Computational Linguistics in the Netherlands Journal 9 (2019) 27–54 Submitted 09/2019; Published 12/2019

A Dutch coreference resolution system with an evaluation on literary fiction

Andreas van Cranenburgh A.W.VAN.CRANENBURGH@RUG.NL

University of Groningen, The Netherlands

Abstract

Coreference resolution is the task of identifying descriptions that refer to the same entity. In this paper we consider the task of entity coreference resolution for Dutch with a particular focus on literary texts. We make three main contributions. First, we propose a simplified annotation scheme to reduce annotation effort. This scheme is used for the annotation of a corpus of 107k tokens from 21 contemporary works of literature. Second, we present a rule-based coreference resolution system for Dutch based on the Stanford deterministic multi-sieve coreference architecture and heuristic rules for quote attribution. Our system (`dutchcoref`) forms a simple but strong baseline and improves on previous systems in shared task evaluations. Finally, we perform an evaluation and error analysis on literary texts which highlights difficult cases of coreference in general, and the literary domain in particular.

The code of our system is made available at `https://github.com/andreasvc/dutchcoref/`

1. Introduction

Coreference resolution is the task of identifying spans in text (mentions) that refer to the same entity. The task is concerned with persons and objects (i.e., entities) referred to by names, pronouns, and descriptions (we do not consider coreference of events and actions in this work). In the following example, brackets indicate referring expressions, while coindexes indicate coreferent expressions:[1]

(1) *‘[Het]$_1$ is [een oude man]$_1$, denk [ik]$_2$,’ zei [ze]$_2$. ‘Heb [je]$_2$ [hem]$_1$ dan gezien?’ vroeg [ik]$_3$.*
‘[It]$_1$ is [an old man]$_1$, [I]$_2$ think,’ [she]$_2$ said. ‘Did [you]$_2$ actually see [him]$_1$?’ [I]$_3$ asked. (Voskuil, *De Buurman*)

Coreference resolution is important for natural language understanding in particular, and language processing tasks beyond the sentence level in general. This paper presents a system for coreference resolution designed to handle book-length documents such as literary texts.

Applying computational methods to the analysis of literature has become increasingly popular under the moniker ‘distant reading’ (Moretti, 2013; Boot, 2014). However, such analyses are, with few exceptions, characterizing texts in terms of surface features—typically word frequencies. While the breadth of research questions that can be addressed using word frequencies is impressive and surprising (for applications to the analysis of style, genre, and authorship cf. Rybicki et al., 2016), a fundamental limitation is that discourse-level phenomena cannot be captured in such bag-of-words models. This paper is part of a research agenda to analyze the characters in fiction. Analyzing coreference will allow us to address questions such as:

- How many distinct entities are referred to in a given amount of text?
- What is the social network of interacting characters?
- What are the attributes and actions associated with a character?

However, automatic coreference resolution is challenging; a typical error rate is at least 30%, but often much higher. Moreover, literature presents its own challenges (Rösiger et al., 2018), such

[1] Throughout this paper, translations are my own.

as long coreference chains, dialogue, and frequent use of pronouns. To get a precise estimate of the performance of a coreference system on literature, we develop a rule-based coreference system, annotate fragments of Dutch novels, and evaluate our system on them.

Many approaches to coreference resolution are based on machine learning; the best performing systems on standard benchmarks use deep learning.[2] However, we use a rule-based approach, based on the Lee et al. (2011, 2013) system which won the CoNLL 2011 shared task (Pradhan et al., 2011). This has several advantages (the first three items are adapted from Krug et al. 2015):

- It is easier to correct specific mistakes by modifying or adding rules, since it is possible to identify the particular rule responsible for a mistake.
- It is not necessary to annotate large volumes of text as training material; the only data used are generic lexical resources on gender and animacy.
- The system can be used as a crutch for semi-supervised annotation by correcting the output of the system and using those annotations to train statistical or neural models.
- High-precision, handcrafted rules typically generalize better to new words and domains than machine learning systems trained on texts from a specific domain (Lee et al., 2013, p. 886).

In summary, the contributions of this paper are:

- A simplified annotation scheme for Dutch coreference (Section 3).
- A rule-based coreference resolution system for Dutch (Section 4).
- An evaluation & error analysis on novels that we annotated for coreference (Section 5, 6).

2. Background

Reference is a relation between a referring expression in text and an entity in the real or mental world. Two or more referring expressions are said to co-refer when they refer to the same entity. Coreference is therefore a relation that holds for two or more referring expressions in text.

For a general overview of coreference resolution, cf. Poesio et al. (2016). Ng (2010) provides an overview of work on coreference resolution in the field of Natural Language Processing (NLP) up to 2010, while more recent machine learning approaches are covered in Ng (2017).

Task description Coreference resolution is a structured prediction task of partitioning detected mentions into entity clusters. A coreference system is typically part of an NLP pipeline including named-entity recognition and syntactic parsing. The input for coreference resolution is then a sequence of syntactic parse trees, combined with auxiliary data from lexical resources. Rule-based systems identify mentions and links by directly matching on parse tree configurations and attributes. Learning-based systems extract features from parse trees that are fed into supervised models such as classifiers or ranking systems.

Architectures The main architectures are:

mention-pair A binary classifier considers all relevant combinations of mentions, and classifies them as coreferent or not.

mention-ranking For each anaphor mention, all potential antecedents are considered and scored at once; the model assigns at most one antecedent.

entity-based Mentions are partitioned into clusters, with each cluster corresponding to an entity; as such, entities are modeled and scored directly.

Recent work on coreference is dominated by neural systems (Wiseman et al., 2015; Clark and Manning, 2016). These systems avoid much of the rule and feature engineering characteristic of earlier architectures since distributed representations are automatically learned; i.e., embeddings of words,

[2] Cf. `http://nlpprogress.com/english/coreference_resolution.html`

mentions, and entities are induced from the training data. End-to-end coreference resolution (Lee et al., 2017b) does not rely on a syntactic parser, a separate mention detection system, or any other external resource. The state of the art in coreference resolution exploits contextualized word embeddings (Lee et al., 2018; Joshi et al., 2019). However, the success of these architectures is determined by the quantity and quality of training data, tuning of hyperparameters, and requires a considerable investment of computing power. In particular, a disadvantage of neural methods for coreference is that they are memory intensive and must deal with longer documents by working on short chunks (Joshi et al., 2019), which may be an issue with literary texts.

Shared tasks Early work on coreference was restricted to pronoun resolution (e.g., Lappin and Leass, 1994). The task of 'unrestricted' coreference, i.e., including coreference of nominals and names in addition to pronouns, was popularized in the MUC-6 and MUC-7 shared tasks (Grishman and Sundheim, 1996; Hirschman and Chinchor, 1998), followed by the ACE shared tasks (Doddington et al., 2004). The shared tasks of SemEval (Recasens et al., 2010) and CoNLL (Pradhan et al., 2011, 2012) introduced the current evaluation practices and multilingual datasets. The datasets of these shared tasks included certain types of events in addition to entities, but the focus was on the latter.

Quote attribution Quote attribution is the task of identifying the speaker (and addressee) of direct speech spans. In example (1), the speakers are as follows:

(2) a. *'[Het]$_1$ is [een oude man]$_1$, denk [ik]$_2$,' zei [ze]$_2$.* Speaker: *[ze]*$_2$
'[It]$_1$ is [an old man]$_1$, [I]$_2$ think,' [she]$_2$ said.
b. *'Heb [je]$_2$ [hem]$_1$ dan gezien?' vroeg [ik]$_3$.* Speaker: *[ik]*$_3$
'Did [you]$_2$ actually see [him]$_1$?' [I]$_3$ asked.

A quote may be attributed to a mention, entity, or a pre-defined character name; potential speakers are either given or automatically detected. Quote attribution is relevant to coreference resolution since the speaker and addressee of dialogue turns must be known to resolve first and second person pronouns in quoted speech correctly.

Elson and McKeown (2010) report results of a quote attribution system applied to literature. O'Keefe et al. (2012) present a sequence labeling system for quote attribution. O'Keefe et al. (2013) present results on the impact of coreference resolution performance on quote attribution. Almeida et al. (2014) report on a joint model of coreference resolution and quote attribution. Muzny et al. (2017b) present a rule-based and statistical quote attribution system. Steinbach and Rehbein (2019) present a system for projecting quote attribution annotations from one language to another.

Beyond direct speech, there are the harder problems of indirect speech and free indirect discourse (cf. e.g., Hammond et al., 2013).

(3) a. indirect speech: John said that he would take care of it.
b. free indirect discourse: John sat down and composed himself. What a day it had been, could he keep it together?

Direct and indirect speech both contain an explicit marker such as "he said" or "she thought"; they differ in whether the speech is quoted verbatim or paraphrased as a clause integrated in the sentence. Free indirect discourse leaves the distinction between the voice of the narrator and the voices of characters implicit.

Pareti et al. (2013) and Brooke et al. (2017) present systems for indirect speech attribution and detection. This paper only considers attribution of direct speech.

Dutch coreference Work on Dutch coreference resolution started with the KNACK 2002 corpus of magazines (Hoste, 2005; Hoste and Pauw, 2006), on which a mention-pair system was trained and evaluated. This was followed by the Corea project (Bouma et al., 2007; Hendrickx et al., 2008a,b), which annotated more data and further developed the aforementioned mention-pair system. The largest Dutch coreference annotation effort is that of the 1 million word SoNaR-1 dataset (Schuurman et al., 2010). De Clercq et al. (2011) presents cross-domain coreference results with this corpus.

The Dutch part of the NewsReader project (Schoen et al., 2014) has also produced coreference annotations and systems. The NewsReader project is an effort to annotate the events, entities, and temporal expressions in news articles, as well as relations between them; entity coreference is one of these relations. These annotations were used in the CLIN 26 shared task,[3] which included a track for Dutch entity coreference.

