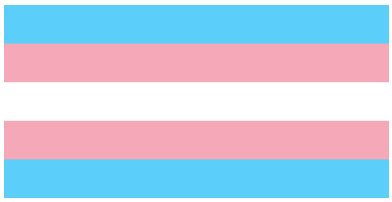
*Trans*forming Dutch: Debiasing Dutch Corefence Resolution Systems for Non-Binary Pronouns



Goya van Boven
6981844

Supervised by:
Yupei Du
Dr. Dong Nguyen
Prof. dr. Antal van den Bosch

A thesis submitted in fulfillment
of the requirements for the degree of
MSc. Artificial Intelligence
44 ECTS

Department of Information and Computing Sciences
Graduate School of Natural Sciences
College of Science

Utrecht University
Utrecht, the Netherlands
January 2024

Contents

Abstract

Gender-neutral pronouns are increasingly being introduced across Western languages, and are continuously more frequently being adopted by non-binary individuals. Recent evaluations have however demonstrated that English language models and coreference resolution systems are unable to correctly process gender-neutral pronouns (Cao and Daumé III, 2021; Baumler and Rudinger, 2022; Dev et al., 2021), which carries the risk of causing harmful consequences such as erasing and misgendering non-binary individuals (Dev et al., 2021). This thesis pioneers an examination of a Dutch coreference resolution sytem's performance on gender-neutral pronouns, specifically *hen* and *die*. In the Dutch context, additional challenges arise from the relative novelty of these pronouns, introduced in 2016, compared to the longstanding existence of singular *they* in English. To carry out this evaluation, a novel Dutch neural coreference model is published, and an innovative evaluation metric, a *pronoun score*, is introduced, which directly represents the percentage of correctly processed pronouns. The results reveal diminished performance on gender-neutral pronouns compared to gendered counterparts. In response to these challenges, this study compares, as a first of its kind, the usage of two debiasing techniques for coreference resolution systems in non-binary contexts: Counterfactual Data Augmentation (CDA) and delexicalisation (Lauscher et al., 2022). Although delexicalisation fails to yield improvement, CDA significantly diminishes the performance gap between gendered and gender-neutral pronouns. A noteworthy contribution is the demonstration that CDA remains effective in low-resource settings, in which a limited set of debiasing documents is applied. This efficacy extends to previously unseen neopronouns, which are currently infrequently used but may gain popularity in the future. This underscores the viability of effective debiasing with minimal resources and low computational costs.

1 Introduction

Recent literature has highlighted the presence of biases in a broad variety of machine learning systems. This work adopts the definition of *bias* provided by Friedman and Nissenbaum (1996), who state it as "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996). Notable examples of bias include the disproportionately high risk scores assigned to Black defendants in comparison to White defendants by a criminal risk assessment system (Angwin et al., 2016) and online advertisements that more frequently display ads implying that a person searched for on Google has a criminal record for people with Black-identifying names (Sweeney, 2013).

Biases can emerge at various stages of the design process of machine learning systems (Friedman and Nissenbaum, 1996), e.g. as a result of the overrepresentation of a certain demographic group in the training data, or due to the reinforcement of patterns present in the training data by the algorithm. It is recognised that no algorithm can be entirely free of bias (Mittelstadt et al., 2016); however, when systems structurally disadvantage already marginalised groups, it may result in the reinforcement of marginalisation and the resulting consequences can be harmful. The harms caused by biases can be classified into two distinct categories: *allocational harms* and *representational harms* (Crawford, 2017). Allocational harms describe the negative outcomes that arise from the discriminatory or unjust allocation of resources or opportunities, such as medical care, job opportunities or loans. Representational harms on the other hand refer to negative consequences that arise from the portrayal of certain groups of people in a system or dataset, including stereotypical depictions or underrepresentation of groups in the output of a system, which in turn can perpetuate stereotypes and discrimination in society.

In the field of Natural Language Processing (NLP), research on bias has largely been centered around gender bias, with a particular emphasis on occupational biases. For instance, Bolukbasi et al. (2016) find that word embeddings for occupation words structurally capture stereotypical gender associations, such as considering *nurse* to be more female and *surgeon* more male. In recent years, research on the topic of gender bias has become increasingly prevalent in NLP, with investigations into various areas such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; de Vassimon Manela et al., 2021), language modeling (Nangia et al., 2020; Nadeem et al., 2021; Gehman et al., 2020; Webster et al., 2020; Bender et al., 2021) and (contextualised) word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Basta et al., 2019; Zhao et al., 2019; Kaneko and Bollegala, 2021) among others.

