Coreference Resolution

Dutch and German coreference resolution
a cross-linguistic corpuslinguistic comparison

Author: Lisa Becker, April 15, 2020

1 Abstract

This project is an exploratory cross-linguistic corpuslinguistic study within the field of coreference resolution. It compares a Dutch and a German coreference system with each other and across two parallel test corpora by analysing the types of annotation incongruences. I exploited the current state-of-the-art coreference system for Dutch which was developed by Cranenburgh (2019) called dutchcoref and is a rule-based sieve approach based on the Stanford CoreNLP coreference system. The German coreference system is an entity-mention approach called CorZu developed by Tuggener (2016). The two test corpora were drawn from Tiedemann (2012) and contain a corpus with news commentaries and a fictive literature corpus which were chosen in regard to the train corpora of the two coreference systems. As expected, each coreference system performs better on the corpus which was thematically closer related to the corpus it was trained on compared to the other corpus. CorZu produces more annotations than dutchcoref which are not necessarily correct.

2 Introduction

With computational analysis of language on the rise, the task of automatically and correctly identifying the links between an entity and its mentions in a discourse is getting more important. This task is called coreference resolution which is carried out intuitively by humans in most cases but appears to be a highly complex task for an algorithm. A coreference is each mention in a discourse and can be any sort of expression that refers to an entity (the referent). A mention is typically a pronoun or a noun phrase but can also be a clause or even a whole sentence. If there is a relation between a mention and a referent, they corefer (Jurafsky and Martin 2020). A set of coreferring expressions which link to the same entity is called a coreference chain or cluster. Depending on whether the entity was introduced before it gets referred to, the relation is either called an anaphora (entity-first) or a cataphora (mention-first).

> **Anaphora:** I like my dog [Xula]$^{1}{}_{\text{antecedent}}$ because [she]$_{\text{anaphor}}$ is very cute.
> **Cataphora:** [She]$_{\text{cataphor}}$ is very cute, that's why I like my dog [Xula]$_{\text{postcedent}}$.[2]

If an entity appears only once in a discourse and is not referred to again, it is called a singleton.
Entity linking describes the reference between a mention in a discourse and a real-life entity, which might require real-life knowledge or discourse deixis.
But not all cases of pronouns are mentions: The pronoun "it" in "it rains" carries a syntactical role but serves no other purpose. It is an expletive (a type of pleonasm), which appears especially in non-drop-languages like English, Dutch or German, where there has to be an overt syntactical subject.

Each language contains different linguistic features to resolve coreference. Many languages exploit morphological markers like number, person and gender agreement between a mention and its referent, which will be discussed in Section **??**. Next to overt grammatical features, binding theory typically restricts the possibility of a coreferential relation within a sentence.

1. My dog Xula is the main character of the examples used in this project paper in memory of her, who passed away on February 20th, 2020.
2. Cataphors tend to pose difficulties for coreference systems. This problem will not be discussed in this paper.

Lastly, the more recently an entity was introduced or mentioned, the more likely it is that it is refered to again. This is closely related to saliency or "prominency" of an entity in a discourse. Grammatical role also influences the saliency: the subject position is generally more salient than the object position. World knowledge and semantics are also important for coreference resolution and disambiguates syntactically and grammatically ambiguous coreference:

> **Grammatical Ambiguity:** I went for a walk to the park with my [mother]$_1$ and my dog [Xula]$_2$. Suddenly, [she]$_x$ starts barking.

World knowledge and the semantics of the verb barking indicate that it was rather my dog Xula (2) than my mother (1) who suddenly started to bark. In the current project, I focus on persons and objects (referred to by names, pronouns and descriptions or appositives[3]). Events and actions are not be included in this paper.

The next Section 2.1 discusses the comparability of Dutch and German. Section 2.2 explains how coreference systems work in general with a focus on the Stanford CoreNLP approach and Section 2.2.1 and 2.2.2 elaborate on the background of Dutch and German coreference systems and present the two coreference systems used for this project. The data is presented in Section 3. Section 4 talks about the implementations, the input and output, Section 4.1 and 4.2 about the analysis and the results. The paper closes with the Discussion and Conclusion with Section 5 and 6.

2.1 Dutch and German Linguistics

I chose Dutch and German for this project since their research is underrepresented and they are closely related to each other. Both languages (as well as English) belong to the family of West Germanic languages.

Syntactically, Dutch and German are closer related to each other than to English because of their simultaneous switching between verb object (VO) and object verb (OV) word order while English sticks to a subject verb object (SVO) word order. In a main sentence, Dutch and German place the finite verb in the second position after the first constituent (verb second word order):

> **NL:** VO: De hond **blaft** luid. OV: Opeens **blaft** de hond luid.
> **DE:** VO: Der Hund **bellt** laut. OV: Plötzlich **bellt** der Hund laut.
> **EN:** VO: The dog **barks** loudly. VO: Suddenly, the dog **barks** loudly.

Both Dutch and German switch to a verb final (VF) word order after a complementizer while English keeps the SVO word order:

> **NL:** VF: Ik hoor dat de hond luid **blafft**.
> **DE:** VF: Ich höre, dass der Hund laut **bellt**.
> **EN:** VO: I hear that the dog **barks** loudly.

The most significant syntactic difference between Dutch and German is the order of finite verbs and infinitives at the end of a sentence but which is not to be expected to make a difference in the task of automatic coreference resolution:

> **NL:** Ik hoor dat de hond luid **kan blaffen**.
> **DE:** Ich höre, dass der Hund laut **bellen kann**.
> **EN:** I hear that the dog **can bark** loudly.

3. Appositives are noun phrases to describe another noun phrase and are thus part of the coreference chain of the mention which they accompany.

Next to syntax, grammatical gender is another important feature of coreference. Both Dutch and German overtly represent the grammatical gender masculine, feminine and neuter in definite and indefinite articles (see Table 1). Plural articles are underspecified, which poses a problem in automatic coreference resolution.

Language	article	masculine	feminine	neuter
Dutch	definite	der	die	das
	indefinite	ein	eine	ein
German	definite	de	de	het
	indefinite	een	een	een
English	definite	the	the	the
	indefinite	a(/n)	a(/n)	a(/n)

Table 1: Singular articles in Dutch and German. English for comparison.

Language	pronoun type	masculine		feminine		neuter	
		singular	plural	singular	plural	singular	plural
Dutch	subject	hij	ze	ze/zij	ze	het	ze
	object	hem	ze	haar	ze	het	ze
	possessive	zijn	hun	haar	hun	zijn	hun
German	subject	er	sie	sie	sie	es	sie
	object	ihn	sie	sie	sie	es	sie
	possessive	sein	ihr	ihr	ihr	sein	ihr
English	subject	he	they	she	they	it	they
	object	him	them	her	them	it	them
	possessive	his	theirs	hers	theirs	its	theirs

Table 2: Third person singular personal pronouns for in Dutch and German. English for comparison.