Computational Linguistics for Literature There has been considerable work on computational linguistics for literature (cf. the workshops of the same name; Elson et al. 2012, 2013; Feldman et al. 2014, 2015). However, systems presented in the field of computational linguistics are typically trained and evaluated on newswire; adapting state-of-the-art models to work well in the domain of literature presents a considerable open challenge (Bamman, 2017).

Most work on computational linguistics for literature focuses exclusively on English language texts (anglocentrism). Moreover, for copyright and convenience reasons, most work studies readily available nineteenth century texts instead of modern or contemporary novels.

There has been work on literary coreference, characters and quote attribution in particular. Elson et al. (2010) automatically extract social networks from literary novels and test specific hypotheses about social networks from literary studies. Bamman et al. (2014) present a model of literary character in English nineteenth century novels; they also make an NLP pipeline available for parsing, NER, quote attribution, and pronoun resolution of book-length documents called BookNLP. The system presented in this paper provides a similar pipeline for Dutch texts. Krug et al. (2015) present a system for coreference resolution of German evaluated on classical literary texts—they implement a system based on the Lee et al. (2011, 2013) architecture, the same approach as used in this paper. Muzny et al. (2017a) test hypotheses from literary studies about dialogue in novels. Bamman et al. (2019b) present a dataset of 210k tokens from 100 English nineteenth century novels in which ACE entities are annotated; Bamman et al. (2019a) additionally annotate this dataset for coreference. Two notable differences with Krug et al. (2015) and Bamman et al. (2019b) are that we focus on contemporary novels, and that the fragments of the novels we annotate are longer (8,000 tokens per novel for the test set, compared to 130 sentences and 2,000 tokens, respectively).

3. Coreference annotation scheme

Annotation of coreference proceeds in two stages: identification of mentions and annotation of coreference relations. Mentions (viz. referring expressions) are spans of text that refer to an entity; spans of text that *potentially* refer to an entity are called markables. Markables are defined syntactically while identifying mentions requires discourse understanding. Non-referring expressions such as pleonastic pronouns or demonstratives referring to verbal clauses are markables but not mentions. A singleton is a special kind of mention, viz. a mention of an entity which is not referred to again. This contrasts with coreferent (or anaphoric) mentions, which are referred to two or more times in the text. A coreference relation indicates that two mentions refer to the same entity (hence, *co*-reference). There are two ways to view coreference relations: (a) the anaphor-antecedent view holds that an anaphoric mention may have a directed link to a particular antecedent mention, while (b) the entity-cluster view holds that coreference indicates that two or more mentions are equivalent and members of the same entity cluster. This paper adopts the latter view. This view is also adopted in the SemEval and CoNLL shared tasks (Recasens et al., 2010; Pradhan et al., 2011, 2012). The anaphor-antecedent view is adopted by most annotation efforts; these annotations may be converted to the tabular SemEval/CoNLL format for purposes of training and evaluation.

The annotation guidelines of the Corea project (Bouma et al., 2007) describe the main annotation scheme for Dutch coreference. They were based on the guidelines presented in Hoste (2005), which

[3] `http://wordpress.let.vupr.nl/clin26/shared-task/`

```
#begin document (example); part 000
0   '       -
1   Ik         (0)
2   ben        -
3   de         (0
4   directeur  -
5   van        -
6   Fecalo     (1)|0)
7   ,          -
8   van        -
9   hierachter -
10  ,          -
11  '          -
12  zei        -
13  hij        (0)
14  .          -
0   '          -
1   Mag        -
2   ik         (0)
3   u          (2)
4   iets       -
5   vragen     -
6   ?          -
7   '          -

0   Ik         (2)
1   vroeg      -
2   hem        (0)
3   binnen     -
4   te         -
5   komen      -
6   .          -

#end document
```

Figure 1: The SemEval/CoNLL tabular format for coreference. The numbers in the last column identify coreference clusters, the start and end of spans is indicated with parentheses.

are in turn derived from the guidelines of the MUC-6 and MUC-7 datasets for English (Grishman and Sundheim, 1996; Hirschman and Chinchor, 1998). The Corea guidelines were later used for the coreference annotation of the 1-million word SoNaR corpus (Schuurman et al., 2010).

The NewsReader project introduced another annotation scheme for Dutch coreference (Schoen et al., 2014), integrated in a larger effort to annotate the events, entities, and temporal expressions in news articles, as well as relations between them. Their (entity) coreference annotation guidelines are based on the Corea guidelines, but differ mainly in that they opt to only annotate particular entities[4] (as in the ACE shared task for English; Doddington et al., 2004) instead of extracting all potentially referring expressions (markables) automatically based on syntactic queries.

For the annotation efforts reported in the present paper, we started with the Corea annotation scheme. However, several issues encountered during annotation led us to formulate a variant of this annotation scheme. Annotation is simplified in several respects to reduce the effort of annotation, while other changes make the annotation more informative by including selected phenomena. The rest of this section describes the simplified annotation scheme and enumerates the differences with the aforementioned annotation schemes.[5]

3.1 Representation

Coreference links are undirected and untyped: all mentions in a cluster are taken to refer to the same entity. In other words, mentions belong to coreference clusters which are equivalence classes. The specific antecedent of an anaphor, the type of entity, the type of coreference relation, and the head of a mention are not part of the annotation. Effectively, the annotation scheme follows the data model of the tabular SemEval and CoNLL 2011/2012 formats. This strategy is also proposed by Rösiger et al. (2018). Conceptually, mentions are clustered into entities:

[4] The entity types are: person, location, organization, product, financial, mixed (Schoen et al., 2014).

[5] Our annotation scheme is documented at `https://github.com/andreasvc/dutchcoref/blob/master/annotationguidelines.pdf`

(4) *'[Ik]$_0$ ben [de directeur van [Fecalo]$_1$]$_0$, van hierachter,' zei [hij]$_0$. 'Mag [ik]$_0$ [u]$_2$ iets vragen?' [Ik]$_2$ vroeg [hem]$_0$ binnen te komen.* (Voskuil, *De Buurman*)
'[I]$_0$ am [the director of [Fecalo]$_1$]$_0$, in the back,' [he]$_0$ said. 'Can [I]$_0$ ask [you]$_2$ something?' [I]$_2$ asked [him]$_0$ to come in.

$$E_0 = \{\text{Ik, de directeur van Fecalo, hij, ik, hem}\}$$
$$E_1 = \{\text{Fecalo}\}$$
$$E_2 = \{\text{u, Ik}\}$$

The SemEval and CoNLL shared tasks introduced a tabular representation for coreference, cf. Figure 1. In the shared tasks, extra columns provided POS, parse tree, and named entity features; such extra columns are however not included in our annotation efforts.

3.2 Mentions

Mentions are manually corrected: all mentions that refer to a person or object are annotated (including singletons), while other spans are excluded. We include generic pronouns and selected indefinite pronouns:

(5) a. Generic pronouns: *[Je] weet maar nooit.*
[You] never can tell.
b. Indefinite pronouns: *"Ik zie [iets]$_1$," zegt Jan. [Een eiland]$_1$ verschijnt aan de horizon.*
"I see [something]," says Jan. [An island] appears on the horizon.

Indefinite pronouns require judgment, since in most causes they do not have a specific referent.

We follow the principle that mentions must refer to an identifiable real or mental entity. We therefore exclude pleonastic pronouns, time-related expressions, and mentions that do not refer to identifiable entities due to being in a modal, negative, figurative, or idiomatic context (the relevant non-mentions are underlined):

(6) a. Pleonastic pronouns: *Het regent*
It is raining
b. Time-related expressions: *gisteren, de langste dag van de zomer*
yesterday, the longest day of summer
c. Modal or negative context: *Maar nee, geen glimmende regenjassen en gleufhoeden 's nachts aan [de deur van [[mijn] hotelkamer]] , geen enkele toespeling op [mijn] geschrijf van de kant van [het Presseamt] , [waar] [ik] [mij] bij ieder bezoek aan [de DDR] nederig meldde , nooit gezeur met visa , . . .* (Springer, Quadriga)
But no, no shiny raincoats and fedoras at night at the door of my hotel room, no allusion on my writing from the Pressamt, where I humbly report every time I visit the DDR, never any visa troubles , . . .
d. Idioms: *Wat is er aan de hand?*
What's the big idea?

The NewsReader annotation guidelines (Schoen et al., 2014) also manually annotate mentions but are more selective about which entities are considered. Other annotation schemes either leave out singletons (e.g., OntoNotes: Hovy et al., 2006; Pradhan et al., 2013) or include all noun phrases (e.g., Corea: Bouma et al., 2007). By manually correcting mentions, our annotations can be used as training material for a classifier that distinguishes coreferent mentions, singletons, and non-mentions.

Mention boundaries Discontinuous mentions and other difficult mention boundaries are avoided by leaving out discontinuous material from the mention (i.e., only the continuous span with the head noun is annotated as mention). While the Corea and NeswReader annotation guidelines prescribe that the complete span of a discontinuous constituent should form the span of a mention, this is incompatible with the tabular SemEval/CoNLL format which only allows continuous spans. This leads to compromises where either the discontinuous spans are carefully annotated but not used in coreference systems and evaluations that cannot handle them, or the discontinuous mention is annotated with the intervening material included. It is difficult to annotate such mentions consistently since discontinuous material is easy to overlook and may lead to arbitrarily long mention spans. Such cases are difficult for annotators as well as for automatic parsers.

