Computational Linguistics in the Netherlands Journal 12 (2022) 235-251 Submitted 12/2022; Published 12/2022

OpenBoek: A Corpus of Literary Coreference and Entities with an Exploration of Historical Spelling Normalization

Andreas van Cranenburgh A.W.VAN.CRANENBURGH@RUG.NL
Gertjan van Noord G.J.M.VAN.NOORD@RUG.NL

University of Groningen, The Netherlands

Abstract

We present OpenBoek: a corpus of 103k tokens of classic Dutch novels with annotated coreference and entities. The corpus has several properties that are challenging for current coreference models: long documents (fragments of 10k+ words each), domain-specific literary phenomena, and 19th century Dutch spelling. Spelling normalization is added to the corpus as an additional annotation layer, using a data-driven rule-based spelling normalization tool. Normalizations are added using meta-annotation, such that evaluation can be performed with annotations on the original texts without losing token alignment. This tool enables the application of parsing and coreference systems originally developed for modern Dutch. We evaluate parsing and coreference systems on the OpenBoek dataset and find that spelling normalization gives a substantial increase in performance. The OpenBoek corpus is available under an open license at `https://andreasvc.github.io/openboek/`

1. Introduction

The literary domain presents unique challenges for NLP tasks such as coreference resolution (Rösiger et al., 2018). Addressing these challenges requires annotated data of sufficient quantity and quality for training and evaluating models. Recent work introduced coreference datasets for classic English novels (LitBank; Bamman et al., 2020) and contemporary Dutch novels (RiddleCoref; van Cranenburgh, 2019).[1] Unfortunately, RiddleCoref is encumbered by copyright; i.e., the annotated texts cannot be made available. We address this by annotating a corpus of Dutch public domain novels which we release under a Creative Commons Attribution license. During annotation of this corpus, the spelling of historical Dutch stood out as a major source of errors by NLP tools. We therefore developed tools to minimize such errors through spelling normalization.

There has been a lot of work on coreference resolution, although most of it has focused on the English OntoNotes benchmark dataset. Recently, there has been more attention for coreference in other domains and languages. A particularly interesting direction is the CorefUD effort (Nedoluzhko et al., 2022), which aims to collect and harmonize coreference annotations for different languages, similar to what the Universal Dependencies project (UD; Nivre et al., 2019) has achieved for syntactic annotations. However, in the remainder of this section we focus on the domain of narrative fiction and the Dutch language.

Several recent works have introduced literary coreference corpora: LitBank (Bamman et al., 2020) consists of classic English novels, FantasyCoref (Han et al., 2021) is a corpus of fairy tiles in English, and GerDraCor-Coref (Pagel and Reiter, 2020) is a corpus of German drama. Baruah et al. (2021) introduces a corpus of (English) screenplays annotated for 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

1. The 1M word SoNaR-1 corpus (Schuurman et al., 2010) contains 2000 tokens of coreference annotations for books, but the majority consists of Wikipedia and various other genres.

©2022 Andreas van Cranenburgh, Gertjan van Noord.

documents	9	mentions / entities	2.66
sentences	5,709	mentions / tokens	0.228
sents per doc	634.3	entities / tokens	0.0857
avg sent len	18.1	% pronouns	40.9
mentions	23,650	% nominals	48.0
entities	8,875	% names	11.1

Table 1: Corpus statistics.

Author, title	# tokens	# n	f	m	fm	sg	pl
Conan Doyle, De Agra Schat	10,002	792	7	34	60	712	181
Couperus, Eline Vere	10,004	746	42	32	55	638	237
Hugo, De Ellendigen	10,022	807	13	65	178	731	332
Multatuli, Max Havelaar	10,008	627	27	34	112	611	189
Nescio, De Uitvreter	14,300	1067	10	48	141	955	311
Nescio, Dichtertje	17,276	1296	55	48	151	1150	400
Nescio, Titaantjes	11,790	689	16	38	68	569	242
Tolstoy, Anna Karenina	10,101	606	21	34	74	553	182
Verne, Reis om de Wereld	10,019	738	4	33	157	680	252
Total	103,522	7368	195	366	996	6599	2326

Table 2: Corpus composition (# tokens, # entities neuter, female, male, female and/or male).

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 RiddleCoref corpus presented a Dutch corpus of literary fragments (van Cranenburgh, 2019). The dataset presented in this work uses the same dutchcoref annotation scheme, which is a variant of the Corea guidelines.

Coreference on long documents has been noted as being challenging. Recently work as started to address this challenge (Toshniwal et al., 2020; Thirukovalluru et al., 2021). These papers train and evaluate on LitBank (Bamman et al., 2020), which consists of documents of up to 2000 tokens. This paper presents an annotated dataset with longer documents.