However, the vast majority of the NLP studies consider gender as binary and immutable (Cao and Daumé III, 2021; Devinney et al., 2022), thereby excluding transgender and non-binary individuals from their evaluations. *Transgender* individuals do not identify with the gender they were assigned at birth, contrasting *cisgender* individuals, who do identify with their assigned gender. The term transgender includes both people with

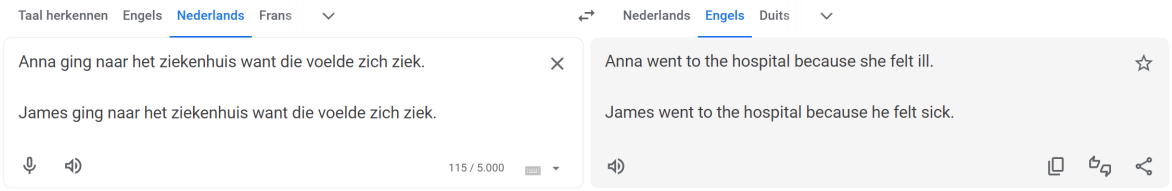


Figure 1.1: Example of misgendering by an NLP system : Google Translate translates the gender-neutral Dutch pronoun *die* as *she* and *he* in English. Screenshot 10-12-2023.

a binary transgender identity (such as transgender women) and people with a non-binary transgender identity. The term *non-binary* refers to individuals who do not conform to the traditional Western binary categorisation of male or female: they might identify as both female and male, as neither or their gender might fluctuate.

Within Western societies, transgender people face various forms of discrimination and marginalisation. They experience high rates of unemployment, homelessness, domestic violence, abuse and poverty, and frequently experience bullying and discrimination at the workplace (Terpstra et al., 2021; James et al., 2016). Furthermore, transgender people often encounter significant barriers in accessing essential institutions, such as healthcare services and the legal system (Zimman, 2018). In order to access healthcare, transgender individuals often need to navigate strict gatekeeping, forcing them to persuade figures of authority of the validity of their identity (Borba, 2017; Speer and Green, 2007). This problem is even more pronounced for non-binary individuals, who frequently struggle with not being seen as "trans enough" (Garrison, 2018): the transition from one binary gender identity to another is often considered as being more “authentic”, making it more likely to be accepted by gatekeepers (Konnelly, 2021).

Dev et al. (2021) point out how NLP models can contribute to the marginalisation of transgender people by perpetuating trans-exclusive practices. They particularly highlight the dangers of *erasing* and *misgendering* non-binary individuals. By erasure, they refer to (a) the stereotypical portrayal of non-binary individuals and (b) “invalidating or obscuring non-binary gender identities” (Dev et al., 2021). Erasure can occur within NLP, for example, when a system predicts a user’s gender but assumes a cisgender identity. Misgendering refers to addressing an individual with a gendered term that does not match their gender identity, which is often experienced as a harmful act (Ansara and Hegarty, 2014). An example of misgendering by an NLP system can be found in the Google Translate[1] translations in Figure 1.1. The Dutch sentences use the gender-neutral pronoun *die*, but the English translations use the gendered pronouns *she* and *he*, corresponding to the binary genders that the names *Anna* and *James* are most frequently associated with. This additionally is an example of erasure, because in the translated sentences, the non-binary identities of these individuals are obscured.

Gender-neutral pronouns are increasingly being introduced and popularised across Western languages, as more suitable alternatives to traditional gendered pronouns for non-binary individuals. In Swedish, the gender-neutral pronoun *hen* was politically introduced

[1] https://translate.google.com/

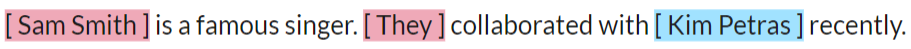


Figure 1.2: Coreference resolution task example, where mentions with the same colour refer to the same entity.

in 2013 (Gustafsson Sendén et al., 2015), the Dutch *hen/die* was democratically chosen by the Transgender community in 2016 (Transgender Netwerk Nederland, 2016) and while English has long known singular *they*, the set of *neopronouns* such as *ze* and *thon* is continuously growing (Lauscher et al., 2022). Given the relatively fast pace of these language changes, and the fact that NLP systems are typically trained on "large, binary-gendered corpora" (Dev et al., 2021), questions arise concerning the ability of existing language models to accurately process emerging pronouns, and if they prove inadequate, how they can be modified to enhance their inclusivity.

Several recent works have started to look into these questions. For instance, several works study the processing of gender-neutral pronouns in large language models (Watson et al., 2023; Martinková et al., 2023). Dev et al. (2021) compare BERT's (Devlin et al., 2019) ability to distinguish between (a) *he* and plural *they* and (b) singular and plural *they*. Brandl et al. (2022) assess the ability of English, Danish and Swedish language models to process gender-neutral pronouns by measuring perplexity as an indicator of processing difficulty. Lastly, Hossain et al. (2023) introduce a framework to comprehensively test the ability of language models to adopt an individual's declared preferred pronouns.