Since relative clauses are often discontinuous, for the sake of consistency we opt to always cut off relative clauses at the relative pronoun to avoid overly long mentions and inconsistencies. The following example sentence from Bouma et al. (2007) illustrates the annotation of relative clauses using the Corea guidelines and using our guidelines:

(7) a. Our guidelines:
[President Alejandro Toledo]$_1$ reisde dit weekend naar Seattle voor een gesprek met [Microsoft topman Bill Gates]$_2$. [Gates]$_2$, [die]$_2$ al jaren bevriend is met [Toledo]$_1$, investeerde onlangs zo'n 550.000 Dollar in Peru.
[President Alejandro Toledo]$_1$ travelled this weekend to Seattle for a discussion with [Microsoft executive Bill Gates]$_2$. [Gates]$_2$, [who]$_2$ has been friends with [Toledo]$_1$ for years, recently invested about 550,000 Dollar in Peru.

b. Corea guidelines:
[President Alejandro Toledo]$_1$ reisde dit weekend naar Seattle voor een gesprek met [Microsoft topman Bill Gates]$_2$. [Gates, die al jaren bevriend is met [Toledo]$_1$]$_2$, investeerde onlangs zo'n 550.000 Dollar in Peru.

The following sentences illustrate the issues avoided by minimal mention boundaries:

(8) a. Relative clauses can be discontinuous: (Springer, *Quadriga*)
Ik kan (...) getrouw [de indrukken]$_1$ weergeven [die]$_1$ deze feiten hebben achtergelaten.
(...) I can faithfully represent [the impressions]$_1$ [which]$_1$ these facts have left me with.

b. Relative clause can be arbitrarily long: (Le Carré, *Our kind of traitor*)
En dit was [de Perry]$_1$ [die]$_1$ vroeg op die ochtend in mei, voordat de zon te hoog stond om nog te kunnen spelen, op de beste tennisbaan in het beste door de recessie getroffen vakantieoord in Antigua stond, met de Russische Dima aan de ene kant van het net en Perry aan de andere.
And this was [the Perry]$_1$ [who]$_1$ early that morning in May, before the sun was too high to play, stood on the best tennis court in the best recession-hit holiday resort in Antigua, with the Russian Dima on one side of the court and Perry on the other.

Other annotation efforts (e.g., MUC, Grishman and Sundheim 1996; ACE, Doddington et al. 2004) annotate both minimal and maximal spans. As a pragmatic consideration to reduce complexity and annotation effort, we annotate only a single span for each mention. The issue of mention boundaries influencing coreference evaluation is investigated by Moosavi et al. (2019), who conclude that using minimal spans in evaluation avoids noise from parser errors in coreference evaluation. We take this conclusion one step further by arguing it should not only apply during evaluation, but also during annotation. However, we do not extend the principle of minimal span boundaries to prepositional postmodifiers; these cause less parsing issues and are informative; we therefore include them in mentions.

3.3 Coreference relations

Coreference relation types Only a single type of coreference relation is annotated, comprising identity/strict coreference, predicate nominals, appositives, and bound anaphora:

(9) a. Strict: *[Jan]$_1$ struikelt. [Hij]$_1$ is boos.*
[Jan]$_1$ trips. [He]$_1$ is upset.
b. Predicative: *[Jan]$_1$ is [de directeur]$_1$*
[Jan]$_1$ is [the director]$_1$
c. Appositive: *[Jan]$_1$ [de schilder]$_1$*
[Jan]$_1$ [the painter]$_1$
d. Bound anaphora: *[Iedereen]$_1$ heeft er [[zijn]$_1$ mening] over.*
[Everyone]$_1$ has [his]$_1$ opinion about it.

The motivation for not annotating the type of coreference relation is that the non-strict relations are less common and hard to distinguish (e.g., Hendrickx et al., 2008a, sec. 2.2). While the distinction is linguistically interesting, it is arguably not crucial for most applications. Bridging coreference (part/whole, subset/superset relations) is outside the scope of this work and therefore not annotated. Bridging relations are harder to annotate and resolve than other relations because they depend on an implicit inference (bridge) derived from world knowledge.

Precise constructs Syntactically obvious coreference links are included in the annotation. Specifically, reflexive, reciprocal, and relative pronouns are annotated for coreference. The motivation is that for any given verb predicate, its syntactic arguments should be linked with entities, such that it is possible to establish *who did what to whom* in a document. For example:

(10) [The man]$_1$ [who]$_1$ sold the world [. . .].

Syntactically, *who* is the agent of sold, but without the coreference link to *the man*, we do not have further information about this entity, for example that the agent is male and singular (and any other information that may be introduced later in the discourse through further mentions of this entity).

Excluded coreference phenomena We exclude coreference to verb phrases and clauses, since our annotation is restricted to entity coreference. Time-indexed coreference receives no special treatment. The following sentence was true at a specific interval in time:

(11) [Barack Obama]$_1$ is [president of the United States]$_1$.

The coreference relation should arguably be restricted to that interval as well. A proper treatment of time-indexed coreference relations is challenging and outside of the scope of this work. Corea makes a compromise of annotating a flag identifying time-indexed coreference relations, without specifying the time of their validity.

Difficult coreference relations Generic mentions are only linked when they refer to the same generic referent in a paragraph. Humorous and figurative references are special cases. These are resolved by applying the principle of always annotating the intended and not the literal referent. For example:

(12) a. Metonomy: *De VS bombardeert meerdere doelen. Moskou reageert woedend.*
The US bombs multiple targets. Moscow is furious.

Here *Moskou* refers to the Russian government, not the city.

An interesting special case is the use-mention distinction from analytic philosophy (Quine, 1940, pp. 23–25). A name is typically used to refer to a person, but can also be used in a meta-linguistic statement such as “John is a common name.” These are distinguished as separate entities (John the person, John the name).

3.4 Harmonization

Introducing a new annotation scheme (variant) has obvious disadvantages. Annotation is expensive and a proliferation of incompatible annotation schemes should be avoided, to maximize the return on investment of annotation efforts. In particular, the SoNaR corpus contains a large amount of coreference annotations using the Corea guidelines. However, we believe that the differences between our annotation scheme and that of Corea can be harmonized. Some of the differences can be converted automatically (e.g., linking relative pronouns), while others (mention annotation) can be solved in a semi-supervised manner by training a mention detection system and correcting its output. However, working out these details is beyond the scope of this paper.

4. Coreference resolution system

The coreference resolution system involves mention detection, quote attribution, and adding coreference links with rule-based, deterministic sieves.

4.1 Preprocessing

Texts are preprocessed and parsed with the Alpino parser (Bouma et al., 2001; van Noord, 2006), which also performs named-entity recognition (NER). We use the Alpino tokenizer for word tokenization and sentence splitting. Our system also uses information on paragraphs, where available. When preprocessing novels, each sentence is assigned an identifier of the form `n-m` where n is a paragraph number and m a sentence number.

4.2 Mention detection

Candidate mentions are extracted from parse trees. Heuristic rules adjust spans and filter out noun phrases that do not refer to persons or objects. If there are multiple candidate mentions with the same head word, only the longest mention is kept. Gender and animacy of mentions is detected using parse tree features and lexical resources. Lee et al. (2013, sec. 3.1) argue that recall is more important than precision for mention detection. However, we find that maintaining good precision of mention detection is also important. An incorrect mention leads to precision errors on coreference links, since pronouns that are actually pleonastic should not be resolved, and incorrectly identified mentions should not be linked to other mentions. We introduce a set of filtering rules to improve precision for mention detection.

Mention boundaries Mention spans are adjusted with the following rules:

1. Gaps: If the span of the mention has a gap (due to punctuation or a discontinuous constituent), drop everything except the continuous span with the head word.
   Exception: commas preceded by conjunct, adjective, or location are allowed:

   (13) a. Alice, Bob, and Charles
   b. The big, red ball
   c. San Jose, California

2. Relative pronoun: only keep span before relative pronoun
3. Verbal complements: do not include clauses
4. Name-initial appositives are split into two mentions, while appositives with the name in second position are a single mention:

   (14) a. *[Jan], [de schilder]*
   [John], [the painter]

b. *[acteur John Cleese]*
[actor John Cleese]

5. Titles are included with the mention of a name

(15) *mevrouw [Steele] ⇒ [mevrouw Steele]*
miss [Steele] ⇒ [miss Steele]

6. Punctuation is dropped if it is at the end or beginning of the mention

Filtering mentions Mentions matching one of the following rules are eliminated:

1. Measure phrases
2. NPs which function as determiner (e.g., "A few" ...)
3. NPs headed by wh-words
4. Mentions consisting solely of indefinite pronouns (e.g., something, anything)
5. Temporal expressions, partitives, and quantifiers
6. Head is a nominalization of a verb
7. Non-referential expressions

Non-referential expressions Pleonastic pronouns and pronouns referring to verbal clauses should be detected and excluded. Pleonastic pronouns consist of weather verbs[6] and fixed expressions. It turned out that devising rules for the pleonastic pronoun *het* (it) is difficult. While the Alpino parse trees contain a lot of detail, they do not distinguish pleonastic pronouns from other pronouns. It is possible to collect a list of weather verbs and fixed expressions, but the there are many more cases which are hard to capture. Compare:

(16) a. *Het is te laat.*
It is too late.
b. *Ga nou maar, [het]$_1$ is [een lange rit]$_1$ en ik wil eigenlijk niet dat je te laat komt.* (James, *Vijftig tinten grijs*)
Just go, [it]$_1$ is [a long ride]$_1$ and actually I don't want you to be late.

Previous work has proposed training a classifier to detect pleonastic pronouns (e.g., Mitkov et al., 2002; Müller, 2006; Lee et al., 2017a). Experiments with a random forest classifier did improve on the rule-based baseline, but did not improve over the majority baseline of always considering the *het* pronoun as non-referential. This is an instance of the difficulty of beating the majority baseline for a classification task.

Similarly, the demonstrative pronoun *dat* (that) is used to refer to verbal clauses in the majority of cases, and our system does not consider event coreference; to avoid spurious mentions and links, we therefore also remove this pronoun from the mentions. Since *het* and *dat* do not refer to persons, discarding them does not affect coreference chains for characters and dialogue extraction. For applications where references to inanimate entities are crucial, a better classifier should be explored.