2. Data

The books for the present corpus have been selected from Project Gutenberg, in order to create a dataset not encumbered by copyright. The OpenBoek corpus consists of 9 texts, both translated and original Dutch novels from the 19th and early 20th century. These texts have been selected to represent a variety of genres of classical works. Tables 1 and 2 provide an overview of the corpus.

The text is based on the plain text UTF-8 format available on Project Gutenberg. The texts were cleaned by removing front and back matter (including epigraphs). Markup (such as for emphasis or underlining) is removed.

Although we address spelling normalization extensively in Section 4, we applied several manual spelling changes before annotation proceeded. This choice was made because at the time the annotation was performed, our spelling normalization tool had not yet been developed, and these changes were found to be necessary to get useful output for the semi-automatic annotation procedure. Specifically, the texts are parsed by Alpino, and the parse trees are used as input for a coreference system. Every parse error leads to many additional errors in the coreference system, which increases the number of annotations that need to be corrected.

For the text by Multatuli, we manually changed the following spellings:

- all occurrences of *y* replaced by *ij*
- *koffi* → *koffie*

As a reviewer pointed out, the normalization of *y* actually results in overnormalization in a few cases, e.g., *Mijthologie* and *hijdraulish*.

For the texts by Nescio, idiosyncratic spellings of the pronoun "he" as *i* are normalized; for example *datti* (a contraction of 'that he') is normalized into *dat -ie*. This change is not only to accommodate Alpino which does not detect *i* as a pronoun, but also to ensure that coreference annotation can proceed on the level of tokens. Note that these spelling variants are idiosyncrasies of the author Nescio, and not representative of the spelling changes that our normalization tools aim to solve.

In the results presented below, the above changes were already included in the dataset that we refer to as "original."

We annotated the full text of the novellas by Nescio, while for the other texts, we annotated the first 10k tokens,[2] rounded upwards to the nearest sentence boundary. The documents are therefore considerably longer than those in other coreference corpora (for SoNaR-1 and LitBank, the mean number of tokens per document is 1000 and 2000, respectively). We opt for longer text fragments of at least 10k tokens with the aim of creating a benchmark dataset for evaluating and tackling the particular challenges of long-document coreference resolution.

3. Annotation

Annotation proceeded with the same semi-automatic method and annotation guidelines[3] as RiddleCoref: each text was parsed by Alpino (van Noord, 2006) and coreference output of the dutchcoref system (van Cranenburgh, 2019) was manually corrected by two annotators, with the second annotator correcting the first, as needed. We used the CorefAnnotator annotation tool (Reiter, 2018). The rest of this section describes the annotation scheme, as presented in van Cranenburgh (2019).

3.1 Mentions

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

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.

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.

2. This token count, and all others reported in this paper, is based on the output of the Alpino tokenizer.
3. https://github.com/andreasvc/dutchcoref/blob/master/annotationguidelines.pdf

3.2 Mention attributes

We annotated the gender (neuter, female, male, gendered but mixed/unknown) and number (singular, plural) of all entities in RiddleCoref and OpenBoek. The gender attribute also distinguishes person and non-person entities (neuter implies non-person, rest implies person). Although the syntactic gender, number, and named entity category available in Alpino parse trees is already informative, we annotated the semantic gender and number; e.g., the grammatically neuter *het meisje* is annotated as female and the singular *de groep* is annotated as plural due to being a collective noun. These annotations are useful for training models to detect mention features for coreference resolution, among other possible applications.

Table 2 shows that in most fragments, there are more male than female mentions, although there are exceptions. Incidentally, the exceptions are the more literary novels (Couperus, Nescio, Multatuli), while the fragments with the least female entities are also less literary, detective and adventure novels (Conan Doyle, Verne).

3.3 Coreference

We annotated unrestricted NP coreference; i.e., we do not select specific entity categories to annotate, but annotate the coreference between all pronominal and nominal entities mentioned in the text.

There were some interesting literary phenomena. For example, in Nescio, *Dichtertje* (Little poet), a distinction is made between two deities:

- *de God van Nederland* (The God of the Netherlands)
- *den echten God van hemel en aarde* (the real God of heaven and earth)

We apply the omniscient reader's point of view; i.e., even though at a point in the story the reader might not be aware that two mentions are part of the same entity since this is only revealed later, we annotate the mentions as the same entity regardless.

The following is an example sentence:

(1) Toen had [zij]$_1$ [Henri Van Raat]$_2$ ontmoet, en sedert verbaasde [zij]$_1$ [zich]$_1$ vaak, hoe [die goede lobbes]$_2$, zooals [zij]$_1$ [hem]$_2$ noemde, [die]$_2$ toch zoo weinig op [den held [[harer]$_1$ droomen]$_3$]$_4$ geleek, [zooveel sympathie]$_5$ in [haar]$_1$ verwekte, dat [zij]$_1$ dikwijls, plotseling, naar [[zijn]$_2$ bijzijn]$_6$ verlangen kon. (Couperus, *Eline Vere*)
Then [she]$_1$ had met [Henri Van Raat]$_2$, and ever since [she]$_1$ often wondered to [herself]$_1$, how [that big dotterel]$_2$, as [she]$_1$ called [him]$_2$, [who]$_2$ so seemed so unlike [the hero [of [her]$_1$ dreams]$_3$]$_4$, arose [so much sympathy]$_5$ in [her]$_1$, that [she]$_1$ often, suddenly, could long for [[his]$_2$ presence]$_6$.