Masculine, feminine and neuter pronouns are used in Dutch and German to refer to animated nouns and persons with natural gender. In contrast to German, the distinction between masculine and feminine inanimate nouns almost completely disappeared in Dutch, leaving the pronoun *hij* as the default pronoun for masculine and feminine nouns (with definite article *de*) and the pronoun *het* for neuter nouns (with definite article *het*) as seen in Table 2 (Berkum 1996). This can vary across Dutch dialects. The distinction between genders occurs only with third-person singular pronouns, which leaves other pronouns more ambiguous in coreference and requires other features like discourse deixis.

As seen in Table 2, pronouns in both languages are underspecified regarding their number and gender attributes especially in feminine singular and plural subject and object pronouns but also in possessive pronouns across all genders. This ambiguity can lead to problems in automated coreference resolution.

These parallels of grammatical and syntactical features makes Dutch and German suitable languages for comparison.

2.2 What is Coreference Resolution

A coreference system is typically part of an natural language processing pipeline which includes named entity recognition (NER) and syntactic parsing. After the tokens are preprocessed, the system has to decide which noun phrases, pronouns and clauses will be considered parts of coreference chains. These so-called markables get extracted and merged into coreference chains with a

discourse processing strategy. This usually contains morphosyntactic agreement as well as binding and distance constraints.

The current state of the art is dominated by neural systems (K. Lee et al. 2017; Lee, He, and Zettlemoyer 2018), but supervised models such as rule-based an statistical approaches are still in use and being developed. One disadvantage of neural models is their requirement of high computing power and intensive memory use which makes working with long samples only possible in small chunks. This may be an issue with literary texts (Cranenburgh 2019) as are being used in this project. Generally, coreference systems are built on a gold standard corpus which means that the corpus' coreferences were manually annotated beforehand. The computer then aims to produce the same coreference chains as a human. The difference between the gold annotation and the system annotation constitutes the basis for the performance of automatic coreference system. Since there were no gold annotated corpora used for this project, the quality of the system output was not measured by the completeness of its annotations but by comparison to another system and across languages.

The Dutch coreference system, which is exploited in this project, is a rule-based sieve-based approach based on the Stanford CoreNLP system.

The multi-pass sieve-based Stanford CoreNLP system is an entity-mention[4] annotation pipeline performing the following steps in this order: sentence splitting, tokenization, constituency and dependency parsing as well as extraction of morphological data. It takes all noun and pronominal phrases of the sample and applies a battery of coreference sieves in order to decide how to cluster them together. The first sieve applies its rules, forms the first (partial) coreference chains and passes them into the next sieve which merges chains, adds new ones or discards ones which were already built. The number and function of the sieves vary from system to system which is explained in Section 2.2.2.

2.2.1 Background on Dutch Coreference Resolution

Hoste (2005) started the work on Dutch coreference resolution by training and evaluating a mention-pair system with the KNACK 2002 corpus of magazines[5] (Hoste and Pauw 2006). Hendrickx et al. (2013) evolved this system and annotated a bigger corpus called Corea project which involved the development of a Dutch parsing system Alpino[6] (Bouma, Noord, and Malouf 2001) which was exploited for this project.
De Clercq, Hoste, and Hendrickx (2011) published cross-domain coreference results with the so far biggest Dutch annotation project SoNaR-1 with about 1 million words (Schuurman, Hoste, and Monachesi 2010).
The Newsreader project is a multilingual approach of annotating entities, events and temporal expressions in news articles as well as the relations between them (Schoen et al. 2014). These annotations were used in the shared task of the CLIN26 Dutch entity coreference track[7], a yearly conference on Dutch computational linguistics.

The most recent coreference system is a Dutch implementation[8] of the Stanford CoreNLP sieve algorithm called dutchcoref (Cranenburgh 2019). It is targeted to analyse the characters in literary fiction. They developed a rule-based coreference system based on the H. Lee et al. (2011) and H.

4. Some researchers call it entity-mention, others call it mention-entity. To avoid confusion, this paper sticks to the notation "entity-mention".

5. The KNACK 2002 corpus seems not to be available online as of April 15, 2020.

6. Alpino Parser for Dutch:
http://www.let.rug.nl/vannoord/alp/Alpino/AlpinoUserGuide.html (last accessed April 15, 2020).

7. CLIN26 shared task: http://wordpress.let.vupr.nl/clin26/shared-task/ (last accessed April 15, 2020).

8. Stanford CoreNLP sieve implementation for Dutch (dutchcoref) :
https://github.com/andreasvc/dutchcoref (last accessed April 15, 2020).

Lee et al. (2013) system which won the CoNLL 2011 shared task[9] (Pradhan et al. 2011; Pradhan et al. 2012). Texts are preprocessed and parsed with the Alpino parser, specifically for tokenization and sentence splitting. The system detects and filters mentions and their boundaries and clusters them into entities. Neoplastic pronouns, time-related expressions and mentions that do not refer to entities are excluded. After quotes are getting attributed, the following 5 of the 7 original Stanford CoreNLP sieves are being applied:

1. String Match: Mentions which are not pronouns but are identical in their string representation are linked as referring to the same referent.
2. Precise constructs: This sieve links mentions that can be inferred from a parse tree and uses relative, reflexive and reciprocal pronouns as well as appositives, predicate nominals and acronyms to establish coreference links. If the parse tree contains errors, the quality of this sieve is impaired.
3. Head match: Nominal mentions with matching heads and modifiers are linked as coreferent. For that, the second mention must be a subset of the first mention. For instance, "the University of Potsdam" gets linked to "the university", but not to "the University of Berlin".
4. Proper head noun match: Different variations of names get linked, for instance: "Lisa Becker", "Mrs. Becker", "Lisa". This sieve is applied to the whole sample while other sieves are only linked to mentions if preceded by an antecedent.
5. Pronoun resolution: Pronouns get linked to referents based on matching features, binding constraints, recency ranking and syntactic prominence/salience. If a pronoun occurs in quoted speech, it is treated separately: first and second person pronouns refer to the detected speaker and addressee. This version of the system does not support the resolution of cataphora (as of April 15, 2020).

The system was tested on Dutch novels from the Riddle of Literary Quality corpus[10] which were manually annotated before to provide a gold standard for evaluation. Dutchcoref was exploited for this project.

2.2.2 Background on German Coreference Resolution

There has been more research about German coreference resolution compared to Dutch. Only the most relevant ones are being mentioned in this paper: Like Cranenburgh (2019), Krug et al. (2015) also based their coreference system for German on H. Lee et al. (2011) and H. Lee et al. (2013) and trained it on historic literary texts (in contrast to dutchcoref which is trained on modern literary texts). Krug et al. (2015) used 11 sieves based on the Stanford CoreNLP algorihm. This was the first coreference system for German that drew more attention but since it is not available online, it could not be used for this project.

Similar to Cranenburgh (2019), Srivastava et al. (2018) provided the first freely available coreference resolution system for German which they called CoRefGer[11] and is also based on the Stanford CoreNLP sieve algorithm. In their work, they adapted a rule-based, a statistical and a knowledge-based coreference system from English to German and performed coreference resolution on both languages.