Gender and animacy information An important source of information for coreference decisions is whether a mention refers to an entity that is human or not, male or female, and singular or plural. When these features of two mentions are incompatible, they are unlikely to be coreferent. This information needs to be inferred for the three types of mentions: pronouns, names, and nouns.

For pronouns this information is partially grammatically marked; e.g., *het* (it) is non-human while *hij* (he) is male and human. The pronoun system of Dutch (cf. Haeseryn et al., 1997; Donaldson, 2017) differs in several relevant respects from the English pronoun system. Briefly, biological gender of persons is expressed by gendered personal pronouns *hij/zij* (he/she) and corresponding possessive pronouns *zijn/haar* (his/her), but these pronouns may also express the linguistic gender of inanimate

[6] Weather verbs such as *it's raining* and *it snows* require a pleonastic pronoun; Dutch exhibits the same pattern.

referents, where English would use it/its. While *zijn* is commonly used for linguistically gendered referents, the use of *haar* to express the linguistic gender of inanimate referents is less common and perceived as archaic, although it is more frequently attested in Flemish dialects. Animals may be referred to by gendered pronouns *hij/zij* (he/she), where English would use "it." The singular third person female pronoun has overlapping forms with the third person plural pronoun: *zij* (she/they) can be female but also plural, in which case it may have a non-human referent.

Names are recognized and classified using named-entity recognition (NER), performed by Alpino; i.e., person names are distinguished from organizations, locations, and other names. The NER component of Alpino is not trained on literary text, a domain which is known to exhibit specific challenges for NER (van Dalen-Oskam et al., 2014; Dekker et al., 2019). To assign gender to names, and to handle names not recognized by Alpino, we use external datasets. We use a list of all first names that occur at least 500 times in the Netherlands, separated by gender (male or female).[7] For names that are not part of this list, we also look up the name in the dataset made available by Bergsma and Lin (2006). This dataset contains names scraped from a large amount of text from the web, categorized by gender and number using pattern-based heuristics. Specifically, for each name, the dataset lists the number of times the heuristics identified the name as being singular male, singular female, singular neuter, or plural. Note that this dataset was extracted from English web text in 2006; repeating the procedure described by Bergsma and Lin (2006) for Dutch can therefore be expected to yield further improvements.

For nouns we use lexical information from Cornetto (Vossen et al., 2013) to distinguish human and non-human mentions, and to infer gender, where applicable. Many nouns in this dataset have multiple senses with conflicting information. As a simple workaround, we manually select the most common sense for these nouns; e.g., the Dutch word *apparaat* is either a physical device (non-human), or a bureaucratic entity (potentially human, much rarer). Implementing and incorporating a proper animacy and gender classifier, or even word-sense disambiguation component, is left for future work.

4.3 Quote attribution

In the CoNLL 2012 shared task a speaker attribute is provided as part of the data, linking utterances to speakers. In a realistic setting with unannotated text, identifying speakers must be done as part of the resolution process. We therefore mark quoted speech and attribute it to its speaker mention where possible.

Direct speech is identified using punctuation: quotation marks or sentences starting with a dash are marked as quoted speech spans. This heuristic is correct in the vast majority of cases; however, there are cases where quotation marks are used for something other than direct speech; these should be corrected with more elaborate rules or a classifier.

Speakers are detected where explicitly mentioned, and this information is extrapolated assuming turn-taking of alternating interlocutors. Explicitly mentioned speakers are identified when the subject of a reported speech verb (says, replies, etc.) is found next to a quotation.[8] The addressee is set to the speaker of the previous or following quotation, if distinct. Paragraph breaks are used to distinguish whether the same speaker continues speaking, or whether another participant takes a turn. By assuming that the same pair of speakers keeps taking turns, the speakers and addressee can be attributed in longer chains of dialogue even when the speakers are not identified explicitly in the text. These heuristic rules are highly similar to those reported by Muzny et al. (2017b), although they do not discuss attributing addressees. Another important difference is that we do not assign unattributed quotes to the majority speaker.

[7] Meertens instituut KNAW, 2010. *Nederlandse Voornamenbank* (Dutch first name database). `https://www.meertens.knaw.nl/nvb/`

[8] The list of reported speech verbs has been mined from a large corpus of parsed novels using a syntactic query of the form "NP verb quoted-speech" in various orders.

4.4 Coreference sieves

Coreference is resolved by applying a sequence of rule-based, deterministic sieves, as presented by Lee et al. (2011, 2013). The name *sieve* refers to the fact that at each step, rules are applied to decide on particular types of coreference, while passing on the rest of the coreference decisions to the next sieve. The sieves are designed to have high precision, and ordered from most to least precise. The rules operate on parse trees and match particular configurations of constituents and available features. In contrast to the commonly used mention-pair model in which pairs of individual mentions are linked based on their features, this model is entity-centric. This means that as mentions are linked into entities, their features are merged and this extra information provides further information for later decisions. The rest of this section introduces the sequence of rules (sieves) that make coreference decisions (for more details, cf. Lee et al., 2013).

String match Non-pronominal mentions with the same surface form are linked by this sieve.

Indexical expressions (his mother, today's special) should be excluded, but are not yet detected.

Precise constructs Intra-sentential coreference links that can be inferred from the parse tree are handled by these rules. This comprises relative, reflexive, and reciprocal pronouns, as well as appositives, predicate nominals, and acronyms and their expansions. Acronyms are detected by mapping noun phrases to candidate acronyms (with and without stop words), and linking the resulting acronym when it is encountered in the sentence:

(17) [The International Standards Organization]$_1$ warns that the space of [Three-Letter Acronyms]$_2$ ([TLA]$_2$) will run out soon ([ISO]$_1$ report).

This sieve is strongly affected by parse tree quality. Errors that trace back to this sieve are typically hard to work around without fixing the parse tree.

Head match This sieves links nominal mentions with matching heads and modifiers. The modifiers of the second mention must be a subset of the first mention. For example, *Yale University* is linked with *the university* ($\{\} \subseteq \{\text{Yale}\}$), but not *Harvard University* ($\text{Harvard} \not\in \{\text{Yale}\}$).

Proper head noun match This sieve links different variations of names; e.g. John Smith, Mr Smith, John. This sieve is applied globally to the whole document, while other sieves are restricted to linking mentions with antecedents preceding them.

Pronoun resolution This sieve links each pronoun to an antecedent with compatible features. Candidate antecedents are ranked by recency and syntactic prominence (Hobbs, 1978); the highest ranked candidate is selected as antecedent. After applying restrictions based on feature compatibility and binding constraints, the closest compatible anaphor is selected. Within each sentence, mentions are ranked by prominence using grammatical functions (in order of prominence: subject, direct object, indirect object, prepositional phrase). Pronouns in quoted speech are treated separately: first and second person pronouns are linked with the detected speaker or addressee while other pronouns are blocked from referring to the speaker. The current implementation does not support cataphora and has no heuristics to exploit frequency or parallelism; these should be added in future work.

An example where binding constraints are needed:

(18) *[Jan]$_1$ vond [[Bens]$_2$ portret van [zichzelf]$_2$]$_3$ mooi. [Hij]$_1$ gaf [hem]$_2$ [een compliment].*
[John]$_1$ liked [[Ben's]$_2$ portrait of [himself]$_2$]$_3$. [He]$_1$ gave [him]$_2$ [a compliment].

The pronoun *Hij* cannot be coreferent with *hem*, since they are co-arguments in the same clause (Lappin and Leass, 1994). Conversely, the reflexive *zichzelf* must refer to its co-argument *Bens*.

Binding constraints also prohibit *i*-within-*i* constructions, specifying that two mentions cannot be coreferent if the span of one is a subset of the span of the other. However, possessive pronouns form an exception to this principle; for example:

(19) *[de raket met [haar]$_1$ massa van 750 ton]$_1$*
[the rocket with [its]$_1$ mass of 750 tons]$_1$

Note the possessive female pronoun *haar*, which is translated as a neuter pronoun in English.

Throughout the implementation of the system, care is taken to minimize asymptotic complexity and inefficiency, to ensure that the code scales to book-length documents. While the implementation was written from scratch, the code of two previous implementations of Dutch coreference systems has been consulted during development: GroRef (van der Goot et al., 2015) and a prototype by Antske Fokkens introduced in Fokkens et al. (2018).[9]

5. Evaluation

Coreference resolution is hard to evaluate because the task involves several levels (mentions, links, entities), and results can be partially correct on each level. Metrics can take mentions, links, or entities as the unit of comparison. Parsing errors and mention detection issues have a large effect on the rest of coreference resolution. Moreover, it is a challenge to come up with a reasonable assignment of importance to different mistakes. A range of metrics have been presented, but all have been shown to exhibit fatal flaws (Moosavi and Strube, 2016):

MUC A link-based metric (Vilain et al., 1995); known to lack discriminative power and biased towards larger entities; awards 100% recall if all mentions are singletons.

B^3 A mention-based metric (Bagga and Baldwin, 1998); awards 100% precision to systems that leave all mentions as singletons; awards 100% recall if all mentions are merged into a single entity.

CEAF An entity (CEAFe) or mention-based (CEAFm) metric (Luo, 2005); scores coreference after aligning system and reference items; an issue is that misaligned items are penalized.

BLANC A link-based metric based on the Rand clustering metric (Recasens and Hovy, 2011); the BLANC score is the arithmetic mean of scores for coreference and non-coreference links, leading to higher scores when the number of gold mentions is increased, regardless of system performance.

In addition, the influence of mention detection performance is large and not intuitive. The CoNLL 2011 shared task (Pradhan et al., 2011) established the compromise of averaging the MUC, B^3, and CEAFe metrics into a score now referred to as the CoNLL score. This cancels out some of the disadvantages, but only increases the difficulty of interpreting scores.