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

(2) 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:

(3) [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:

(4) [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.

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).

4. Spelling normalization

Dutch spelling has undergone a number of reforms (Nunn, 2006) to the effect that linguistic tools developed for current Dutch texts will have trouble with texts from the nineteenth century, since many words, written in the old spelling, will not be recognized by tools for modern Dutch. For this reason, we have implemented a tool chain which essentially predicts what the modern spelling of a given word would be. Our goal is not to fully normalize all orthography to a uniform standard, but rather to minimize the number of errors in downstream NLP tools due to spellings that are unfamiliar to those tools. Words with common alternative spellings, e.g., *elektriciteit* vs *electriciteit*, can therefore be left unnormalized.

Previous work on dealing with divergent spelling in NLP tools can be divided into two categories: adapting NLP tools to the data, and adapting the data to the NLP tools. In this paper, we follow a variant of the latter approach. We augment the data with meta-annotation to ensure that the original source text remains easily accessible, without the need to adapt modern tools. The meta-annotation act as instructions for modern tools how to treat particular word forms.

There has been work on spelling normalization of historical English (Baron and Rayson, 2008; Schneider et al., 2015; Yang and Eisenstein, 2016), as well as work on social media text (van der Goot and van Noord, 2017a,b). There is also the related problem of correcting OCR errors, for which tools are available such as TICCL (Reynaert, 2009, 2011; Reynaert et al., 2012), which is a data-driven, language-independent system. Such tools attempt to recognize spelling errors where

such errors can be very diverse. The problem we attempt to solve here is more limited in scope: recognizing the systematic changes in spelling that were implemented during the last 100 to 200 years. For this reason, we expected to be able to adapt the spelling more easily and effectively by writing rule templates for the anticipated systematic spelling changes in our data.

A recent paper addresses coreference in historical Irish texts (Darling et al., 2022) by applying normalizations and retraining NLP tools. As far as we are aware, there has been no other work on the particular challenges of coreference in texts with non-standard orthography.

Since there is, to the best of our knowledge, no parallel data available of texts in both spellings (19th and 21st century),[4] we have not applied a machine learning method, but instead we developed a data-driven rule-based approach. We have generated a number of rules based on a small set of hand-written rule templates and the DBNL novels corpus (van Cranenburgh et al., 2022). This corpus consists of 1346 Dutch novels from DBNL (a digital archive of Dutch literature), comprising all originally Dutch novels published in the period 1800–2000 which are available in DBNL. The corpus contains almost 6 million sentences and more than 130 million tokens.

The normalization tool will not change the input sentences, but instead it will add meta-notation which instructs the linguistic tools how particular words or word sequences should be treated. Preserving the input sentences is important for evaluation purposes, and means we do not have to change the other annotation layers. The meta-notation is described as follows. Note that this meta-notation is supported by the Alpino parser (van Noord, 2006). We also implemented a tool (available from the Github link given at the end of this section) that can be used to convert the text with meta-notation into a text in modern spelling (without any meta-notation).

4.1 Meta annotation

The most frequently used meta-notation is the `@alt` keyword. The sequence `[ @alt AltWord Word ]` indicates that the word `Word` should be interpreted as if it were written as `AltWord`.

Examples:

```
" Meld het aan Darja Alexandrowna en doe [ @alt zoals zooals ] zij beveelt . "
Ach , het moest alles [ @alt zo zoo ] komen !
```

In some cases, the old spelling of a word may be ambiguous. For instance, the word *den* might be the old spelling of the determiner "de", but it could also be the noun "den" (pine). This can be indicated in the meta-notation for `@alt` by using a sequence of alternatives all preceded by the tilde. Examples:

```
Matjeff stak de handen in [ @alt ~de~den den ] zak
Achter [ @alt ~dezen~deze dezen ] verscheen de barbier met alle
    [ @alt benodigheden benoodigdheden ] voor zijn meester .
```

In some cases, expressions in 19th century Dutch leave out a word that is now obligatory. For instance, the complex preposition *te midden* followed by a genitive case-marked noun is now often expressed by including the preposition *van* and a neutral form of the noun. In order to indicate that the tools should work as if a particular word was present, the meta notation `@phantom` can be used. Examples:

```
Eensklaps riep Fantine , te midden [ @phantom van ] [ @alt deze dezer ] stilte :
[ @alt Vind Vin-je ] [ @phantom je ] 't goed ?
Maar onze pa wil niet [ @alt dat da'k ] [ @phantom ik ] er bijkom .
```