The current version of the program implemented six of the seven sieves from Stanford CoreNLP system (as of April 15, 2020):

1. Exact Match: This is equivalent to the "String match" sieve from dutchcoref except that they

9. The CoNLL 2011 shared task produced the CoNLL-style format which verticalizes the text and adds tokens annotations horizontally as columns. An example can be seen in Table 4 and 5.
10. The Riddle of Literary Quality corpus: http://literaryquality.huygens.knaw.nl/ (last accessed April 15, 2020).
11. Stanford CoreNLP sieve implementation for German (CoRefGer):
https://github.com/dkt-projekt/e-NLP/tree/master/src/main/java/de/dkt/eservices/ecorenlp/modules (last accessed April 15, 2020).

use a sliding window of five sentences that compares the strings of noun phrases to another. If they match, they are linked to the same coreference chain. This sieve includes variations in endings including name extensions and morphology, for instance "der Hund" (DE: the dog) and "des Hundes" (DE: the dog's) get linked by this sieve.
2. Precise constructs: This sieve matches the Dutch sieve of Precise Constructs which uses the information from a parse tree, relative, reflexive and reciprocal pronouns and appositive constructs.
3. / **4.** / **5. Noun Phrase Head Matching:** This sieve matches all mentions with the same head, such that it also links "the University of Potsdam" and "the University of Berlin" as coreferent although the modifiers make it clear that they refer to different entities. These sieves employs stemming.
6. Integration of Named Entity Recognition: This sieve is built on the in-house Named Entity Recognition engine DBpedia-Spotlight[12] (Daiber et al. 2013) and discards referential links between instances like "the University of Potsdam" and "the University of Berlin".

Because of the similarity of sieves between the dutchcoref and CoRefGer, they are interesting candidates to compare in a crosslinguistic study. Additionally to the mentioned sieves, Srivastava et al. (2018) included stemming to avoid false negative matches due to different case markers in German articles and noun phrases. They applied their system to parallel corpora of English and German of digital curation scenarios such as digital archives, newspaper reports and museum exhibits. In conclusion, Srivastava et al. (2018) determined that their deterministic rule-based system achieves the best performance for out-of-domain use cases.

What Srivastava et al. (2018) failed to include is a morphological processor for gender and number information for pronoun matching. Their method merely matches pronouns that are morphologically identical.

In comparison to Krug et al. (2015), they left out the following sieves or comprised them in other sieves by using the DBpedia-Spotlight tool:

Nameflexion: Detection of variations from names and co-referring them (e.g. derivatives like "Lisa", "Lissie", "Lischen").
Attributes: Modifiers used as names are matched and coreferred (e.g. *"Die alte Lisa"* (DE: the elderly Lisa), *"die Alte"* (DE: the elderly).
Semantic: Semantic synonyms get matched and coreferred (e.g. *"Gatte"* (= spouse) and *"Gemahl"* (DE: consort).
Pronoun resolution: Pronouns are getting linked with the use of the following features: saliency, morphology, context (whether they appear in direct speech or not), binding constraints.
Detection of the addressed person in direct speech: Finding the addressed named entity in a speech annotation. Firstly, the speaker gets detected, then the possible addressed person in the dialogue can be detected (e.g. by expressions like "Xula, you are great").
Pronouns in direct speech: Change of pronouns due to deixis (e.g. "I" changes to "you" when the speaker changes). In contrast, this sieve is included in the 5th sieve from dutchcoref.

Srivastava et al. (2018) also implemented a statistical approach based on the Stanford CoreNLP statistical system Mention Ranking Model (Clark and Manning 2015) which they called CoRefGer-stat and another projection based approach: One projection method computes coreference on English texts and projects the annotations on a parallel German text via word alignments (transferring model). The other projection method translates a German text to English, computes coreference on the translated text and projects the annotations back to the German text via word alignment.

One of their biggest problem was that the rule-based system does not exploit number and gender information as it is lacking a morphological analyser which they left for future work. Nonetheless

12. DBpedia-Spotlight: https://github.com/dbpedia-spotlight/dbpedia-spotlight (last accessed April 15, 2020).

Srivastava et al. (2018) observed that their deterministic rule-based system performs better for out-of-domain use cases than their other methods, like the statistical and the projection-based approach, although not being deep learning based. Thus, the rule-based system CoRefGer-rule was the choice for the German system for this project.

However, as much as this coreference system for German would have been comparable to the Dutch system that I chose for this project, it was not possible to install and use it as it did not produce coreference chains across sentences. Consulting and discussing with the author of the system did not resolve this issue. Therefore I opted for another state-of-the-art coreference system for German, Corzu[13] (Tuggener 2016):

CorZu is trained and tested on the TüBa-D/Z treebank[14] which is a collection of newspaper articles from the German newspaper "die Tageszeitung" (Telljohann, Hinrichs, and Kübler 2004). The TüBa-D/Z version 9.1 contains 3,644 articles (with 95,595 sentences or 1,787,801 tokens) and has a gold standard. CorZu uses ParZu[15] for parsing the input which was also exploited for this project (Sennrich, Volk, and Schneider 2013; Noord 2006). ParZu preprocesses the input, tags parts-of-speech (POS), parses syntax, recognizes named entities, detects animacy, labels semantic classes and converts all of this data into the CoNLL format.
CorZu is an entity-mention model that incrementally disambiguates (see Figure 1). It partitions the mentions into clusters with each cluster corresponding to an entity such that entities are modeled and scored directly. In the main loop lines (1-17), it passes the markables from left to right. The coreference partition (lines 2-4) and the buffer list (lines 5-7) provide compatible antecedent candidates for each markable. The buffer list mainly contains nominals that can serve as antecedents but only contains markables that have not been linked to an antecedent. Lines 13-15 pair a markable from the buffer list with a new markable and appends it to the coreference partition (line 14) which is also queried for compatible antecedents candidates. In contrast to other entity-mention algorithms, CorZu restricts the accessibility of mentions in an entity to the last mention while the other mentions are hidden. This last mention gets all observed features of the coreference chain projected to it. So if a new markable is to be linked to the available coreference chains, it gets compared only to the most recent mention of a chain:

> **DE:** Im März 2020 hat die Universität Potsdam$_1$ entschieden, ihre$_1$ Semesterferien wegen der Pandemie zu verlängern.
> **EN:** In March 2020, the University of Potsdam$_1$ decided to extend their$_1$ semester break due to the pandemic.