For comparison with earlier work, we do report results with these flawed metrics. For our own dataset, we report results with the Link-based Entity Aware (LEA) metric presented by Moosavi and Strube (2016) to address the aforementioned problems. The LEA metric scores coreference based on links between pairs of mentions and assigns higher weight to mistakes with larger entities. All results include singleton mentions as part of the evaluation.[10] We use the reference scorer implementation (Pradhan et al., 2014) as well as the scorer by Moosavi and Strube (2016).

5.1 Shared task data

There have been two shared tasks on coreference resolution in which Dutch was evaluated: SemEval 2010 (Recasens et al., 2010) and CLIN 26 (2015).[11] Table 1 presents an evaluation on the data of

[9] Cf. `https://bitbucket.org/robvanderg/groref/` and `https://github.com/antske/coref_draft`

[10] While the LEA metric is clearly the best available coreference metric, we did run into counter-intuitive behavior with the LEA metric. When singleton mentions are included in the evaluation, adding a link to the system output may lower recall, while one would expect that the addition of a link may lower precision but never lower recall. This is because singleton mentions are evaluated using artificial self-links; adding a link to a singleton mention replaces the self-link which will cause a recall error for that self-link if the mention is a singleton in the gold standard file (in addition to the precision error for the added link).

[11] `http://wordpress.let.vupr.nl/clin26/shared-task/`

CLIN26 shared task	**Mentions**	**BLANC**
GroRef, Boeing test set	59.3	31.0
`dutchcoref`, Boeing test set	**59.5**	**31.5**
GroRef, GM test set	**60.4**	**31.3**
`dutchcoref`, GM test set	59.3	31.1
GroRef, Stock test set	53.7	25.4
`dutchcoref`, Stock test set	**54.7**	**26.1**
GroRef, dev. set	60.7	31.5
`dutchcoref`, dev. set	**62.2**	**33.5**

SemEval 2010, Dutch, test set	**Mentions**	**BLANC**	**MUC**	B^3	**CEAFm**
SemEval 2010: Sucre	42.3	**46.9**	29.7	11.7	15.9
SemEval 2010: UBIU	34.7	32.3	8.3	17.0	17.0
`dutchcoref`	**64.3**	41.5	**52.0**	**45.9**	**51.2**
SemEval 2010: Sucre, gold mentions	100	65.3	69.8	67.0	58.8
`dutchcoref`, gold mentions	100	**78.1**	**74.0**	**75.1**	**69.2**

Table 1: Evaluation on shared tasks.

these shared tasks. Our system (`dutchcoref`) improves on previously reported results on most of the metrics and datasets. The GroRef system from the CLIN 26 shared task is an implementation of the same rule-based sieve architecture as our system. We compare against the two systems reporting Dutch scores in the SemEval 2010 shared task: Sucre and UBIU. The Sucre system (Kobdani and Schütze, 2010) supports various machine learning classifiers and supports features for mention pairs but also individual mentions and words. The UBIU system (Zhekova and Kübler, 2010) uses memory-based learning with a mention-pair architecture.

We were able to reproduce the GroRef results with their code; we obtain the same results as reported in van der Goot et al. (2015) except for differences in rounding. Note that differences in the annotation schemes cause some additional errors. We evaluate the same way as the reported results of GroRef. Results of the files are micro averaged using the scripts of the shared task. For the development set, some annotations are missing and we therefore limit the coreference output to the first 6 sentences; the gold standard data is not modified, leading to recall errors for several files which do include 7 annotated sentences. The test set does not exhibit this issue and is therefore evaluated as intended.

For the SemEval evaluation we add a post-processing step to filter out singleton mentions (except names) and links from precise constructs since these are not annotated in the SemEval data. The other relevant differences between our evaluation and that of the SemEval shared task are as follows. Our results make use of external resources (Alpino parse trees, NER, and lexical resources for gender and animacy information). The only reported scores for Dutch in the shared task are for the closed track, meaning that they only use predicted parses, POS and NE tags in the provided files; these are of lower quality than the Alpino parses. Except for BLANC, the coreference scores for the SemEval systems are strikingly low—there might have been an evaluation issue in the shared task, or this could be due to the previous point; we report the scores as they are in Recasens et al. (2010) verbatim. The scores for SemEval include partial credit for mentions with a matching head but an incorrect span; we only count mentions with the right boundaries.

Manual Corea evaluation The Corea system is available as a web service.[12] We perform a manual evaluation on the first document of the SemEval 2010 development set. We manually convert

[12] `http://corea.tst-centrale.org/`

	Mentions			LEA			CoNLL
	recall	prec	F1	recall	prec	F1	
Corea web service	50.0	58.8	54.1	25.0	25.5	25.2	35.2
`dutchcoref`	70.0	87.5	77.8	55.0	75.0	63.5	71.0

Table 2: Manual evaluation of `file_7` of SemEval 2010 shared task data.

the XML output of the web service to the tabular format expected by the evaluation script.[13] The results are compared with our system, see Table 2. Our system performs significantly better. Caveats:

- The document consists of only 11 sentences (20 mentions and 10 entities in gold).
- The Corea system may have been trained on this data.
- The Corea web service is tuned for speed, not accuracy.
- Our system uses a better parser.

5.2 Novels

Coreference annotation We annotated a selection of popular, contemporary, Dutch-language novels (translated and originally Dutch) from the Riddle of Literary Quality[14] corpus for coreference. The novels are syntactically parsed with Alpino (i.e., the parse trees are not manually corrected). Coreference annotation is not done from scratch but proceeds by correcting the output of the automatic coreference resolution system presented in the present paper. Based on the output of this system, the annotations are manually corrected with the CorefAnnotator tool (Reiter, 2018).

Annotation proceeded in two phases. In the first phase, a selection of 10 novels was annotated by the author of the present paper. For each novel, a fragment consisting of the first 100 sentences was annotated for coreference. This set was used as a development set.[15] In the second phase, students annotated longer fragments from 11 different novels of 8,000 tokens each (rounded up to the nearest sentence) for coreference, and corrected each other's annotations in a second pass. In total, the annotated novel fragments consist of more than 107k tokens; see Table 3 for an overview of basic statistics of the corpus.

Results The performance of coreference resolution on the novel fragments with manually corrected coreference annotations is reported in Table 4. The scores for individual novels show considerable variance in performance. Cursory inspection shows that this variance cannot be explained by the number of mentions and entities in the fragments. However, the variance may well be explained by more complex sentence and discourse structure in more literary novels, which is an interesting topic for future research.

The difference in scores between the shared task data and the novels is large. This difference remains when gold mentions are used, although this does reduce the gap considerably. Both recall and precision are low for the shared task data. In order to see whether there is a marked difference in domain or annotation style, we consider basic statistics of the datasets, reported earlier in Table 3. The lower scores of the shared task data cannot be explained by a higher density of mentions or entities, since the novels exhibit the highest densities (0.095 vs 0.068 and 0.048 for the shared tasks), while the average sentence length is comparable (16–19 tokens per sentence). However, document length could be an important factor, since our novel fragments are much longer (100–491 sentences, compared to 7–21 for the shared tasks). The novels may well contain more frequent, repeated use of the same prominent entities (i.e., protagonists), which could be easier to resolve. Indeed, there is

[13] `https://github.com/andreasvc/dutchcoref/tree/master/data/manualeval`

[14] Cf. `http://literaryquality.huygens.knaw.nl/`

[15] The system presented in this work is unsupervised and does not use machine learning. The development set is therefore not needed to tune parameters, but was used during development of the system to test the effectiveness of rules empirically.

	CLIN26 dev set	**SemEval 2010 dev**	**Novels, dev set**	**Novels, test set**	**Springer, Quadriga**
documents	30	23	10	11	1
sentences	208	496	1,000	5,406	443
tokens	4,018	9,164	19,051	88,092	8,012
sents per doc	7	21.4	100	491.5	443
avg sent len	19.3	18.4	19.0	16.3	18.0
mentions	663	1,010	4,243	20,873	1,873
entities	273	424	1,798	8,337	698
mentions/tokens	0.17	0.11	0.22	0.24	0.23
mentions/entities	2.43	2.38	2.36	2.50	2.68
entities/tokens	0.068	0.046	0.094	0.095	0.087
% pronouns	7.69	14.45	43.3	36.5	36.5
% nominal	52.34	54.35	46.2	52.2	43.1
% names	39.97	31.20	10.5	11.2	20.4

Table 3: Statistics of the coreference datasets. Springer, Quadriga is part of the test set but reported separately because it is used in the error analysis of Section 6.

	Mentions		**LEA**		**CoNLL**
	F1	recall	precision	F1	
CLIN26 (Boeing test set)	59.5	29.8	34.0	31.8	41.2
SemEval 2010 (test set)	64.3	36.0	40.0	37.9	48.4
Novels (dev. set)	87.1	57.1	61.7	59.3	70.3
Novels (test set)	87.1	49.3	57.5	53.1	66.7
CLIN26 (Boeing test set, gold mentions)	100	50.7	62.5	56.0	69.0
SemEval 2010 (test set, gold mentions)	100	50.4	63.6	56.2	71.2
Novels (dev. set, gold mentions)	100	64.2	73.4	68.5	80.8
Novels (test set, gold mentions)	100	57.3	65.1	60.9	76.0

Table 4: Results on shared tasks and novels, with predicted and gold mentions.

	Mentions	**MUC**	**B^3**	**CEAFe**	**CoNLL**
Bamman et al. (2019a), English novels	89.1	84.3	62.7	57.3	68.1
`dutchcoref`, Dutch novels (dev. set)	87.1	74.9	67.6	67.6	70.3
`dutchcoref`, Dutch novels (test set)	87.1	71.4	62.1	66.7	66.7
Krug et al. (2015), German novels	100?	85.5	56.0		
Bamman et al. (2019a), English novels	100	88.5	72.6	76.7	79.3
`dutchcoref`, Dutch novels (dev. set)	100	81.0	79.2	82.1	80.8
`dutchcoref`, Dutch novels (test set)	100	77.5	72.1	78.4	76.0

Table 5: Comparison with other work on literary coreference with predicted and gold mentions.

a marked difference in the number of pronouns, with the novels having a much higher proportion of pronoun mentions (36-43% vs 7-14% for the shared tasks). To see whether the LEA evaluation metric is affected by document length (longer documents may score higher due to length alone), we evaluate the first 20 sentences of each of the novels. This does not make an appreciable difference in the scores, with or without gold mentions.