In some cases, single words were spelled as two separate words in older spelling. In such cases, the words can be combined by the `@mwu` meta-notation (short for multi-word unit). Examples:

4. Texts by Multatuli and Couperus have been published in modernized spelling, called *hertaling* in Dutch. However, our task has a few particular requirements: we require word aligned text, in which only the spelling of words is normalized, without changing the existing sentence structure.

```
Wat is er [ @mwu van daag ] met je ?
[ @mwu Als je blieft ] .
```

The `@mwu_alt` meta-notation indicates that a sequence of words should be treated as the given alternative single word. For instance, the old fashioned phrase `des daags` can be treated as if the word `overdag` was used instead:

```
Zwarte [ @alt dromen droomen ] pijnigden mij [ @mwu_alt overdag des daags ] en des
    nachts .
Witte schrok van [ @mwu_alt zichzelf zich eiges ] .
```

The modern Dutch word *haar* is ambiguous between the pronoun ("her") and the noun ("hair"). In the old spelling, the noun variant was sometimes spelled as *hair*. In order to indicate that this word should be treated as *haar*, but without introducing ambiguity, the `@postag` notation can be used. This indicates the part-of-speech of the word. Examples:

```
Zij durven me geen [ @postag noun(het,mass,sg) hair ] krenken ... .
Aylva's handen werkten voort , [ @postag complementizer zonderdat ] hij zag .
```

Note that in the case of "hair" the meta-notation does not include the modern spelling of the word, "haar." This would be straightforward to add (in Alpino), but so far we did not (there was only a single instance of this type of transformation in our current rule set anyway).

4.2 Semi-automatic construction of rules

Rules have been generated on the basis of a small set of rule templates, and on the basis of a large collection of texts. Rule templates include spelling changes such as "`sch` is replaced by `s`". Some of the more frequently applicable rule templates are (upper case letters act as variables here, the actual rules do not differentiate between lower case and upper case):

- `sch` → `s`
- `gch` → `ch`
- `y` → `ij`
- `ae` → `aa`
- `VVlijk` → `Vlijk` where $V \in \{a, e, o, u\}$
- `VVCig` → `VCig` where $V \in \{a, e, o, u\}, \text{consonant}(C)$
- `VVCeN` → `VCeN` where $\text{vowel}(V), \text{consonant}(C), N \in \{n, r\}$
- `zoo` → `zo`

A rule mapping a word to its alternative is constructed if the following conditions are met:

1. the word is unknown to Alpino
2. the word occurs more than a threshold number of times in the DBNL novels corpus (van Cranenburgh et al., 2022), a 130M token corpus of novels published 1800–2000; we use a threshold of 10
3. the application of the rule templates leads to a word that is known to Alpino.

The lexical lookup procedure of Alpino has access to a full form lexicon of some 200 thousand words and 350 thousand names. In addition, there is a simple rule component that recognizes named entities such as dates and times. Finally, there is a large set of heuristics which are used for words that are otherwise not recognized. For the current procedure, this set of heuristics is naturally not used.

For instance, the word `aerdsch` occurs 57 times in the DBNL novels corpus, is not known to Alpino, and the application of the rule templates leads to the form `aards` which *is* known to Alpino. The rule `aerdsch` → `aards` therefore is added to the list of rules. These rules can then be used to

change an occurrence of `aerdsch` to the sequence `[ @alt aerdsch aards ]`. This mechanism leads to a rule set of slightly more than 4000 rules. Some typical examples are:

```
aandeelen aandelen
aangekleede aangeklede
aangeloopen aangelopen
aangenaame aangename
aangenaamer aangenamer
aangestooten aangestoten
```

By manual inspection of one of the novels, *Eline Vere* by Louis Couperus, we collected an additional set of 432 rules, as well as a set of more complicated and more ad hoc transformations which we expressed as edit statements for the Unix utility `sed` (in total 76 `sed` statements). Some examples are:

```
s/ da'k / [ @alt dat da'k ] [ @phantom ik ] /g
s/ is-t -ie / [ @alt is is-t ] [ @alt hij -ie ] /g
s/ als-t -ie / [ @alt als als-t ] [ @alt hij -ie ] /g
s/ as-t -ie / [ @alt als as-t ] [ @alt hij -ie ] /g
s/ aa's tie / [ @alt als aa's ] [ @alt hij tie ] /g
s/ was-t -ie / [ @alt was is-t ] [ @alt hij -ie ] /g
s/\([| ]\)Dat 's /\1Dat [ @alt is 's ] /g
s/ vin 'k / [ @alt vind vin ] [ @alt ik 'k ] /g
```

4.3 Final result and manual correction

The corpus-based rules, manual rules, and manually developed `sed` statements are applied to the text of the novels of our interest. In Table 3, the first column indicates how many words were altered by this procedure. Some of the older spelling variants are not caught by our approach, and in addition some of the rules lead to ambiguous output (using `@alt` with several options prefixed by tilde). For this reason, it is beneficial to go over the result manually and adapt the texts further. Further manual adaptation has been done for three novels.

