version v1.4.0. DOI: 10.5281/zenodo.6556064. URL: https://doi.org/10.5281/zenodo.6556064.

Schweitzer, Katrin, Kerstin Eckart, and Markus Gärtner (May 2018). *GRAIN: German Radio Interviews*. Version 1.0. URL: http://hdl.handle.net/11022/1007-0000-0007-C632-1.

Van Cranenburgh, Andreas (2019b). *andreasvc/dutchcoref*. URL: https://github.com/andreasvc/dutchcoref.

– (2022). *andreasvc/openboek*. URL: https://github.com/andreasvc/openboek.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston (Oct. 2013). *OntoNotes*. Version 5.0. DOI: 10.35111/xmhb-2b84.

Winston, Patrick and Mark Alan Finlayson (2015). *Supplementary materials for "PropplLearner: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory"*. URL: http://hdl.handle.net/1721.1/100054.

Software

Pagel, Janis (May 2020). *pagelj/DramaCoref*. URL: https://github.com/pagelj/DramaCoref.

Pradhan, Sameer, Xiaoqiang Luo, and Marta Recasens (June 2016). *conll/reference-coreference-scorers*. Version v8.01. URL: https://github.com/conll/reference-coreference-scorers.

Reiter, Nils, Börge Kiss, and Andreas van Cranenburgh (June 2022). *nilsreiter/CorefAnnotator*. Version v2.1.1. DOI: 10.5281/zenodo.6694817. URL: https://doi.org/10.5281/zenodo.6694817.