"Ihre" (DE: her/their) is underspecified in this example. CorZu resolves the possessive pronoun to the antecedent "die Universität Potsdam" and projects the features of the antecedent to the pronoun, such that the markable "ihre" does not only gets disambiguated to feminine singular, but also holds the entity class "organisation" and is classified as an inanimate entity. The pronoun "ihre" is now the most recent mention of the coreference chain and "die Universität Potsdam" is removed from the buffer list. This prevents the pronoun "ihre" to be mistakenly linked to later occuring, incompatible pronoun markables, such as a plural instance of "sie" (DE: "they"). In a mention-pair model, for instance, those two pronouns would have been considered compatible. However, CorZu can not guarantee to prevent pronouns to be correctly disambiguated in the first place, which would subsequently lead to more wrongly annotated or missing mentions in coreference chains.

13. Coreference Resolver for German from Zurich: https://github.com/dtuggener/CorZu (last accessed April 15, 2020).
14. Coreference Resolver for German from Zurich: https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/resources/corpora/tueba-dz/ (last accessed April 15, 2020).
15. Zurich Dependency Parser for German: https://github.com/rsennrich/ParZu (last accessed April 15, 2020).

Algorithm 1 Incremental entity-mention model

Input: Markables
Output: Coreference partition

for $m_i \in Markables$ **do**
  for $e_k \in CorefPartition$ **do**
    if $compatible(e_{k_n}, m_i)$ **then** ▷ e_{k_n} is the most recent mention of entity e_k
      $Candidates \oplus e_{k_n}$
  for $m_j \in BufferList$ **do**
    if $compatible(m_j, m_i)$ **then**
      $Candidates \oplus m_j$
  $ante \leftarrow get\_best(Candidates)$
  if $ante \neq \varnothing$ **then** ▷ An antecedent has been identified
    $ante, m_i \leftarrow disambiguate(ante, m_i)$ ▷ Propagate animacy, NE class, morphology
    if $\exists e_k \in CorefPartition : ante \in e_k$ **then** ▷ Antecedent is part of a coref. chain
      $e_k \oplus m_i$
    else
      $CorefPartition \oplus \{ante \oplus m_i\}$ ▷ Open new coreference chain
      $BufferList \ominus ante$ ▷ Remove antecedent from buffer list
  else
    $BufferList \oplus m_i$ ▷ No antecedent, append m_i to buffer list
return $CorefPartition$

Figure 1: CorZu Entity-Mention Algorithm as described by Tuggener (2016). Markables are tokens that are marked as possible parts of a coreference chain.

3 Data

A news media corpus and a literature corpus were drawn from OPUS: the open parallel corpus[16] (Tiedemann 2012). I chose corpora from two different genres to provide even conditions for both coreferent systems in order to compare them to another:

CorZu was trained and tested on newspaper articles, thus I opted for the News Commentary Parallel Corpus v11[17]. It includes 21,435 aligned sentences (1.04 million tokens), from which 656 sentences (16,579 tokens) were used for this project. 18 News Commentary samples were randomly chosen from the corpus with an average of 921.06 tokens per sample (see Table 3).

Since dutchcoref was trained and tested on literary fiction, the parallel corpus of free literary works annotated by Andras Farkas[18] was used. It consists of 15,569 aligned sentences (0.53 million words), from which 908 sentences (16,701 tokens) from the novel "The Adventures of Tom Sawyer" from Mark Twain were used for this project (see Table 3). This corresponds to 12 chapters which were chosen such that their individual length matches the length of a News Commentary sample with an average of 1,391 token per sample.

16. OPUS: the open parallel corpus: http://opus.nlpl.eu/ (last accessed April 15, 2020).
17. News Commentary Parallel Corpus v11 (2016):
http://www.casmacat.eu/corpus/news-commentary.html (last accessed April 15, 2020).
18. Books by Andras Farkas: http://farkastranslations.com/bilingual_books.php (last accessed April 15, 2020).

	News				Literature			
	# NL	# DE	Avg NL	Avg DE	# NL	# DE	Avg NL	Avg DE
Tokens	16,579	15,140	921.06	841.11	16,701	14,957	1,391.75	1,246.42
Sentences	656	670	36.44	37.22	908	796	75.67	66.33
Sentence Length	467.86	413	25.99	22.94	235.34	234.31	19.61	19.53
Chapters	18				12			

Table 3: The subset of the News Commentary and Literature corpus used for this project.

4 Current Project

The German sieve-based coreference system was installable and produced coreference chains, but not across sentences. Therefore it was discarded after a week of trying to get it to run sufficiently by its developer and me. That is when the decision was made to change to CorZu which is less comparable but functional.

Both coreference systems were installed as described in their corresponding github repositories. The input samples were formatted according to the requirements of the coreference systems. Since both systems take dependency parsed samples in the CoNLL format, all Dutch samples were parsed by Alpino and all German samples by ParZu before passing them to the coreference systems. No further preprocessing was performed. The two coreference systems were applied on their matching language version of each corpus. The output files were CoNLL formated files with the coreference chains as the last column (see Tables 4 and 5, with the example "they can't choose their parents, let alone the circumstances in which they are born." from the News Commentary corpus):

ID	Form	Lemma	Features	Head	POS	Coref
6	Zij	zij	VNW(pers,pron,nomin,vol,3p,mv)	6	nsubj	-3
7	kunnen	kunnen	WW(pv,tgw,mv)	6	aux	-
8	hun	hun	VNW(bez,det,stan,vol,3,mv,prenom,zonder,agr)	4	nmod:poss	(6|(3)
9	ouders	ouder	N(soort,mv,basis)	6	obj	6)
10	niet	niet	BW()	6	advmod	-
11	uitkiezen,	uit_kiezen	WW(inf,vrij,zonder)	0	root	-
12	laat	laat	WW(pv,tgw,ev)	10	mark	-
13	staan	staan	WW(inf,vrij,zonder)	7	fixed	-
14	de	de	LID(bep,stan,rest)	10	det	(7
	omstandigheden	omstandigheid	N(soort,mv,basis)	6	advcl	7)
1	waarin	waarin	BW()	13	amod	-
2	ze	ze	VNW(pers,pron,stan,red,3,mv)	13	nsubj:pass	-3
3	geboren	geboren	WW(vd,vrij,zonder)	10	acl:relcl	-
4	worden.	worden	WW(pv,tgw,mv)	13	aux:pass	-

Table 4: Example from News Commentary corpus for CoNLL formatted output file from dutchcoref.

For this project, only the columns "Form" and "Coref" were exploited.