As illustration, consider the following examples of automatically corrected sentences:

```
Sherlock Holmes nam zijn [ @alt fles flesch ] van [ @alt ~de~den den ]
    schoorsteenmantel en zijn werktuig voor onderhuidsche inspuitingen uit zijn
    marokijnen foudraal .
Met [ @alt ~zijn~zijne zijne ] lange , witte , zenuwachtige vingers bracht hij de
    fijne naald in orde , en schoof de linkermouw van zijn overhemd omhoog .
```

After manual correction, alternatives are reduced to one, and an additional correction is added:

```
Sherlock Holmes nam zijn [ @alt fles flesch ] van [ @alt de den ]
    schoorsteenmantel en zijn werktuig voor [ @alt onderhuidse onderhuidsche ]
    inspuitingen uit zijn marokijnen foudraal .
Met [ @alt zijn zijne ] lange , witte , zenuwachtige vingers bracht hij de
    fijne naald in orde , en schoof de linkermouw van zijn overhemd omhoog .
```

As can be seen in Table 3, this leads to a slightly larger number of transformations. In the final column of that table, we provide the ratio of the number of transformations found by the automatic procedure. For these three novels, the automatic procedure finds 86.5% of the required corrections.

To get an idea of the performance of the automatic spelling normalization, we need to make a more detailed comparison. Table 4 includes a breakdown of differences between the automatic and manual normalizations, distinguishing the following types:

Text	Automatic	Manual	Ratio
Nescio, Titaantjes	461 (3.7%)	514 (4.4%)	89.7%
Multatuli, Max Havelaar	287 (2.7%)	378 (3.8%)	75.9%
Conan Doyle, De Agra Schat	466 (4.4%)	511 (5.1%)	91.2%
Total	1214 (3.8%)	1403 (4.4%)	86.5%

Table 3: Number of tokens with spelling corrections.

Metric	Nescio	Multatuli	Conan Doyle	*Total*
Insertions	53	90	73	216
Deletions	2	2	32	36
Changes	25	17	65	107
Specifications	170	70	117	357
Common	278	210	261	749
Precision (%)	94.3	93.7	79.6	88.6
Recall (%)	89.4	75.7	70.5	83.7
ERR (%)	79.6	83.8	62.6	72.8

Table 4: Evaluation of the automatic spelling normalizations with precision, recall, and the error reduction rate (ERR).

Insertions normalizations missed by the automatic method that were added in the manual correction

Deletions normalizations that were made by the automatic method but removed in the manual correction.

Changes words that are normalized in both versions, but which are normalized differently in the manual correction.

Specifications cases where the automatic normalization suggests multiple alternatives, and the manual correction picks the correct one.

Common normalizations on which the automatic and manual method agree.

Insertions can be viewed as recall errors of the system, while deletions and changes can be viewed as precision errors of the system. On further inspection, most cases of the latter are actually debatable annotation choices in the manual version of Conan Doyle, in which correct normalizations were inadvertently removed. However, a handful of cases represent genuine precision errors by the system; e.g., the names *Eeden* or *Hope* should not have been normalized. Both Specifications and Common can be interpreted as true positives of the system, since the ambiguous cases with multiple alternatives cannot be detected by a rule-based system.

Based on this classification, Table 4 presents an evaluation of the spelling normalizations. We report word-based precision and recall (Reynaert, 2008), as well as the Error Reduction Rate (van der Goot et al., 2021), which is the proportion of correct normalizations out of all words that should be normalized. This evaluation is meant to give an indication of the amount of overlap between the normalizations of the automatic system and the manual corrections. However, note that not every disagreement with the manual correction is necessarily a mistake, since multiple spellings may be acceptable for the downstream NLP systems.

Figure 1 shows an example of a sentence that is parsed incorrectly without spelling normalization, and correctly after automatic spelling normalization. In the following section, we report the effect of spelling normalization (both automatic and manual) on the quality of both the parsing system on the one hand, and the coreference system on the other hand.