Moreover, in machine translation, recent studies have investigated the translations of traditional gender-neutral pronouns (Cho et al., 2019; Ghosh and Caliskan, 2023), neopronouns and newly introduced gender-neutral pronouns (Lauscher et al., 2023) between English and other languages. In Part of Speech (POS) tagging, Björklund and Devinney (2023) compare the performance of Swedish taggers on gendered and gender-neutral pronouns. Finally, in coreference resolution, Baumler and Rudinger (2022) systematically test whether systems can disambiguate between plural *they* and singular *they* (i.e. the gender-neutral pronoun), and Cao and Daumé III (2021) introduce an evaluation dataset of naturally occurring data pertaining to and authored by queer people, encompassing gender-neutral pronouns and neopronouns. Across these studies, consistently poor performances are observed for gender-neutral pronouns when compared to gendered pronouns.

The current project contributes to this body of work by evaluating and debiasing the ability of a Dutch coreference resolution system to process gender-neutral pronouns, zooming in on the pronouns *hen* and *die*. The task of coreference resolution entails identifying all expressions in a text that refer to the same entity, as Figure 1.2 illustrates. This is a fundamental NLP task, because it forms the basis of a wide range of applications, such as question answering, information extraction and summarisation (Ng, 2017). Therefore, any structural mistakes for non-binary individuals in coreference resolution systems, such as failing to recognise these pronouns – and thereby failing to extract information about these individuals – can lead to their erasure in a broad set of downstream applications.

While earlier studies have evaluated the performance of English coreference resolution

systems on gender-neutral pronouns (Baumler and Rudinger, 2022; Cao and Daumé III, 2021), this is the first study to perform such an evaluation for a Dutch system. Particularly, I train a new neural Dutch system, using the wl-coref architecture (Dobrovolskii, 2021). In order to precisely compare the model's performance on different pronouns, I introduce a *pronoun score*: a novel evaluation metric that quantifies the percentage of pronouns correctly resolved by the model.

The Dutch context differs from the English one because (a) Dutch gender-neutral pronouns are less frequent than English singular *they* and (b) there are generally fewer NLP resources available for Dutch than for English. The debiasing results might therefore be indicative examples for other Western languages that have coreference corpora of similar sizes ($\sim$ 1M tokens) available. Moreover, to my best knowledge, this is the first study to systematically compare methodologies for making coreference resolution systems more inclusive for non-binary individuals.

1.1 Research question

I ask the following research question: *Can the debiasing techniques Counterfactual Data Augmentation and delexicalisation improve the ability of Dutch coreference resolution systems to process gender-neutral pronouns?* In order to answer this question, I formulate five subquestions.

- *SQ1: How good is an existing Dutch coreference resolution system at processing gender-neutral pronouns compared to gendered pronouns?*
- *SQ2: Can the debiasing method Counterfactual Data Augmentation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns?*
- *SQ3: Can the debiasing method delexicalisation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns?*
- *SQ4: Can the debiasing method Counterfactual Data Augmentation improve system performance on previously unseen neopronouns?*
- *SQ5: Can the debiasing method delexicalisation improve system performance on previously unseen neopronouns?*

I now describe each of the subquestions in more detail.

SQ1: How good is an existing Dutch coreference resolution system at processing gender-neutral pronouns compared to gendered pronouns? In order to answer this question, I create four *pronoun-specific* datasets by transforming the Dutch coreference resolution corpus SoNaR-1 (Schuurman et al., 2010). This transformation involves replacing all third-person pronouns in the datasets with specific sets of pronouns, namely: 1. *hij/hem/zijn, 2. zij/haar/haar, 3. hen/hen/hun* and 4. *die/hen/diens* (subject/object/possessive). Subsequently, I evaluate the wl-coref model (Dobrovolskii, 2021), trained on Dutch data, on each of these *pronoun-specific* test sets. A description of the model and its adaptation to Dutch is provided in Chapter 4. Because Dutch gender-neutral pronouns are infrequent, and similar evaluations of English models have identified poor results on the processing of gender-neutral pronouns (Baumler and Rudinger, 2022; Cao and Daumé III, 2021), I

expected the model to perform worse on gender-neutral pronouns than on gendered pronouns. The results of the experiment validate this expectation, revealing notably lower performances for gender-neutral pronouns in comparison to their gendered counterparts.