ID	Form	Lemma	CPOSTAG	POSTAG	Features	Head	DEPREL	Coref
1	Sie	sie	PRO	PPER	3|Pl|_|Nom	2	subj	0
2	suchen	suchen	V	VVFIN	3|Pl|Pres|_	0	root	-
3	sich	sie	PRO	PRF	3|_|Dat	2	objd	-
4	ihre	ihre	ART	PPOSAT	_|Acc|Pl	5	det	(0)|(17
5	Eltern	Elter	N	NN	_|Acc|Pl	2	obja	17)
6	nicht	nicht	PTKNEG	PTKNEG	_	2	adv	-
7	aus	aus	PTKVZ	PTKVZ	_	2	avz	-
8	,	,	$,	$,	_	0	root	-
9	geschweige	geschweige	V	VVFIN	_|Pl|_|_	2	kon	-
10	denn	denn	ADV	ADV	_	9	adv	-
11	die	die	ART	ART	Def|Masc|Acc|Pl	13	det	(1
12	allgemeineren	allgemein	ADJA	ADJA	Comp|Masc|Acc|Pl|Wk|	13	attr	-
13	Umstände	Umstand	N	NN	Masc|Acc|Pl	9	obja	1)
14	,	,	$,	$,	_	0	root	-
15	in	in	PREP	APPR	Acc	18	pp	-
16	die	die	PRO	PRELS	Masc|Acc|Pl	15	pn	-1
17	sie	sie	PRO	PPER	3|Pl|_|Nom	19	subj	0
18	hineingeboren	hineingebären	V	VVPP	_	19	aux	-
19	werden	werden	V	VAFIN	3|Pl|Pres|_	13	rel	-

Table 5: Example from News Commentary corpus for CoNLL formated output file from CorZu.

4.1 Analysis

Since there is no gold standard for the coreference systems and the parallel corpora that were used for this project, I opted for a manual analysis in the manner of Lapshinova-Koltunski et al. (2019). The lack of a gold standard also led to focusing on the comparison of the automatically annotated coreference chains between languages and systems alone instead of evaluating the performance of a single system in terms of whether all possible chains in a sample got annotated. Lapshinova-Koltunski et al. (2019) analysed in their work the annotation errors made by coreference systems. They presented four types of annotation incongruences which is how they called the differences in annotated parallel texts: (1) explication, (2) implication, (3) annotation interpretation and (4) annotation error. An (1) explication is the appearance of linguistic unit that is more specific in the translation than in the original text. An (2) implication happens when the original text contains a more specific linguistic unit than the translation. Annotation interpretation errors (3) occured when ambiguous cases in the original text create different identities and thus, different coreference chains and an annotation error (4) is an error due to manual annotation. For this project the concepts of (1) explication and (4) annotation error was used. Since the concept of source and target language does not apply to the corpora used for this project, I did not differentiate between explication (more mentions in target language) and implication (more mentions in source language). Instead, the concept of explication was divided into Dutch and German, depending on in which language were more annotations observed per coreference chain. Annotation interpretations (3) did not apply to this project since all annotation were done automatically. The extraction and analysis of the automatically generated coreference chains was done manually.

4.2 Results

Table 6 shows the results of the manual analysis.

	News				Literature			
	# NL	# DE	Avg NL	Avg DE	# NL	# DE	Avg NL	Avg DE
Tokens	16,579	15,140	921.06	841.11	16,701	14,957	1,391.75	1,246.42
Sentences	656	670	36.44	37.22	908	796	75.67	66.33
Sentence Length	467.86	413	25.99	22.94	235.34	234.31	19.61	19.53
Mentions	861	1161	47.83	64.5	904	697	75.33	58.08
Chains	354	402	19.67	22.33	597	449	49.75	37.42
Pronouns	232	338	12.89	18.78	461	382	38.42	31.83
Noun Phrases	422	565	23.44	31.39	452	311	37.67	25.92
Names	207	258	11.5	14.33	99	94	8.25	7.83

Table 6: Manual analysis of automatically annotated coreference chains for both corpora and both coreference systems. The "#" column indicates the total amount per corpus, the "Avg" column the average per sample.

The average sample of the News Commentary corpus contains less tokens than the Literature corpus (News: NL 921.06 / DE 841.11 versus Literature: NL 1,391.75 / DE 1,246.42) due to the nature of the data. Simultaneously, the Literature corpus contains about twice as many sentences per sample, but the sentences are slightly shorter on average compared to the News Commentary corpus (News: NL 36.44 / DE 37.22, Literature: NL 75.67 / DE 66.33 sentences per sample). The Dutch coreference system annotated singletons as a referent without further mentions. These were ignored and are not part of the row "Mentions" analysis. Following Lapshinova-Koltunski et al. (2019), the manually extracted coreference chains were divided into three categories: matching, overlapping and unpaired (see Table 7).

Match: Two coreference chains are considered "matching" if they contain the same amount of mentions. This also applies in case the type of mention changes: i.e. if an entity is mentioned by its name in one language but is paraphrased by a nominal phrase in the other language.

Overlap: If two coreference chains share at least one mention but differ in length, they are marked as "overlapping". Chains can be overlapping due to different reasons: (i) one or more mentions of one or both chains can be **wrongly annotated**, (ii) one or more mentions of one or both chains can be **missing** or, (iii) one chain is longer than another due to **explication**.

While (i) and (ii) are caused by the automatic annotation of the coreference system, (iii) usually occurs due to differences in language or translation, which will be discussed in Section 5.

Unpaired: If there is a coreference chain detected in one sample but not the other, this chain is considered "unpaired". The three types of incongruency mentioned in the previous paragraph also apply to unpaired coreference chains.

Chain category		News		Literature		Total	
		#	%	#	%	#	%
Matching	same number of mentions	107	25.97	40	5.12	147	12.32
Overlapping	more Dutch mentions	66	16.02	187	23.94	253	21.21
	more German mentions	173	41.99	249	31.88	422	35.37
Unpaired	Dutch chains	11	2.67	167	21.38	178	14.92
	German chains	55	13.35	178	22.79	233	19.53
Total number of chains		412	100	781	100	1,193	100

Table 7: Results of the manual extraction of the three different chain categories.

The coreference systems annotate more matching chains in the News Commentary corpus compared to the Literature corpus (107 / 25.97% versus 40 / 5.12%). The biggest chain category are the overlapping chains with more German mentions across both corpora (in total 422 / 35.37%). Next, the overlapping and unpaired chains were analysed according to the three types of incongruences (missing, wrong and explicit) (see Table 8):

Types of incongruences		News		Literature		Total	
		#	%	#	%	#	%
Missing	Dutch mention missing	121	35.07	129	13.75	250	20.19
	German mention missing	48	13.91	98	10.45	146	11.38
Wrong	Dutch mention wrongly annotated	22	6.38	206	21.96	228	17.77
	German mention wrongly annotated	59	17.10	263	28.04	322	25.10
Explicit	Additional Dutch mention	24	6.96	142	15.14	166	12.94
	Additional German mention	71	20.58	100	10.66	171	13.33
Total number of incongruences		345	100	938	100	1,283	100

Table 8: Manual analysis of incongruencies of automatically detected coreference chains (overlapping and unpaired chains).

The sum of incongruences does not equal the sum of overlapping and unpaired chains because each chain can contain one or more missing or wrong mentions or an explication due to language specifics or artistic freedom in translation. The full tables for both corpora can be seen in appendix, Table 9 and 10.

In general, there are more Dutch mentions missing than German mentions (NL 35.07% / DE 13.91% for the News Commentary corpus and NL 13.75% / DE 10.45% for the Literature corpus) but simultaneously more wrongly annotated German mentions in both corpora (NL 6.38% / DE 17.10% for the News Commentary corpus and NL 21.96% / DE 28.04% for the Literature corpus). There is one inconsistency in the incongruences: there are more German explications than Dutch ones in the News Commentary corpus (NL 6.96% / DE 20.58%) but more Dutch explications in the Literature corpus (NL 15.14% / DE 10.66%).