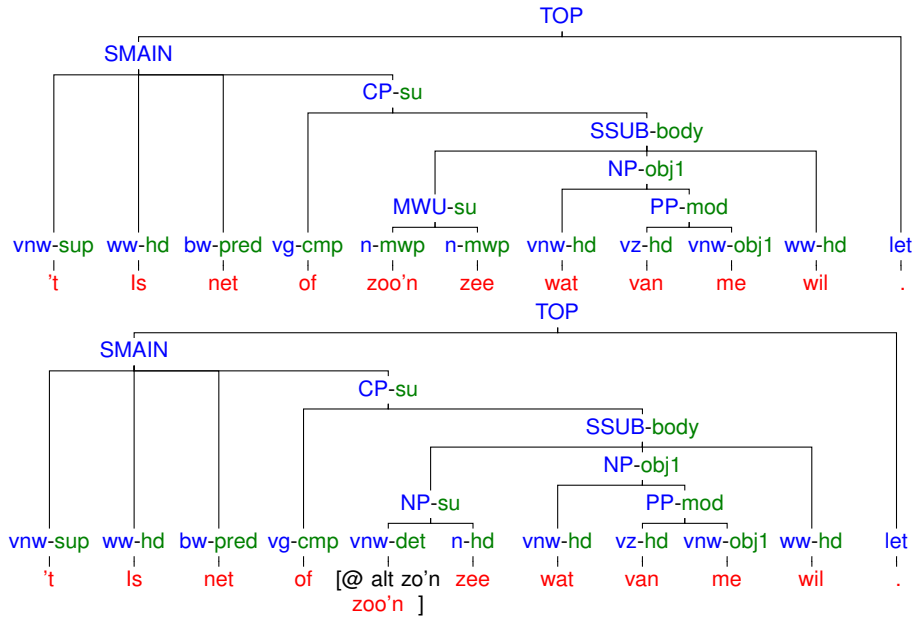

Figure 1: An example of a parse error which is avoided by using the automatic spelling normalization system (above: original spelling, below: normalized spelling). The sentence is from Nescio, Titaantjes.

The code of the spelling conversion is available for free and as open source at `https://github.com/gertjanvannoord/oudeboeken`. A web interface to the tool is available at `https://urd2.let.rug.nl/~vannoord/oudeboeken/`.

5. Evaluation of Parsing and Coreference Systems

We report coreference performance using three metrics. Mention performance reflects the performance of mention identification. The CoNLL score is the arithmetic mean of the F1 scores of three coreference metrics (MUC, B^3, and CEAFe), as used for the CoNLL-2012 task (Pradhan et al., 2012). Lastly, we report the LEA F1 score, which is a coreference metric that addresses issues discovered in previous coreference metrics (Moosavi and Strube, 2016), by giving more weight to larger (and thus more important) entities.

We evaluate the dutchcoref system[5] (van Cranenburgh, 2019) on the corpus. The dutchcoref system is an implementation of the Stanford deterministic multi-pass sieve coreference system (Lee et al., 2011, 2013), adapted for the Dutch language. It is a rule-based system that takes texts parsed by Alpino (van Noord, 2006) as input and applies rules to detect mentions and coreference.

5.1 The effect of spelling normalization

In order to see the effect of spelling normalization, we evaluate on three versions of the texts:

1. the original text
2. an automatically normalized version
3. a manually normalized version

5. Available at `https://github.com/andreasvc/dutchcoref/`

Text	Spelling	POS	DEP F1	Mention F1	CoNLL
Titaantjes	original	95.39	87.15	86.68	66.91
	automatic	**96.81**	90.67	88.82	67.83
	manual	96.75	**90.89**	**88.83**	**68.18**
Max Havelaar	original	95.11		84.81	64.92
	automatic	**96.46**		86.10	65.93
	manual	96.17		**86.43**	**66.18**
De Agra Schat	original	94.64		84.05	57.00
	automatic	**96.95**		86.63	59.30
	manual	96.93		**87.38**	**59.87**

Table 5: The effect of spelling normalization.

Although the dutchcoref system has optional neural modules, in this subsection we only evaluate the rule-based coreference system, since the manually corrected texts overlap with the training set of the neural modules. Besides evaluating the coreference performance, we also consider Part-of-Speech (POS) tag accuracy of the CGN coarse tags produced by Alpino. We have manually corrected the POS tags on the evaluation set using brat (Stenetorp et al., 2012). This correction was performed based on the POS tags from the automatically normalized text. The *Titaantjes* novella has also been manually annotated with Alpino dependency annotation. For this text we can therefore also compare how well the Alpino parser performs in terms of F1 score on labeled dependencies.

The correction of automatically normalized spelling turned out to be a much less labor intensive process than the correction of POS tags. The spelling of the Conan Doyle text was corrected in about five hours, while the POS tag correction took 14 hours. Given the improvement in parsing and coreference performance from spelling normalization, spelling corrections therefore seem to have a better return on annotation investment.

See Table 5 for the results. For all scores, we see that the automatically corrected version substantially improves on the results with the original version. For coreference, there is an additional but smaller improvement with the manually corrected spelling; this improvement appears to correlate with the spelling normalization scores reported in Table 4, i.e., most spelling corrections are already found by the automatic method.

Most of the coreference performance improvement can be attributed to better mention identification after spelling normalization, since we see these scores improve considerably and it is well known that mention identification is an important factor in coreference performance. Examples of incorrectly detected mentions in the original spelling are *rooken, hooren, den*; these are no longer detected after spelling normalization.