SQ2: Can the debiasing method Counterfactual Data Augmentation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns? Counterfactual data augmentation (CDA), established as a useful debiasing method for mitigating gendered pronoun biases in coreference resolution (Zhao et al., 2018, 2019), involves the generation of modified instances within the training data, by altering specific features or labels to create hypothetical scenarios, including *gender swapping*: changing all female entities to male entities and vice versa (Zhao et al., 2018). I adapt this method to the non-binary context by creating a *gender-neutral* training set, wherein all third-person pronouns are replaced by gender-neutral pronouns (see Chapter 3). Debiasing is subsequently performed by training the system on this *gender-neutral* training set.

SQ3: Can the debiasing method delexicalisation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns? The only work to date that experiments with debiasing a coreference resolution system for gender-neutral pronouns is by Lauscher et al. (2022), who propose a method called *delexicalisation*. This method entails replacing *all* pronouns in the text with their part of speech tag, which they argue prevents the model from relying on gender-related lexical clues and instead learn a unified representation for all pronouns. I apply this debiasing approach by training the model on a *delexicalised* version of the data.

I experiment with delexicalisation and CDA in two settings: (i) fine-tuning the model from scratch on the respective debiasing dataset and (ii) further fine-tuning the original wl-coref model on the debiasing dataset. The effectiveness of both methods is evaluated by comparing the performance of the debiased systems on the *pronoun-specific* test sets for the two gender-neutral pronouns against the performance of the original model. Applying delexicalisation does not improve the performance on gender-neutral pronouns. Conversely, CDA demonstrates noteworthy debiasing results in both the fine-tuning from scratch and further fine-tuning settings. With the application of CDA, the performance gap between gendered and gender-neutral pronouns closes almost entirely. Importantly, a follow-up experiment in Section 5.3.3 shows that this method maintains effective in low-resource scenarios with just a handful of debiasing documents available.

SQ4 & SQ5: Can the debiasing methods Counterfactual Data Augmentation / delexicalisation improve system performance on previously unseen neopronouns? Considering the emergence and potential future introduction of new gender-neutral pronouns (Lauscher et al., 2022), it is crucial for any effective debiasing method to provide the model with the ability to handle previously unseen pronouns. For this reason, an additional test set is created and evaluated, encompassing Dutch neopronouns absent in the debiasing datasets. The findings reveal that neither debiasing technique succeeds in improving performance on neopronouns. However, in a supplementary debiasing experiment explicated in Section 5.4.3, it is observed that effective neopronoun debiasing can be achieved through the application of CDA with a limited number of debiasing documents containing neopronouns. This observation is encouraging as it signifies that, despite the

non-automatic future-proof nature of current debiasing techniques, modest interventions may suffice to ensure the accurate handling of novel pronouns.

I finally conclude that CDA proves to be an effective means of enhancing performance on gender-neutral pronouns. The noteworthy insight, that this method yields substantial improvements with only a small number of debiasing documents, stands out as a primary contribution of this study. This underscores the viability of debiasing in low-resource contexts with low computational costs. This finding opens up interesting directions for future explorations, such as applying this methodology across different languages and NLP tasks. Furthermore, by investigating debiasing within non-binary contexts, the present study adds to the advancement of inclusive AI systems.

This thesis is structured as follows. Chapter 2 provides a theoretical background for the current study. Subsequently, a description of the data (Chapter 3) and model (Chapter 4) are provided. The experiments and their results are reported in Chapter 5. Additionally, a test suite for evaluating pronoun-related model behaviour is presented in Section 5.5, wherein the models undergo testing on four core capabilities related to gender-neutral pronouns. Finally, I present my conclusions and discuss the results in Chapter 6.

The code used for this project can be found at `https://github.com/gvanboven/Transforming_Dutch`.

2 Theoretical background

This chapter is structured as follows. Section 2.1 gives a detailed account of gender and its manifestations in Dutch, Section 2.2 describes gender bias in NLP, and Section 2.3 elaborates on the coreference resolution task. The literature overview concludes with a discussion of gender bias in coreference resolution in Section 2.4.

2.1 Gender and language

In this section, I first provide an explanation of the concepts of *gender* and *misgendering* (Section 2.1.1), because an understanding of these concepts is essential for studying how language models handle gender-related phenomena, such as gender-neutral pronouns. Subsequently, I provide an account of the manifestation of gender in Dutch (Section 2.1.2), wherein gender-neutral nouns and pronouns are discussed.

2.1.1 Definitions

Gender

When discussing gender, it is important to make the distinction between *gender identity*, *gender expression* and *sex assigned at birth*. *Gender identity* refers to one's subjective experience of gender, i.e. being female, non-binary, agender, genderqueer or another gender identity (Stryker, 2017). *Gender expression* refers to how one expresses their gender through their physical appearance and mannerisms (Rajunov and Duane, 2019). *Sex assigned at birth* refers to the classification of a person as being female, intersex, male or another sex based on the combination of their genitals, anatomy, chromosomes, hormones, reproductive organs and secondary sex characteristics (Trans Student Educational Resources; Mey, 2014; Rajunov and Duane, 2019). For cisgender individuals, gender identity and sex assigned at birth align, while for transgender individuals they are distinct (Rajunov and Duane, 2019). Continuing, one's gender identity is not static but can change over time. The umbrella term *non-binary* encompasses all genders that are outside of the female-male binary (Rajunov and Duane, 2019). It is noteworthy to acknowledge that individuals identifying as non-binary are consistently categorised as transgender but that the reverse is not always the case, as transgender individuals may also identify with a binary gender identity.