5 Discussion

As seen in Table 6, CorZu annotated more mentions and chains in the News Commentary corpus than dutchcoref while the opposite shows for the Literature corpus. These numbers are equally distributed over the pronouns, noun phrases and names of both corpora. The German coreference system seems to achieve a better performance on the News Commentary corpus and the Dutch coreference system on the Literature corpus, based on the missing and wrong annotations for German summing up to 107 (31.01%) missing or wrongly annotated mentions for News and 361

(38.49%) for the Literature corpus while dutchcoref sums up to 143 (41.45%) for the News corpus and 335 (35.71%) for the Literature corpus. This could be tied back to the reason for which the corpora were chosen: dutchcoref was trained on a literature corpus and CorZu on news articles.

Dutchcoref misses more mentions across both corpora but CorZu generates more wrong annotations. More explications are observed in the News Commentary corpus by CorZu and more in the Literature corpus by dutchcoref.

In total, the News Commentary corpus has far fewer errors in its annotation than the Literature corpus (250 versus 696) and more matching chains (107 / 25.97% versus 40 / 5.12%). Several differences in annotation can be traced back to the differences of the coreference systems or the genres.

5.1 Annotation differences due to the coreference systems

There are several reasons for explications to occur between the two coreference systems.[19]

For instance, dutchcoref successfully detects and annotates appositives while CorZu does not. This leads to longer coreference chains in the Dutch annotation and missing mentions for CorZu:

NL: [...] het begrip [...] dat [atoomwapens]$_1$ een [geostrategisch statussymbool]$_1$ blijven.
DE: [...] sein Verständnis von [Atomwaffen]$_1$ als eins geostrategisches Statussymbol.
EN: [...] the understanding that atomic weapons are a geostrategic status symbol.

Similar to appositives, CorZu also misses to annotate definite articles referring to clauses or full sentences:

NL: [...] [het juiste gebruik van antibiotica]$_2$. [Dit]$_2$ is [...]
DE: [...] [die richtige Anwendung von Antibiotika]$_2$. Das ist [...]
EN: [...] the correct use of antibiotics. This is [...]

CorZu oftentimes fails to catch reflexive pronouns in 3rd person singular, especially if the reflexive pronoun is placed before its noun:

NL: [...] terwijl [de partijen]$_3$ [zich]$_3$ [...]
DE: [...] während sich [die Parteien]$_3$ [...]
EN: [...] while the parties [...] themselves.

CorZu annotates nouns related to time, such as "day", "week", "month", "year", "yesterday", "tomorrow", and so on. Apart from most chains including temporal mentions being wrongly annotated, they were not included into the analysis:

NL: [...] maar weinig landen is de laatste jaren [...]
DE: [...] in [den letzten Jahren]$_4$ blieb es nur [...]
EN: [...] during the past years [...]

5.2 Annotation differences due to genre

As seen in Table 7, 8, 9 and 10, the Literature corpus has fewer matching chains than the News Commentary corpus (5.12% versus 25.97%). One possible reason for that are that the samples (chapters) from the Literature corpus are longer than the ones from the News Commentary corpus (NL 921.06 / DE 841.11 versus NL 1,391.15 / DE 1,246.42 tokens per sample) which makes the sample possibly more complex and provides more options for coreference. In addition to that, a lot of entities were split, partially due to paraphrased mentions, leading to almost twice the amount of chains on average (NL 49.75 / DE 37.42 versus NL 19.67 / DE 22.33 per sample).

19. The examples in this Section are extracted from the corpora used in this project.

Written in 1876, the writing style of Mark Twain's "The Adventures of Tom Sawyer" varies from that of modern literature. This might reflect in the number of wrong and missing annotations since not only the syntax slightly changed but also the orthography of some words. This leads to the annotation of words like "zoodra" (NL: "as soon as") as a mention while it in its modern orthography "zodra" was not annotated. In comparison, the news media is usually phrased to be cohesive and informative with little leeway for author(s) and translator(s).

Direct and indirect speech are problematic features for coreference resolution due to the shift of deixis. This makes the Literature corpus more difficult to annotate. On the other hand, the average sentence length of the News Commentary corpus was longer (NL 19.61 / DE 19.54 versus NL 25.99 / DE 22.94) which possibly raises the structural complexity of a sentence and therefore more difficult to annotate.

5.3 Annotation differences due to language and translation

Fictive literature tends to allow for more artistic freedom than news media: this leads to the paraphrasing of entity mentions which often do not get caught by the coreference systems. For instance, mentions like "de oude vrouw" (NL: "the old woman") were not annotated to be part of the same coreference chain as "tante Polly" (NL: "aunt Polly"), although both mentions refer to the same entity in "The Adventures of Tom Sawyer". In comparison, "die Bundeskanzlerin" (DE: "the German Chancellor") got annotated as a mention of "Angela Merkel" by both coreference systems in the News Commentary corpus.

Since not only the author of fictive literature uses artistic freedom but to a certain point also the translator, it occured that in some cases whole paragraphs between the Dutch and the German translation of the Literature corpus did not match in the matter that they were significantly shortened or did not appear in one translation at all, which leads to missing annotations in the translation which does not contain the clause or phrase:

> **NL:** [...] toen hij, heel in de verte **[eene opening]$_5$** ontekt had, waaruit een blauw stipje schemerde.
> **EN:** [...] as he discovered in far distance **an opening** from where a blue glimpse of light shimmered.
>
> **DE:** [...] als er in weiter Ferne einen schwachen Lichtschimmer entdeckte.
> **EN:** [...] as he discovered in far distance a weak glimpse of light.

In comparison, the news media allow less freedom which showed in the German and Dutch version of the News Commentary corpus being almost identical in syntax. This caused a significantly higher amount of unpaired chains in the Literature corpus (NL 21.38% / DE 22.79%) compared to the News commentary corpus (NL 2.67% / DE 13.35%).

Lastly, since Dutch tends to produce clause constructions that contract "waar" (NL: "where") and a preposition (similar to the English "wherein", "whereas") while German prefers relative clauses with pronouns, dutchcoref misses to annotate those coreferences, leading to more overlapping and unpaired chains with more German mentions:

> **NL:** [De OESO]$_6$, waarvan de leden in feite de 34 rijkste landen ter wereld zijn [...]
> **DE:** [Die OECD]$_6$, [die]$_6$ im Prinzip aus den 34 reichsten Volkswirtschaften der Welt besteht [...]
> **EN:** The OECD which consists of the 34 richest countries in the world [...]

6 Conclusions and future work

This project aimed to compare a state-of-the-art model from each Dutch and German and their annotation incongruences on two parallel corpora. As expected, the German coreference system annotates less incongruencies on the News Commentary corpus compared to the Literature corpus and the Dutch coreference system vice versa.