The results for dependency parsing on Titaantjes are very similar to the coreference results: a substantial improvement on the automatically normalized text, with a further small improvement if the parser proceeds on the basis of the manually corrected spelling.

Surprisingly, the POS performance is slightly but consistently lower after the additional manual spelling correction. However, inspection of the errors shows that this is in part due how split and merged words are evaluated. Other cases can be attributed to POS annotation mistakes. The gold POS annotations were based on the output of the automatically normalized spelling; therefore, if the annotator overlooks an incorrect POS tag, this will lead to a slight bias for the automatic normalization over the manual normalization. Examples of each of these error types:

split words When a word AB is split into A and B in the manual annotation using `[ @phantom A ] [ @alt B AB ]` the predicted POS tag for A is assigned as the POS tag for AB:

```
auto:     met petten op inplaats van helmen
gold POS:               VZ
pred POS:               VZ
manual:   met petten op [ @phantom in ] [ @alt plaats inplaats ] van helmen
pred POS:                                                 N
```

merged words When a word is merged with `[ @mwu_alt AB A B ]` the single predicted POS tag for AB is assigned to both of the original tokens A and B:

```
auto:     ... wie dat toe juichen gemeend had
gold POS:             VZ  WW
pred POS:             VZ  WW
manual:   ... wie dat [ @mwu_alt toejuichen toe juichen ] gemeend had
pred POS:                                    WW  WW
```

POS annotation mistakes When an incorrect POS tag based on automatically normalized spelling is not corrected, and a manual spelling correction triggers the correct POS tag to be predicted, this POS tag is counted as an error:

```
auto:     ... een ouden schoolkameraad
gold POS:     TW
pred POS:     TW
manual:   ... een [ @alt oude ouden ] schoolkameraad
pred POS:     LID
```

5.2 Rule-based versus neural coreference

The rule-based dutchcoref system was later extended with neural modules (van Cranenburgh et al., 2021). The neural modules take care of mention identification, predicting mention attributes (gender, animacy, number), and pronoun resolution; we refer to these modules as span, attr, and pron, respectively. These modules consist of neural classifiers that replace the rule based modules for the respective subtasks. The neural classifiers use a combination of handpicked features and BERT token embeddings. We use the monolingual Dutch BERTje model (Vries et al., 2019) to create the embeddings.

We use *Max Havelaar* as test set, in the original spelling (this is the spelling used during annotation, but as described in Section 2, a few manual changes have already been made to this text), as well as the automatically and manually normalized versions. The neural modules have been trained on the other texts in the corpus (in the original spelling),[6] as well as the contemporary novel fragments in the RiddleCoref corpus (van Cranenburgh, 2019).[7]

Table 6 presents the evaluation comparing the purely rule-based system with its neural extensions. We show the cumulative effect of adding the three modules. Similar to previous results on contemporary novels (van Cranenburgh et al., 2021), the neural modules obtain a substantial boost in performance. The neural modules are probably adapting to the spelling differences to an extent, since they are trained on parse trees from the original texts. However, spelling normalization gives a small but consistent boost in performance.

Contrary to the scores reported in van Cranenburgh et al. (2021) for contemporary novels, the neural module for pronoun resolution gives a substantial boost in pronoun accuracy. The best CoNLL score of 67.60 can be compared with the result of 68.1 reported for the English LitBank dataset (Bamman et al., 2020) using an end-to-end neural system. However, it is lower than the

6. It is probably advantageous to train the neural modules on spelling normalized versions of the OpenBoek texts; however, we leave this experiment for future work.
7. The trained models and the version of dutchcoref used in these experiments are available at `https://github.com/andreasvc/dutchcoref/releases/tag/v0.2`

System	mention			LEA			CoNLL	pron
	recall	prec	F1	recall	prec	F1		acc
ORIGINAL								
dutchcoref	89.47	80.61	84.81	53.61	46.97	50.07	64.92	53.44
dutchcoref+span	89.38	83.57	86.38	52.72	48.08	50.29	65.67	56.11
dutchcoref+span,attr	89.38	83.57	86.38	53.78	48.38	50.94	66.50	59.92
dutchcoref+span,attr,pron	89.38	83.57	86.38	55.22	48.11	51.42	67.14	62.98
AUTOMATIC SPELLING NORMALIZATION								
dutchcoref	90.09	82.44	86.10	54.06	47.79	50.73	65.93	55.73
dutchcoref+span	90.23	84.02	87.01	53.12	47.71	50.27	65.93	56.87
dutchcoref+span,attr	90.23	84.02	87.01	54.02	47.70	50.66	66.63	59.54
dutchcoref+span,attr,pron	90.23	84.02	87.01	55.58	47.65	51.31	67.34	62.98
MANUAL SPELLING NORMALIZATION								
dutchcoref	90.23	82.94	86.43	54.22	47.92	50.87	66.18	56.11
dutchcoref+span	90.50	84.12	87.19	53.38	48.01	50.55	66.26	57.63
dutchcoref+span,attr	90.50	84.12	87.19	54.31	47.83	50.87	66.85	59.92
dutchcoref+span,attr,pron	90.50	84.12	87.19	55.97	47.87	51.61	67.60	63.36

Table 6: Coreference evaluation of Max Havelaar.

score for the same system on contemporary Dutch literature (RiddleCoref): 71.0 (van Cranenburgh et al., 2021). Besides spelling, this difference may also be attributable to document length and text complexity.