Misgendering

Misgendering entails using gendered language to refer to someone in a way that does not correspond to their gender identity, which can be both intentional or accidental.

Examples of misgendering include using the wrong pronouns, nouns or honorifics, for instance by addressing a (trans)man as *she*, *Lady* or *Miss*. Even though it can occur to anyone, it is a common experience for transgender and gender-nonconforming individuals (Ansara and Hegarty, 2014). In a Dutch survey, 60% of the non-binary respondents indicates commonly being misgendered in the workplace (Terpstra et al., 2021). This act is harmful, as it can perpetuate the viewpoint that your gender identity is not perceived as real or valid by society (Keyes, 2018). Continuing, being misgendered correlates with rumination, emotional distress, self-doubt and internalised shame (Johnson et al., 2019), increased feelings of stigmatisation and devaluation, lower self-esteem (McLemore, 2015) and higher expectations of rejection (Rood et al., 2016).

2.1.2 Gender in Dutch

In this section, I discuss gender in the Dutch language. I first give an account of how gender is traditionally manifested in Dutch. Subsequently, I discuss Dutch gender-neutral nouns and gender-neutral pronouns.

In Dutch, gender distinction is present in third-person personal pronouns and nouns, but there is no gendered verb agreement or case inflection. The traditional usage of Dutch third-person singular pronouns distinguishes between feminine and masculine pronouns, without a gender-neutral alternative. In plural, a single group of pronouns is used for all genders. Table 2.1 below gives an overview of the third-person pronouns.

	Singular		**Plural**
Gender	Feminine	Masculine	All genders
Personal (subject)	*zij*	*hij*	*zij*
Personal (direct object)	*haar*	*hem*	*hen*
Possessive	*haar*	*zijn*	*hun*

Table 2.1: Overview of the traditional Dutch third-person pronouns

Nouns have a *grammatical gender* in Dutch, which can be feminine, masculine or neuter. Feminine and masculine nouns use the definite article *de* and nouns with a neuter gender use definite article *het*. Feminine nouns furthermore use feminine pronouns *zij/haar/haar*, while masculine nouns take masculine pronouns *hij/hem/zijn* and neuter nouns take *het/het/zijn*. In plural there is no gender distinction: all nouns use determiner *de* and pronouns *zij/hen/hun*.

We can further distinguish *referential gender*, the gender of the real-world entity that a linguistic expression refers to (Cao and Daumé III, 2021). In Dutch, referential and grammatical gender overlap in some cases, for example for the feminine word *dochter* (*daughter*). This is not always the case however. For instance the neuter word *het meisje* (*the girl*) is used for female referents. Additionally, some words with a masculine grammatical gender, like *de arts* (*the doctor*) and *de minister* (*the minister*) can be used for male, female and non-binary individuals. In such cases, the referential gender overrules the grammatical gender in terms of pronouns. For instance, one would say:

(1) De arts gaat naar *zijn* afspraak
(The doctor goes to *his* appointment)

if the doctor is male, and

(2) De arts gaat naar *haar* afspraak
(The doctor goes to *her* appointment)

when the doctor is female, despite the fact that *arts* (*doctor*) has a masculine grammatical gender.

Gender-neutral nouns

In Dutch, nouns referring to occupations and family members are usually gendered. For many occupations the masculine form is the root (e.g. *eigenaar* (*owner*), *schrijver* (*artist*), *student* (*student*)) and the feminine form is formed by adding a suffix (e.g. *eigenares*, *schrijfster*, *studente*) (Gerritsen, 2002). In other cases the male form is used for all genders, such as for *professor* (*professor*). For some occupational terms gender-neutral alternatives exist, e.g. replacing *lerares* (*female teacher*) and *leraar* (*male teacher*) with *leerkracht* (*teacher*). But, such gender-neutral forms do still not exist for all nouns.

Continuing, most words describing relatives only have a feminine and masculine version. For instance, no term such as the English *sibling* exist in Dutch, only providing the options *broer* (*brother*) and *zus* (*sister*). This is a problem for non-binary people, since there is no alternative that matches their gender identity.