One next step for this project would be to manually annotate the two corpora to provide a gold standard. This can then be compared to the automatically annotated coreference chains in order to evaluate the performance of both coreference systems. It would also grant the possibility to analyse which kind of mentions and coreference chains were not analysed by either corerefence system.
Another analysis could exploit the information about the head of a phrase or coreference chain as well as the features of a form of both CoNLL output files in order to get a deeper understanding of the internal structure and the decision making of both coreference systems and which features of a language possibly lead to overlapping or unpaired chains. It would also be interesting to have a closer look at the annotation and linking of underspecified pronouns.

Other systems like the German sieve-based approach by Srivastava et al. (2018) could be applied on both corpora to compare to the Dutch sieve-based approach used in this project and CorZu. This would provide better comparable results across languages and a basis to evaluate two recent German coreference systems. Since Klenner and Tuggener (2011) and Klenner, Tuggener, and Fahrni (2010) claim that their coreference system is applicable to other languages, one could apply it to the Dutch corpora and the English version of both corpora and do a cross-linguistic analysis as well as a comparison in performance depending on the language version of the CorZu system.

Future coreference models could try to aim on cross-genre applications: instead of training on one genre specific corpus it might be interesting to train a coreference system on a genre-mixed corpus or train it on a variety of corpora and test the systems cross-genre. As seen in this project, German and Dutch face similar difficulties when facing coreference resolution.

Lastly, Anglocentrism continues to dominate computational linguistics which reflects in the performance of coreference systems. This is due to the less funding for other languages than English which leads to less researchers annotating data with gold standards and developing coreference systems. This explains why there are less deep learning coreference systems for other languages than English. Future work should aim to develop more diversified systems for other languages as well to catch up with the state-of-the-art models for English. Cross-linguistic research does not only help understanding the weaknesses and needs of a coreference system for a specific language but also helps understanding the similarities and differences between different languages.

		# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12	# 13	# 14	# 15	# 16	# 17	# 18	Averaged	Total
Tokens	NL	838	843	913	1,124	673	866	958	874	1,199	719	957	863	1,026	984	1,045	929	880	888	921.06	16,579
	DE	772	726	829	1,012	625	755	882	847	1,095	719	814	808	968	906	812	830	890	850	841.11	15,140
Sentences	NL	29	38	42	40	24	34	47	44	40	25	37	43	45	40	30	34	33	31	36.44	656
	DE	31	38	39	43	27	35	50	45	42	27	38	43	46	41	30	34	29	32	37.22	670
Avg Sen Length	NL	28.97	22.29	21.88	28.3	28.17	25.76	20.45	20.09	30.15	28.96	26.27	20.21	23.31	24.8	35.37	27.44	26.73	28.71	25.99	467.86
	DE	25.28	19.16	21.46	23.79	23.3	21.83	17.74	19.13	26.17	26.78	21.58	18.93	21.52	22.29	27.53	24.62	25.03	26.86	22.94	413
Mentions	NL	38	56	50	65	37	32	68	31	36	38	35	33	77	58	43	53	60	51	47.83	861
	DE	67	86	66	86	48	42	104	51	69	28	50	46	72	86	53	81	59	67	64.5	1,161
Chains	NL	14	27	20	27	16	15	26	11	22	10	15	18	25	19	20	22	26	21	19.67	354
	DE	17	28	21	27	16	15	28	19	26	14	21	19	29	27	20	27	24	24	22.33	402
Pronouns	NL	2	11	10	12	5	10	29	9	5	24	6	14	29	14	6	15	21	10	12.89	232
	DE	11	29	16	24	10	13	41	17	17	12	13	19	29	21	5	23	18	20	18.78	338
Nominals	NL	19	32	15	45	23	20	25	8	17	14	25	16	42	22	29	27	25	18	23.44	422
	DE	34	37	24	57	28	26	38	24	34	16	32	22	35	25	40	45	23	25	31.39	565
Names	NL	17	13	25	8	9	2	14	14	14	0	4	3	6	22	8	11	14	23	11.5	207
	DE	22	20	26	5	10	3	25	10	18	0	5	5	8	40	8	13	18	22	14.33	258
Matching chains		7	9	5	14	3	4	5	4	2	5	4	4	4	7	7	12	6	5	5.94	107
Overlapping chains	NL longer	0	4	6	6	5	3	1	3	4	2	3	1	10	4	2	4	5	3	3.67	66
	DE longer	7	13	8	7	8	7	20	3	16	2	8	14	14	7	8	13	5	13	9.61	173
Unpaired chains	NL	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3	0	7	0	0.61	11
	DE	3	0	1	0	0	1	2	7	4	4	6	0	0	8	3	4	9	3	3.06	55
Missing mentions	NL	4	2	5	5	2	4	14	3	12	2	6	7	9	10	8	5	8	15	6.72	121
	DE	1	2	4	1	2	2	0	3	4	2	2	0	8	1	2	2	6	6	2.67	48
Wrong mentions	NL	0	2	1	4	3	0	1	0	0	0	0	1	1	2	2	3	0	2	1.22	22
	DE	2	9	1	3	4	2	5	7	1	2	4	3	5	1	0	4	1	5	3.28	59
Explication	NL longer	0	0	1	0	0	1	1	0	0	0	3	0	0	6	2	8	2	0	1.33	24
	D longer	1	2	4	0	2	3	6	3	9	3	5	4	3	0	5	5	8	8	3.94	71

Table 9: News Commentary corpus per sample.

		MT 2	MT 7	MT 12	MT 15	MT 17	MT 18	MT 20	MT 21	MT 23	MT 28	MT 29	MT 33	Averaged	Total
Tokens	NL	1,785	2,027	1,763	1,868	1,218	1,243	813	1,736	1,087	952	1,103	1,106	1,391.75	16,701
	DE	1,746	1,898	1,473	1,694	1,143	1,089	775	1,566	979	753	937	904	1,246.42	14,957
Sentences	NL	102	142	92	104	63	46	41	91	52	59	76	40	75.67	908
	DE	78	112	89	85	67	39	36	86	58	52	57	37	66.33	796
Avg Sen Length	NL	17.68	14.58	19.28	18.07	19.36	27.17	20.24	19.19	20.94	16.35	14.73	27.75	19.61	235.34
	DE	22.51	17.15	16.62	19.97	17.08	28	21.94	18.25	16.93	14.73	16.59	24.54	19.53	234.31
Mentions	NL	86	118	92	112	69	73	29	96	63	49	43	74	75.33	904
	DE	65	87	76	68	56	68	39	74	39	38	43	44	58.08	697
Chains	NL	60	83	62	74	41	47	22	66	33	31	33	45	49.75	597
	DE	43	63	48	44	33	41	25	48	26	22	28	28	37.42	449
Pronouns	NL	39	58	39	46	30	34	14	40	28	24	16	93	38.42	461
	DE	31	47	35	26	24	31	23	33	14	20	20	78	31.83	382
Nominals	NL	40	52	47	54	39	34	11	50	31	23	19	52	37.67	452
	DE	27	33	31	32	32	29	14	30	19	15	18	31	25.92	311
Names	NL	7	8	6	12	0	5	4	6	4	2	8	37	8.25	99
	DE	7	7	10	10	0	8	2	11	6	3	5	25	7.83	94
Matching chains		8	3	1	4	5	3	1	3	4	1	3	4	3.33	40
Overlapping chains	NL longer	15	32	18	31	8	13	6	23	9	10	9	13	15.58	187
	DE longer	26	34	26	35	23	24	10	19	5	16	8	23	20.75	249
Unpaired chains	NL longer	13	30	16	10	5	11	11	26	14	7	17	7	13.92	167
	DE longer	16	27	12	21	16	17	11	22	11	9	7	9	14.83	178
Missing mentions	NL	13	23	9	25	6	7	4	11	6	4	13	8	10.75	129
	DE	7	22	17	9	5	3	8	12	1	2	4	8	8.17	98
Wrong mentions	NL	21	27	25	23	19	20	8	22	9	14	6	12	17.17	206
	DE	11	16	19	6	9	15	9	151	5	10	7	5	21.92	263
Explication	NL longer	13	13	13	18	7	17	6	17	11	7	6	14	11.83	142
	DE longer	11	14	8	13	13	10	2	7	7	3	8	4	8.33	100