Table 5 compares our system with other coreference results on literature. Krug et al. (2015) evaluate a similar rule-based system on classic, German literary texts. Note that Krug et al. (2015) only consider coreference of person entities; it is not clear whether they perform mention detection or use gold mentions. Bamman et al. (2019a) evaluate an end-to-end neural coreference system with BERT embeddings on English novels published 1719–1922. These are the most relevant results to compare with, since the texts are from a similar domain. Although there are differences in annotation schemes and their novels are not contemporary, it is encouraging to see that for this domain, a rule-based model is still competitive with a state-of-the-art model trained on contextualized word embeddings.

5.3 Quote attribution

To get an estimate of the performance of the quote attribution component, we annotate the first 1,000 direct speech quotations from the Voskuil novel (this represents almost a fifth of the complete novel which has 5,239 direct speech spans). Each quotation is attributed to the name of one of the six speakers in this part of the novel. Before evaluation, each speaker is greedily mapped to a single coreference cluster based on matching names.

We obtain a low recall score of 43.3% since almost half of the quotations are not assigned a speaker, but when a speaker is assigned, it is correct 81.7% of the time (high precision). The low recall score is due to the fact that we opted not to assign unattributed quotes to the majority speaker since we want to favor precision. These results can be contrasted with the majority baseline accuracy score of 36% if every quotation is attributed to the most common speaker (“I,” 364 quotations).

The quote attribution rules work well when speakers are explicitly mentioned in the dialogue. Errors occur on harder cases where the speaker is implicit:

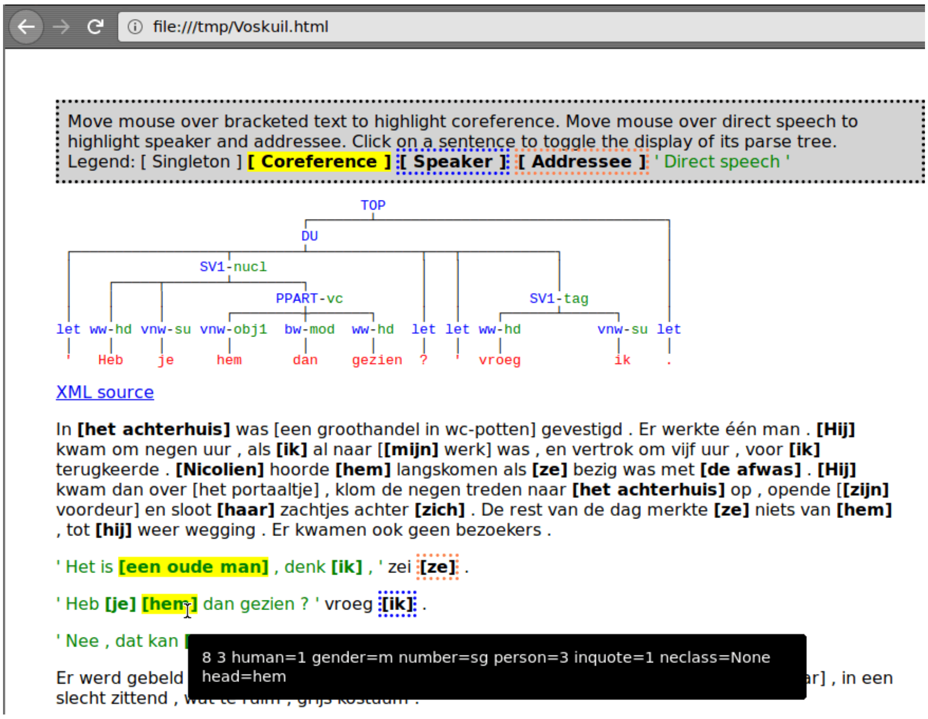
- Dialogues with more than two participants, where the turn taking heuristics fail. In practice, a dialogue turn may only be attributable to a speaker by inference from the discourse context and which participant is most likely to make a particular utterance.
- Multiple consecutive turns by the same speaker, which also causes turn taking heuristics to fail; again, this often relies on the context.

When looking more closely at the text, it becomes apparent that this novel is a difficult case due to an unusually large amount of dialogue. It turned out that during preprocessing of the text of this novel, mistakes have been introduced in paragraph marking, which is a crucial feature for the quote attribution rules. Moreover, chapter breaks should be used as a feature to block the heuristic of turn-taking, but were not preserved in the sentence identifiers or the cleaned text.

Our quote attribution results are lower than the results of Muzny et al. (2017b) on English, who apply almost the same heuristic rules. We have not trained classifiers with which they obtain even better results. Perhaps there is a difference in the style of dialogue reporting in the datasets (our novels are contemporary, theirs are 19th century novels). A preliminary experiment with training a fastText (Joulin et al., 2017) classifier on the text of the attributed quotes to classify the unattributed quotes did not yield encouraging results. A larger annotated dataset is needed.

6. Error analysis

Manual error analysis is greatly aided by visualization, since the tabular SemEval/CoNLL format is cumbersome to read directly. To this end, a web-based visualization was developed; see Figure 2.



file:///tmp/Voskuil.html
TOP
DU
SV1-nucl
PPART-vc
SV1-tag
let ww-hd vnw-su vnw-obj1 bw-mod ww-hd let let ww-hd vnw-su let
' Heb je hem dan gezien ? ' vroeg ik .

Move mouse over bracketed text to highlight coreference. Move mouse over direct speech to highlight speaker and addressee. Click on a sentence to toggle the display of its parse tree.
Legend: [Singleton] **[Coreference]** **[Speaker]** **[Addressee]** ' Direct speech '

XML source

In **[het achterhuis]** was [een groothandel in wc-potten] gevestigd . Er werkte één man . **[Hij]** kwam om negen uur , als **[ik]** al naar [**[mijn]** werk] was , en vertrok om vijf uur , voor **[ik]** terugkeerde . **[Nicolien]** hoorde **[hem]** langskomen als **[ze]** bezig was met **[de afwas]** . **[Hij]** kwam dan over [het portaaltje] , klom de negen treden naar **[het achterhuis]** op , opende [**[zijn]** voordeur] en sloot **[haar]** zachtjes achter **[zich]** . De rest van de dag merkte **[ze]** niets van **[hem]** , tot **[hij]** weer wegging . Er kwamen ook geen bezoekers .

' Het is **[een oude man]** , denk **[ik]** , ' zei **[ze]** .

' Heb **[je]** **[hem]** dan gezien ? ' vroeg **[ik]** .

' Nee , dat kan

8 3 human=1 gender=m number=sg person=3 inquote=1 neclass=None head=hem

Er werd gebeld ar] , in een slecht zittend ,

Figure 2: Visualization of coreference and quote attribution.

The tooltip shows features of the mention under the mouse cursor. Coreference of the respective entity for this mention is highlighted in yellow. The speaker and addressee of the quotation under the mouse cursor are highlighted with red and blue. Since the source of errors can often be traced back to the parse tree of a sentence, the latter can be inspected as well.

We considered several available tools for automatic error analysis of coreference systems (Kummerfeld and Klein, 2013; Gärtner et al., 2014; Martschat et al., 2015). However, a challenge is that such tools depend on the particular syntactic annotation scheme to classify errors (i.e., the Penn treebank annotation scheme, limiting these tools to English, Arabic, and Chinese). In addition, tools may require system specific output (e.g., mention features such as gender, anaphor-antecedent links), beyond the tabular SemEval/CoNLL output on which the evaluation metric is based. For future research, it would be good to develop coreference tools based on Universal Dependencies, with the goal of creating language independent tools.

After considering the three systems, we determined that the system by Kummerfeld and Klein (2013) was easiest to adapt to our Dutch coreference data.[16] The system takes gold standard coreference with parse trees as input, and compares it to the system output for the same sentences. Each error is categorized as being of one of several types, described in Table 6. Span errors are often parse errors, but may also be caused by incorrect mention detection rules. Extra mentions and conflated entities are precision errors, since they occur when the output has an extra mention or link. Conversely, missing mentions and divided entities are recall errors.

We will analyze the errors of the system on the first 8,000 tokens of the novel *Quadriga* by F. Springer. A breakdown of the errors classified by type is given in Table 6. The rest of this section discusses representative examples of each type.

[16] Our version is publicly available as a fork at `https://github.com/andreasvc/berkeley-coreference-analyser`

#	Error type	Description
35	Span error	A mention with the same head as a gold mention exists, but the boundary is incorrect.
31	Extra mention	A system mention not in the gold annotation.
147	Missing mention	A gold mention not in system output.
178	Conflated entities	Two separate entities in the gold data are linked in the system output.
203	Divided entity	A single entity in the gold data appears as two or more entities in the system output.

Table 6: Breakdown of automatically identified error types.

6.1 Span errors

Looking at the span errors, the first thing that jumps out is that 14 out of the 35 span errors are with German-language mentions. These German phrases are instances of code-switching: they are part of the Dutch-language novel as written by the original (Dutch) author. The POS tagger and parser are not trained to handle this. Examples (showing the system boundaries with brackets, gold boundaries underlined):

(20) a. *Eine [grosse Ehre für uns]*
A [big honour for us]
b. *die besten [Beziehungen im Staatsapparat]*
the best [jobs in government]
c. *[mit lebhaftem Interesse]*
with [lively interest]

These errors are a specific case of parse errors propagating downstream.