Gender-neutral pronouns

In languages that traditionally have binary gendered pronouns, like Dutch, gendering others is close to unavoidable. This constitutes a problem for non-binary people, since neither of the binary options applies to them. For this reason additional pronouns that bypass the gender binary have been introduced, and are increasingly adopted across Western languages in recent years.[1] The English singular *they*, which has a long history of being used as a generic singular, has become a popular gender-neutral pronoun[2] (Conrod, 2019; Konnelly and Cowper, 2020), and was even voted Word of the Decade by the American Dialect Society (Roberts, 2020). An alternative to repurposing existing words is the use of *neopronouns*. Neopronouns are sets of pronouns that are newly introduced in a language (McGaughey, 2020), such as the Spivak pronouns *e/em/es* in English (Spivak, 1990). Another example of neopronouns is Swedish *hen*, which was politically introduced in 2013 (Gustafsson Sendén et al., 2015) and has quickly gained popularity since: it is now commonly used among the wider population (Gustafsson Sendén et al., 2021).

[1] See for instance `https://nonbinary.wiki/wiki/Pronouns` for an overview of gender-neutral pronouns in various languages

[2] `https://www.gendercensus.com/results/2021-worldwide-summary/`

In 2016, Transgender Netwerk Nederland organised a vote to determine what the Dutch gender-neutral pronoun should be, in which 500 community members participated. Here, *hen/hen/hun* was favoured over *die/die/diens* and neopronouns *dee/dem/dijr* (Transgender Netwerk Nederland, 2016). But, as of 2023, both *hen* and *die* are increasingly being adopted by non-binary people (EditieNL, 2021; Becker, 2020). Additionally, a broader set of neopronouns has been proposed, including *zhij* and *ij* (Hurkens, 2021; Het Neutrale Taal collectief), but these are not as widely used yet.

Traditionally, *die* is a demonstrative and relative pronoun, and *hen* is a third-person plural personal pronoun (i) for direct objects and (ii) succeeding pronouns. When used as gender-neutral pronouns, there is no difference in meaning between the two, and they can be used interchangeably. Contrasting the English singular *they* that remains conjugated as plural, both gender-neutral *hen* and *die* are conjugated as singular. An example usage of *hen* and *die* in Dutch is:

(3) Noa geeft *hun* studieboek weg omdat *hen* is afgestudeerd.
Noa geeft *diens* studieboek weg omdat *die* is afgestudeerd.
(Noa gives *their* study books away because *they* have graduated.)

The female version of this sentence would be:

(4) Noa geeft *haar* studieboek weg omdat *zij* is afgestudeerd.
(Noa gives *her* study books away because *she* has graduated.)

Despite the numerous reference guides on how to use these pronouns[3] (Schlief, 2021; Van Dale, 2021; Woelkens and de Vries, 2021), opponents claim gender-neutral *hen* is confusing (Haijtema, 2021), grammatically incorrect (Vogel, 2021) and feels unnatural because it is already used as a plural (Kamphuis and Akse, 2021; Becker, 2020; Geels, 2022). Others simply ignore its existence (Europees Parlement, 2018; Haverkamp, 2021). The introduction of gender-neutral pronouns commonly evokes negative responses (Gustafsson Sendén et al., 2015), but studies in the Swedish context show that these can quickly turn around to more positive attitudes (Gustafsson Sendén et al., 2015, 2021). Moreover, eye-tracking studies (Vergoossen et al., 2020a) debunk the prevalent argument among opponents (Speyer and Schleef, 2019; Vergoossen et al., 2020b), claiming that gender-neutral pronouns would lead to increased processing times.

Recently, Dutch official institutions have started to partially acknowledge the gender-neutral pronouns. In 2020 the online Van Dale dictionary[4] altered their definition of *hen* to include its gender-neutral usage, albeit with a *non-general* mark to indicate that the term is not used by the general public. Continuing, in 2022 TaalAdvies.net – a collaboration of four official language institutions – publised a reference guide on language and gender,[5] composed by a commission of 5 experts, of which one member is transgender

[3] https://denieuweliefde.com/genderneutraal-taalgebruik/, https://www.transgenderinfo.nl/wp/wp-content/uploads/2020/10/genderneutrale-voornaamwoorden-in-het-nederlands.pdf, https://www.langzaldieleven.nl/, https://weten.site/genderneutrale-voornaamwoorden/

[4] https://www.vandale.nl/

[5] https://taaladvies.net/taal-en-gender-algemeen/

and no members use gender-neutral pronouns themselves. They claim there is not yet a consensus about a gender-neutral pronoun in Dutch that can be used to refer to (i) non-binary people, (ii) people whose gender is unknown, and (iii) people in general; but they do list *hen* and *die* as potential candidates.