Table 10: Literature corpus per chapter of Mark Twain's "The Adventures of Tom Sawyer".

References

Berkum, Jos J. A. van. 1996. “The psycholinguistics of grammatical gender: Studies in language comprehension and production.” PhD diss., Max Planck Institute for Psycholinguistics.

Bouma, Gosse, Gertjan van Noord, and Robert Malouf. 2001. *Alpino: Wide-coverage Computational Analysis of Dutch.* Technical report, University of Groningen, June. http://www.let.rug.nl/vannoord/papers/alpino.pdf.

Clark, Kevin, and Christopher D. Manning. 2015. “Entity-Centric Coreference Resolution with Model Stacking.” In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 1405–1415. Beijing, China: Association for Computational Linguistics. https://www.aclweb.org/anthology/P15-1136.

Cranenburgh, Andreas van. 2019. “A Dutch coreference resolution system with an evaluation on literary fiction.” *Computational Linguistics in the Netherlands Journal* 9:27–54. https://www.clips.uantwerpen.be/clinjournal/clinj/article/view/91.

Daiber, Joachim, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. “Improving Efficiency and Accuracy in Multilingual Entity Extraction.” In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics).*

De Clercq, Orphée, Véronique Hoste, and Iris Hendrickx. 2011. “Cross-Domain Dutch Coreference Resolution.” In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011,* 186–193. Hissar, Bulgaria: Association for Computational Linguistics, September. https://www.aclweb.org/anthology/R11-1026.

Hendrickx, Iris, Gosse Bouma, Walter Daelemans, and Véronique Hoste. 2013. “COREA: Coreference Resolution for Extracting Answers for Dutch.” In *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme,* edited by Peter Spyns and Jan Odijk, 115–128. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hoste, Véronique. 2005. “Optimization Issues in Machine Learning of Coreference Resolution.” PhD diss., Universiteit Antwerpen. Faculteit Letteren en Wijsbegeerte. https://biblio.ugent.be/publication/598135/file/1876875.

Hoste, Véronique, and Guy De Pauw. 2006. “KNACK-2002: a Richly Annotated Corpus of Dutch Written Text.” In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06).* Genoa, Italy: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2006/pdf/342_pdf.pdf.

Jurafsky, Dan, and James H. Martin. 2020. *Coreference Resolution.* https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf.

Klenner, M, Don Tuggener, and Angela Fahrni. 2010. “Inkrementelle Koreferenzanalyse für das Deutsche.”

Klenner, Manfred, and Don Tuggener. 2011. “An Incremental Entity-Mention Model for Coreference Resolution with Restrictive Antecedent Accessibility.” In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011,* 178–185. Hissar, Bulgaria: Association for Computational Linguistics, September. https://www.aclweb.org/anthology/R11-1025.

Krug, Markus, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. “Rule-based Coreference Resolution in German Historic Novels.” In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature,* 98–104. Denver, Colorado, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W15-0711.

Lapshinova-Koltunski, Ekaterina, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. "Cross-lingual Incongruences in the Annotation of Coreference." In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference,* 26–34. Minneapolis, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-2805.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules." *Computational Linguistics* 39 (4): 885–916. https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00152.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 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,* 28–34. Portland, Oregon, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W11-1902.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. "End-to-end Neural Coreference Resolution." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 188–197. Copenhagen, Denmark: Association for Computational Linguistics. https://www.aclweb.org/anthology/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer. 2018. "Higher-Order Coreference Resolution with Coarse-to-Fine Inference." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* 687–692. New Orleans, Louisiana: Association for Computational Linguistics. https://www.aclweb.org/anthology/N18-2108.

Noord, Gertjan van. 2006. "At Last Parsing Is Now Operational," 20–42. http://www.let.rug.nl/vannoord/papers/taln.pdf.

Pradhan, Sameer, 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,* 1–40. Jeju Island, Korea: Association for Computational Linguistics, July. https://www.aclweb.org/anthology/W12-4501.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes." In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task,* 1–27. Portland, Oregon, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/W11-1901.

Schoen, Anneleen, Chantal van Son, Marieke van Erp, and Hennie van Vliet. 2014. *NewsReaderdocument-level annotation guidelines: Dutch.* Technical report, VU University. http://www.newsreader-project.eu/files/2013/01/8-AnnotationGuidelinesDutch.pdf.

Schuurman, Ineke, Véronique Hoste, and Paola Monachesi. 2010. "Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch." January. http://www.lrec-conf.org/proceedings/lrec2010/pdf/162_Paper.pdf.

Sennrich, Rico, Martin Volk, and Gerold Schneider. 2013. "Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis." In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013,* 601–609. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. https://www.aclweb.org/anthology/R13-1079.

Srivastava, Ankit, Sabine Weber, Peter Bourgonje, and Georg Rehm. 2018. “Different German and English Coreference Resolution Models for Multi-domain Content Curation Scenarios.” In *Language Technologies for the Challenges of the Digital Age,* edited by Georg Rehm and Thierry Declerck, 48–61. Cham: Springer International Publishing.

Telljohann, H, E Hinrichs, and S Kübler. 2004. *The Tu ba-D/Z Treebank : Annotating German with a Context-Free Backbone.* Lisbon. `https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/resources/corpora/tueba-dz/`.

Tiedemann, Jörg. 2012. “Parallel Data, Tools and Interfaces in OPUS.” In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* edited by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA).

Tuggener, Don. 2016. “Incremental Coreference Resolution for German.” PhD diss., University of Zurich. `http://pub.cl.uzh.ch/purl/coreference-resolution`.