A common error for which a rule should be added is when adverbs are included in a noun phrase:

(21) a. *[ook Hamburg]*
[also Hamburg]
b. *[de brave Behrman erbij]*
[the good Behrman there as well]
c. *[steeds meer bijzaak]*
[increasingly an incidental matter]

6.2 Missing/extra mentions

Missing and extra mentions form a precision-recall trade-off (see Table 7). In the case of potentially non-referential pronouns we opted to favor precision by not including them as mentions, since we do not have a good way to detect them. Aside from non-referential pronouns, other pronouns and names are (relatively) easy to identify, and therefore the highest number of missing mentions is with nominal phrases, which have the most syntactically diversity. Many of the extra mentions contain time-related expressions or German text, while missing mentions are often due to parse errors.

6.3 Divided/merged entities

When the gold and system annotations are compared, a single cluster in one may correspond to multiple clusters in the other, and vice versa. Such errors are counted as divided and merged entity errors. For each divided or merged entity error, we can distinguish the incorrect part and the rest of the entity. Table 8 shows a breakdown of such errors, comparing them in terms of the composition of the different parts, listing the 7 most common combinations.

	extra	**missing**	**total**
name	13	42	55
nominal	12	73	85
pronoun	6	32	38
total	31	147	178

Table 7: Breakdown of missing/extra mentions.

Incorrect part			**Rest of entity**				
Na	No	Pr	Na	No	Pr	Conflated	Divided
-	1+	1+	1+	1+	1+	3	19
-	-	1+	1+	-	-	19	4
-	1+	-	1+	1+	1+	5	19
-	-	1+	-	1+	1+	20	6
-	1+	-	-	1+	-	28	9
-	-	1+	-	1+	-	37	11
-	-	1+	1+	1+	1+	7	85
Other						59	50
Total						178	203

Table 8: Counts of conflated and divided entities errors grouped by the Name / Nominal / Pronoun composition of the parts involved.

The most common type of conflated entity error is when a pronoun is incorrectly merged with an entity containing only a nominal (37 times). Conversely, the most common divided entity error is when a pronoun is incorrectly split from an entity containing names, nominals, and pronouns (85 times). These results confirm the finding by Kummerfeld and Klein (2013), who also report that pronoun link errors dominate the conflated/divided entity errors. Dutch has an additional challenge with the third person singular pronouns *hij, hem, zijn, haar* (he, him, his, her): their gender may refer to either biological or linguistic gender (Hoste, 2005, p. 168). Concretely, a gendered pronoun often refers to an animate referent with that gender, but may also refer to an object with that linguistic gender.

No errors with names in the incorrect part are in the table, which makes sense because names are relatively easy to link by surface forms (this does not hold for all genres; Dekker et al., 2019). When the incorrect part contains a nominal, the entities are most often divided when they should be together, except in cases where both the incorrect part end the rest of the entity consists solely of nominals. An example of a divided entity where a nominal is separated from a mixed entity:

(22) *[de belangrijkste gast aan boord]* vs *[ik], [mij]*, ...
[the most important guest on the plane] vs [I], [me], ...

Here a descriptive NP referring to the first-person narrator is missed. This particular link requires a high level of semantics/discourse understanding that is intrinsically hard. Most of the errors with this composition are such descriptive NPs, including metaphors:

(23) *[de gleufhoeden]* vs *[de Stasi]*, ...
[the fedora guys] vs [the Stasi], ...

An example of a conflated entity where both parts are nominals:

(24) *[zijn chef]* vs *[zijn hoogste chef]*
[his boss] vs [his highest boss]

Here the sieve linking mentions with matching heads has overlooked the significance of a modifier.

7. Discussion and Conclusion

We have a presented a system for coreference resolution of Dutch texts. Our evaluation on shared task data and novels shows that a simple rule-based system attains good performance while being efficient and easy to debug. Our results are competitive with a state-of-the-art deep learning system trained and evaluated on English literature (Bamman et al., 2019a). The evaluation and error analysis of Section 5 and 6 confirmed the following general challenges of coreference resolution, well known from previous research:

- Parse quality and mention detection has a large effect on later decisions. A large part of the performance of our system can be attributed to quality of parse trees from the Alpino parser.
- Pleonastic pronouns and pronouns with non-nominal referents are hard to identify.
- Coreference links which depend on world knowledge are hard.
- Commonly used coreference evaluation metrics are fatally flawed; subsequent research should give priority to the LEA metric.

In addition, we argue that coreference tools should use Universal Dependencies (Nivre et al., 2019) to strive for language independence where possible. The following challenges are particularly relevant for the literary domain:

- Dialogue: coreference resolution and quote attribution are intertwined; errors in the former propagate to the latter and vice versa.
- Long coreference chains, pervasive use of pronouns. The limitations of distance/salience-based heuristics for pronoun resolution are more apparent in longer texts.
- Novels set up a lot of implicit context when compared to short, self-contained news stories, which need to spell out their context more explicitly.

Despite these challenges, we observed the surprising result that the performance with novels is much better than the newswire from the shared tasks. However, this does not imply that coreference of literature is inherently easier. While we excluded document length as a factor, different annotation conventions may play a role. The most notable difference between the domains is that literature has a much higher number of pronouns. Additionally, we observed a large variance in performance across the novels; this warrants further investigation, since it suggests interesting stylistic differences on the discourse level.

In future work, we plan to explore two directions for improving the system to deal with these challenges. First, we will extend the rule-based system with statistical classifiers, as in the approach by Lee et al. (2017a). Such classifiers can be used to improve the detection of mentions and singletons, animacy and gender detection of mentions, quote attribution, and pronoun resolution. Second, we will train and evaluate a deep learning system; i.e., the end-to-end neural approach presented by Lee et al. (2017b)[17] and used by Bamman et al. (2019a). The advantage of deep learning methods over statistical machine learning approaches is that no feature engineering is required. While deep learning methods have proven to be extremely effective in benchmarks, the rule-based/statistical approach might still be more cost-effective for addressing the specific challenges of literary coreference, given the limited amount of annotated literary text available for training and the desirability of taking global information from the whole document into account.

[17] Note that the last two papers are written by two distinct people called Lee.

Acknowledgments: The author is grateful to the students for their annotation work, and to Stéphan Tulkens and four anonymous reviewers for their detailed comments and criticisms which substantially improved this paper.

References

Almeida, Mariana S. C., Miguel B. Almeida, and André F. T. Martins (2014). A joint model for quotation attribution and coreference resolution. In *Proceedings of EACL*, pages 39–48. http://aclweb.org/anthology/E14-1005.

Bagga, Amit and Breck Baldwin (1998). Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation*, pages 563–566. Granada.

Bamman, David (2017). Natural language processing for the long tail. In *Proceedings of Digital Humanities*. https://dh2017.adho.org/abstracts/408/408.pdf.

Bamman, David, Olivia Lewke, and Anya Mansoor (2019a). An annotated dataset of coreference in English literature. https://arxiv.org/abs/1912.01140.

Bamman, David, Sejal Popat, and Sheng Shen (2019b). An annotated dataset of literary entities. In *Proceedings of NAACL*, pages 2138–2144. http://aclweb.org/anthology/N19-1220.

Bamman, David, Ted Underwood, and Noah A. Smith (2014). A Bayesian mixed effects model of literary character. In *Proceedings of ACL*, pages 370–379. http://aclweb.org/anthology/P14-1035.

Bergsma, Shane and Dekang Lin (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of COLING-ACL*, pages 33–40. http://aclweb.org/anthology/P06-1005.

Boot, Peter (2014). Review of Distant Reading by Franco Moretti. *Digital Scholarship in the Humanities*, 30(1):152–154. https://doi.org/10.1093/llc/fqu010.

Bouma, Gosse, Walter Daelemans, Iris Hendrickx, Véronique Hoste, and Anne-Marie Mineur (2007). The COREA-project: manual for the annotation of coreference in Dutch texts. Technical report, University of Groningen.

Bouma, Gosse, Gertjan Van Noord, and Robert Malouf (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59. http://www.let.rug.nl/vannoord/papers/alpino.pdf.

Brooke, Julian, Adam Hammond, and Graeme Hirst (2017). Using models of lexical style to quantify free indirect discourse in modernist fiction. *Digital Scholarship in the Humanities*, 32(2):234–250. https://doi.org/10.1093/llc/fqv072.

Clark, Kevin and Christopher D. Manning (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of ACL*, pages 643–653. http://aclweb.org/anthology/P16-1061.

De Clercq, Orphée, Véronique Hoste, and Iris Hendrickx (2011). Cross-domain Dutch coreference resolution. In *Proceedings of RANLP*, pages 186–193. http://aclweb.org/anthology/R11-1026.

Dekker, Niels, Tobias Kuhn, and Marieke van Erp (2019). Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5(e189). https://doi.org/10.7717/peerj-cs.189.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840. http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

Donaldson, Bruce (2017). *Dutch: A comprehensive grammar.* Routledge. https://doi.org/10.4324/9781315620787.

Elson, David, Nicholas Dames, and Kathleen McKeown (2010). Extracting social networks from literary fiction. In *Proceedings of ACL*, pages 138–147. http://aclweb.org/anthology/P10-1015.

Elson, David, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors (2012). *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature.* Association for Computational Linguistics, Montréal, Canada. http://aclweb.org/anthology/W12-25.

Elson, David, Anna Kazantseva, and Stan Szpakowicz, editors (2013). *Proceedings of the Workshop on Computational Linguistics for Literature.* Association for Computational Linguistics, Atlanta, Georgia. http://aclweb.org/anthology/W13-14.

Elson, David K. and Kathleen R. McKeown (2010). Automatic attribution of quoted speech in literary narrative. In *Twenty-Fourth AAAI Conference on Artificial Intelligence.* https://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/viewFile/1945/2137.