6. Conclusion and Future Work

We have presented OpenBoek: a Dutch corpus of long literary fragments annotated for coreference. The long fragments are challenging for current coreference systems and the corpus is therefore a useful benchmark for long document coreference. In future work, more manually corrected annotation layers can be added to the corpus, such as named entities, syntactic dependencies, and events.

We have shown that old spelling of Dutch has a substantial effect on NLP tools, but can be mitigated to a large extent with simple rule-based methods. More advanced methods should be evaluated in future work.

Acknowledgments

We are grateful to the Information Science students who helped with the annotation of the corpus. The spelling normalization tool was developed for parsing the DBNL novels corpus, as part of a researcher-in-residence project at the National Library (KB). We thank an anonymous reviewer for extensive comments.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of LREC*, pages 44–54.

Alistair Baron and Paul Rayson. 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK.

Sabyasachee Baruah, Sandeep Nallan Chakravarthula, and Shrikanth Narayanan. 2021. Annotation and evaluation of coreference resolution in screenplays. In *Findings of ACL-IJCNLP*, pages 2004–2010.

Gosse Bouma, 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.

Mark Darling, Marieke Meelen, and David Willis. 2022. Towards coreference resolution for early Irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 85–93, Marseille, France. European Language Resources Association.

Orphée De Clercq, Véronique Hoste, and Iris Hendrickx. 2011. Cross-domain Dutch coreference resolution. In *Proceedings of RANLP*, pages 186–193.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of CRAC*, pages 24–35.

Iris Hendrickx, 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*.

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

Véronique Hoste. 2005. *Optimization issues in machine learning of coreference resolution.* Ph.D. thesis, Universiteit Antwerpen. Faculteit Letteren en Wijsbegeerte.

Véronique Hoste and Guy De Pauw. 2006. KNACK-2002: A richly annotated corpus of Dutch written text. In *Proceedings of LREC*.

Heeyoung Lee, 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.

Heeyoung Lee, 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.

Nafise Sadat Moosavi 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.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Joakim Nivre, 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 Mititelu, 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, Adédayọ̀ 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.

Anneke Marijke Nunn. 2006. *Dutch orthography: A systematic investigation of the spelling of Dutch words*. Ph.D. thesis, Katholieke Universiteit Nijmegen.

Janis Pagel and Nils Reiter. 2020. GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of LREC*, pages 55–64.

Sameer Pradhan, 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.

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

Nils Reiter. 2018. CorefAnnotator — a new annotation tool for entity references. In *Abstracts of EADH: Data in the Digital Humanities*.

Martin Reynaert. 2008. All, and only, the errors: more complete and consistent spelling and OCR-error correction evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Martin Reynaert. 2009. Parallel identification of the spelling variants in corpora. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09, pages 77–84, New York, NY, USA. Association for Computing Machinery.

Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2):173–187.

Martin Reynaert, Iris Hendrickx, and Rita Marquilhas. 2012. Historical spelling normalization: A comparison of two statistical methods: TICCL and VARD2. In *Proceedings of ACRH-2*, pages 87–98.

Ina Rösiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of LaTeCH-CLfL*, pages 129–138.

Gerold Schneider, Hans Martin Lehmann, and Peter Schneider. 2015. Parsing early and late modern English corpora. *Digital Scholarship in the Humanities*, 30:423–439.

Ineke Schuurman, 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of EACL demonstrations*, pages 102–107.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. In *Findings of ACL-IJCNLP*, pages 3921–3931.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of EMNLP*, pages 8519–8526.

Andreas van Cranenburgh. 2019. A Dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

Andreas van Cranenburgh, Esther Ploeger, Frank van den Berg, and Remi Thüss. 2021. A hybrid rule-based and neural coreference resolution system with an evaluation on Dutch literature. In *Proceedings of CRAC*, pages 47–56.

Andreas van Cranenburgh, Sara Veldhoen, and Michel de Gruijter. 2022. Textual features and metadata for DBNL novels 1800-2000. Zenodo data set.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509.

Rob van der Goot and Gertjan van Noord. 2017a. MoNoise: Modeling noise using a modular normalization system. *Computational Linguistics in the Netherlands Journal*, pages 129–144.

Rob van der Goot and Gertjan van Noord. 2017b. Parser adaptation for social media by integrating normalization. In *Proceedings of ACL*.

Gertjan van Noord. 2006. At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. arXiv:1912.09582.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of NAACL*.