Certainly, the set of Dutch gender-neutral and neopronouns is anticipated to undergo further development in the coming years, encompassing an expanded range of options to effectively denote diverse identities. This evolutionary process is a recurring phenomenon observed across various languages (Brandl et al., 2022). The definitive designation of the most widely accepted gender-neutral Dutch pronoun remains an issue to be resolved over time. Nevertheless, the irrevocable integration of gender-neutral pronouns into the Dutch language is evident, as their adoption continues to grow.

2.2 Gender bias in NLP

NLP works that investigate gender bias typically only consider *binary* gender bias. For instance, in a literature review on gender bias, Sun et al. (2019) define gender bias as "the preference or prejudice toward one gender over the other (Moss-Racusin et al., 2012)", a definition that is strongly rooted in a binary conception of gender. Only more recently, studies have pointed out this cis-normativity in bias studies (Cao and Daumé III, 2020) and its accompanying harms, including misgendering and erasure (Dev et al., 2021).

In this section, I first briefly discuss the causes of gender bias in NLP systems in Section 2.2.1. Second, I discuss existing techniques for detecting binary gender bias in Section 2.2.2, as well as their suitability to be extended to non-binary gender evaluations. Last, I discuss works that consider gender bias for non-binary individuals in Section 2.2.3, where I explain the harms of cis-normativity in NLP in more detail. This section does not discuss bias evaluations and debiasing methods for coreference resolution systems, as those topics are discussed in Section 2.4.

2.2.1 Causes of gender bias in NLP systems

Gender bias can arise at various stages of the NLP pipeline. For example, during data collection, imbalanced representations of different gender identities may occur due to population biases (Olteanu et al., 2019). Additionally, the use of words related to certain gender identities in the training data may be limited to stereotypical contexts (Dinan et al., 2020), leading the NLP model to replicate these stereotypes, or to even amplify them (Zhao et al., 2017; Foulds et al., 2020; Wang and Russakovsky, 2021; Hall et al., 2022). Furthermore, the way in which NLP tasks are defined can also introduce bias, such as assuming a binary understanding of gender when setting up a coreference resolution task, which can lead to the reinforcement of cis-normativity (Cao and Daumé III, 2020). Moreover, during the annotation phase, factors such as the annotation guidelines (Geiger et al., 2020; Olteanu et al., 2019; Sap et al., 2019), the characteristics of the annotators (Olteanu et al., 2017; Patton et al., 2019) and the method used to aggregate the annotations (Pavlick and Kwiatkowski, 2019; Poirier, 2018) can be sources of bias. The used

bias evaluation metric can introduce bias as well, since bias evaluation metrics rely on a specific definition of bias and may not detect other forms of bias (Olteanu et al., 2019; Orgad and Belinkov, 2022). Finally, failing to take the underlying social roots of biases into account can lead to superficial debiasing methods that fail to address the causes of bias (Elsafoury and Abercrombie, 2023).

2.2.2 Binary gender bias detection

Methods to detect bias can be categorised into intrinsic and extrinsic bias metrics (Orgad and Belinkov, 2022). Intrinsic evaluations measure bias in the internal representations of NLP systems, while extrinsic evaluations, such as the evaluation carried out in the current study, consider biases in downstream tasks. I first describe intrinsic evaluations methods and then discuss extrinsic evaluation metrics. Finally I give an overview of gender bias detection studies for Dutch models and datasets.

Intrinsic evaluations

Two common methods for identifying bias in the internal representations of NLP systems, such as word embeddings, include (a) comparing the similarity between the representations of gender-neutral words and a *gender subspace* (Bolukbasi et al., 2016); and (b) the WEAT test (Caliskan et al., 2017), which involves comparing the similarity between the representations of two *target* word lists (e.g. male and female words) and two *attribute* word lists (e.g. positive and negative terms). Both of these methods are rooted in a binary conception of bias however, since they involve gender *pairs* in their methodology. Manzini et al. (2019) extend the WEAT test to multiclass setups, but despite evaluating racial and religious biases with more than two classes, they continue to treat gender as binary. Dev et al. (2021) do extend WEAT to analyse the representations of binary versus non-binary gender terms. They combine female and male terms to form one target word list of binary concepts, and create an additional target list of non-binary concepts. They find that terms relating to non-binary gender identities are more associated with negative sentiments than binary gender terms.

Extrinsic evaluation

Orgad and Belinkov (2022) categorise extrinsic evaluation methods into two groups: extrinsic prediction evaluations and extrinsic performance evaluations. The current study can be categorised in the latter group.