Feldman, Anna, Anna Kazantseva, and Stan Szpakowicz, editors (2014). *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL).* Association for Computational Linguistics, Gothenburg, Sweden. http://aclweb.org/anthology/W14-09.

Feldman, Anna, Anna Kazantseva, Stan Szpakowicz, and Corina Koolen, editors (2015). *Proceedings of the Fourth Workshop on Computational Linguistics for Literature.* Association for Computational Linguistics, Denver, Colorado. http://aclweb.org/anthology/W15-07.

Fokkens, Antske, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter van Atteveldt (2018). Studying muslim stereotyping through microportrait extraction. In *Proceedings of LREC.* https://www.aclweb.org/anthology/L18-1590.

Gärtner, Markus, Anders Björkelund, Gregor Thiele, Wolfgang Seeker, and Jonas Kuhn (2014). Visualization, search, and error analysis for coreference annotations. In *Proceedings of ACL System Demonstrations*, pages 7–12. http://aclweb.org/anthology/P14-5002.

Grishman, Ralph and Beth Sundheim (1996). Message understanding conference 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.* http://aclweb.org/anthology/C96-1079.

Haeseryn, Walter, K. Romijn, G. Geerts, J. De Rooij, and M.C. Van den Toorn (1997). *Algemene Nederlandse Spraakkunst [General Dutch Grammar].* Martinus Nijhoff, Groningen. Electronic version: http://ans.ruhosting.nl/e-ans/index.html.

Hammond, Adam, Julian Brooke, and Graeme Hirst (2013). A tale of two cultures: Bringing literary analysis and computational linguistics together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8. http://aclweb.org/anthology/W13-1401.

Hendrickx, Iris, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri van der Vloet, and Jean-Luc Verschelde (2008a). A coreference corpus and resolution system for Dutch. In *Proceedings of LREC.* http://www.lrec-conf.org/proceedings/lrec2008/pdf/49_paper.pdf.

Hendrickx, Iris, Veronique Hoste, and Walter Daelemans (2008b). Semantic and syntactic features for Dutch coreference resolution. In Gelbukh, Alexander, editor, *Computational Linguistics and Intelligent Text Processing*, pages 351–361, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hirschman, Lynette and Nancy Chinchor (1998). Appendix F: MUC-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. `http://aclweb.org/anthology/M98-1029`.

Hobbs, Jerry R (1978). Resolving pronoun references. *Lingua*, 44(4):311–338. `https://doi.org/10.1016/0024-3841(78)90006-2`.

Hoste, Véronique (2005). *Optimization issues in machine learning of coreference resolution*. PhD thesis, 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 LREC*. `http://www.lrec-conf.org/proceedings/lrec2006/pdf/342_pdf.pdf`.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics. `http://aclweb.org/anthology/N06-2015`.

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (2019). BERT for coreference resolution: Baselines and analysis. In *Proceedings of EMNLP-IJCNLP*, pages 5807–5812. `http://aclweb.org/anthology/D19-1588`.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov (2017). Bag of tricks for efficient text classification. In *Proceedings of EACL*, pages 427–431. `http://aclweb.org/anthology/E17-2068`.

Kobdani, Hamidreza and Hinrich Schütze (2010). SUCRE: A modular system for coreference resolution. In *Proceedings of SemEval*, pages 92–95. `http://aclweb.org/anthology/S10-1018`.

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 Computationa Linguistics for Literature*, pages 98–104. `http://aclweb.org/anthology/W15-0711`.

Kummerfeld, Jonathan K. and Dan Klein (2013). Error-driven analysis of challenges in coreference resolution. In *Proceedings of EMNLP*, pages 265–277. `http://aclweb.org/anthology/D13-1027`.

Lappin, Shalom and Herbert J. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561. `http://aclweb.org/anthology/J94-4002`.

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. `http://aclweb.org/anthology/J13-4004`.

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 CoNLL*, pages 28–34. `http://aclweb.org/anthology/W11-1902`.

Lee, Heeyoung, Mihai Surdeanu, and Dan Jurafsky (2017a). A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, 23(5):733–762. https://doi.org/10.1017/S1351324917000109.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017b). End-to-end neural coreference resolution. In *Proceedings of EMNLP*, pages 188–197. http://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 NAACL*, pages 687–692. http://aclweb.org/anthology/N18-2108.

Luo, Xiaoqiang (2005). On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP*, pages 25–32. https://www.aclweb.org/anthology/H05-1004.

Martschat, Sebastian, Thierry Göckel, and Michael Strube (2015). Analyzing and visualizing coreference resolution errors. In *Proceedings of NAACL Demonstrations*, pages 6–10. http://aclweb.org/anthology/N15-3002.

Mitkov, Ruslan, Richard Evans, and Constantin Orasan (2002). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 168–186. Springer. https://doi.org/10.1007/3-540-45715-1_15.

Moosavi, Nafise Sadat, Leo Born, Massimo Poesio, and Michael Strube (2019). Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of ACL*, pages 4168–4178. http://aclweb.org/anthology/P19-1408.

Moosavi, Nafise Sadat and Michael Strube (2016). Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of ACL*, pages 632–642. http://aclweb.org/anthology/P16-1060.

Moretti, Franco (2013). *Distant Reading.* Verso Books.

Müller, Christoph (2006). Automatic detection of nonreferential *it* in spoken multi-party dialog. In *Proceedings of EACL*. http://aclweb.org/anthology/E06-1007.

Muzny, Grace, Mark Algee-Hewitt, and Dan Jurafsky (2017a). Dialogism in the novel: A computational model of the dialogic nature of narration and quotations. *Digital Scholarship in the Humanities*, 32(Supplement 2):ii31–ii52. https://doi.org/10.1093/llc/fqx031.

Muzny, Grace, Michael Fang, Angel Chang, and Dan Jurafsky (2017b). A two-stage sieve approach for quote attribution. In *Proceedings of EACL*, pages 460–470. http://aclweb.org/anthology/E17-1044.

Ng, Vincent (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of ACL*, pages 1396–1411. http://aclweb.org/anthology/P10-1142.

Ng, Vincent (2017). Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Thirty-First AAAI Conference on Artificial Intelligence.* https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14995/13999.

Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John

Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mițitelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van

Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu (2019). Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11234/1-2988`.

O'Keefe, Tim, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal (2012). A sequence labelling approach to quote attribution. In *Proceedings of EMNLP-CoNLL*, pages 790–799. `http://aclweb.org/anthology/D12-1072`.

O'Keefe, Tim, Kellie Webster, James R. Curran, and Irena Koprinska (2013). Examining the impact of coreference resolution on quote attribution. In *Proceedings of ALTA*, pages 43–52. `http://aclweb.org/anthology/U13-1007`.

Pareti, Silvia, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of EMNLP*, pages 989–999. `http://aclweb.org/anthology/D13-1101`.

Poesio, Massimo, Roland Stuckardt, and Yannick Versley (2016). *Anaphora resolution.* Springer. `https://doi.org/10.1007/978-3-662-47909-4`.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of ACL*, pages 30–35. `http://aclweb.org/anthology/P14-2006`.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong (2013). Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics. `http://aclweb.org/anthology/W13-3516`.

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*, pages 1–40. `http://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 CoNLL*, pages 1–27. `http://aclweb.org/anthology/W11-1901`.

Quine, Willard V. (1940). *Mathematical Logic.* Cambridge: Harvard University Press.

Recasens, Marta and Eduard Hovy (2011). BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510. `https://doi.org/10.1017/S135132491000029X`.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley (2010). SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of SemEval*, pages 1–8. `http://aclweb.org/anthology/S10-1001`.

Reiter, Nils (2018). CorefAnnotator — a new annotation tool for entity references. In *Abstracts of EADH: Data in the Digital Humanities.* `http://dx.doi.org/10.18419/opus-10144`.

Rösiger, Ina, Sarah Schulz, and Nils Reiter (2018). Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of LaTeCH-CLfL*, pages 129–138. http://aclweb.org/anthology/W18-4515.

Rybicki, Jan, Maciej Eder, and David L. Hoover (2016). Computational stylistics and text analysis. In Crompton, Constance, Richard J. Lane, and Ray Siemens, editors, *Doing Digital Humanities: practice, training, research*, pages 159–180. Routledge, New York.

Schoen, Anneleen, Chantal van Son, Marieke van Erp, and Hennie van Vliet (2014). NewsReader document-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. In *Proceedings of LREC*, pages 2471–2477. http://www.lrec-conf.org/proceedings/lrec2010/pdf/162_Paper.pdf.

Steinbach, Uli and Ines Rehbein (2019). Automatic alignment and annotation projection for literary texts. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 35–45. https://www.aclweb.org/anthology/W19-2505.

van Dalen-Oskam, Karina, Jesse de Does, Maarten Marx, Isaac Sijaranamual, Katrien Depuydt, Boukje Verheij, and Valentijn Geirnaert (2014). Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal*, 4:121–136. https://clinjournal.org/clinj/article/download/45/39.

van der Goot, Rob, Hessel Haagsma, and Dieke Oele (2015). GroRef: Rule-based coreference resolution for Dutch. In *CLIN26 shared task*. http://kyoto.let.vu.nl/clin26_presentations/paper73.pdf.

van Noord, Gertjan (2006). At last parsing is now operational. In Mertens, Piet, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. http://www.let.rug.nl/vannoord/papers/taln.pdf.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.* https://www.aclweb.org/anthology/M95-1005.

Vossen, Piek, Isa Maks, Roxane Segers, Hennie Van Der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In *Essential Speech and Language Technology for Dutch*, pages 165–184. Springer. https://link.springer.com/chapter/10.1007/978-3-642-30910-6_10.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, and Jason Weston (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of ACL*, pages 1416–1426. http://aclweb.org/anthology/P15-1137.

Zhekova, Desislava and Sandra Kübler (2010). UBIU: A language-independent system for coreference resolution. In *Proceedings of SemEval*, pages 96–99. http://aclweb.org/anthology/S10-1019.