Extrinsic prediction methods evaluate bias through considering output probabilities, for instance by creating a male (e.g. *he is a doctor*) and female (e.g. *she is a doctor*) version of a sentence, and measuring the proportion of sentence pairs where the male version is assigned a higher probability (Nangia et al., 2020; Nadeem et al., 2021). Another option is to measure the *prediction gap*, by calculating the difference between the probabilities assigned to the two versions (Kiritchenko and Mohammad, 2018). This can

also be done by masking a word and comparing the probabilities of different continuations (Bordia and Bowman, 2019; Kurita et al., 2019; Nangia et al., 2020; Nadeem et al., 2021; Kiritchenko and Mohammad, 2018; Bartl et al., 2020). For instance, a sentence s could be

$$s = [MASK] \textit{ is a doctor}$$

where $p([MASK] = he|s)$ is compared with $p([MASK] = she|s)$.

On the other hand, extrinsic performance evaluations, such as the current study, quantify the effect of gender on downstream model performance by comparing performance scores for female and male sentences. Here, bias is computed as the *performance gap* between a female and male version of the same test set.This type of evaluation has been conducted for various tasks, such as abusive language detection (Park et al., 2018), coreference resolution (Webster et al., 2018; Rudinger et al., 2018; Zhao et al., 2018; Cao and Daumé III, 2020) and occupation classification (De-Arteaga et al., 2019).

Creating such a *gender-balanced test set*, with a female and male version of each sentence, can be done in two ways. One option is to apply a form of *Counterfactual Data Augmentation* called *gender swapping*, in which all male entities in an existing dataset are replaced with female entities and vice versa (e.g. Webster et al., 2018; Cao and Daumé III, 2020). This method has the benefit that naturally occurring text can be evaluated. Another option is to create templates, which can be filled with lists of words (e.g. Rudinger et al., 2018; Zhao et al., 2018; Dixon et al., 2018; Kurita et al., 2019). The benefit of this approach is that the templates can be precisely constructed to test difficult cases, and that a large number of test sentences can quickly be generated. Both types of datasets can be extended to non-binary individuals, for instance by adding a third version of each sentence with non-binary entities (Rudinger et al., 2018; Brandl et al., 2022).

Orgad and Belinkov (2022) argue that it is preferable to evaluate for extrinsic biases directly because research shows intrinsic biases do not necessarily correlate with extrinsic biases (Goldfarb-Tarrant et al., 2021; Elsafoury et al., 2022; Kaneko et al., 2022). Moreover, they contend that the harms caused by intrinsic biases are often unclear.

Dutch bias evaluations

Three works evaluate bias in Dutch NLP models directly. Firstly, Chávez Mulsa and Spanakis (2020) evaluate binary gender bias in static and contextualised Dutch word embeddings through WEAT-based tests (Caliskan et al., 2017; May et al., 2019). Secondly, McCurdy and Serbetci (2020) also use WEAT tests to compare binary gender biases across languages with different levels of grammatical gender saliency. Thirdly, Delobelle et al. (2020) investigate binary gender occupation bias in a Dutch RoBERTa (Liu et al., 2019) based model, through a template based association test (Kurita et al., 2019; May et al., 2019). They identify a correlation between the lexical gender of occupation words and the probability of the third-person singular pronoun of the same gender (*hij*/*zij*), but they find the male pronoun *hij* to generally be more probable than its female counterpart across most occupations.

Other studies evaluate biases in downstream tasks as part of multilingual evaluations: Hovy et al. (2020) investigate stylistic biases in commercial machine translation systems and Ghaddar et al. (2021) evaluate biases in named entity recognition systems. Finally, several studies apply NLP methods to identify biases in Dutch bodies of text. Koolen and van Cranenburgh (2017) analyse gender in two corpora of Dutch literary novels. Other works analyse gender bias (Wevers, 2019), cultural biases (Kroon et al., 2020; Kroon and van der Meer, 2021) and stereotypes (Fokkens et al., 2018) in news paper texts. To my best knowledge, no Dutch bias study to date looks into non-binary gender biases.

2.2.3 Non-binary gender bias

More recently, authors have started to question the binary conception of gender that is common in NLP studies and NLP bias research. In this section I firstly discuss the harms that can follow from cis-normativity in NLP, followed by a discussion of bias evaluations that target non-binary gender bias. Non-binary gender bias evaluations in coreference resolution systems are discussed in Section 2.4.

Harms

Cao and Daumé III (2020) discuss cis-normativity in NLP. From a random sample of 150 NLP studies that mention the word *gender*, they find that 92.8% considers gender as binary, and only 3.5% considers the use of personal singular *they* or neopronouns. In a similar evaluation Devinney et al. (2022) furthermore observe that a large portion of NLP papers fails to define gender. Cao and Daumé III (2020) argue that these practices can lead to the erasure of transgender individuals.

Dev et al. (2021) furthermore discuss the specific harms *misgendering* and *erasure*. An example of misgendering in downstream tasks is a coreference resolution system that links gendered pronouns to non-binary individuals. Continuing, erasure occurs when non-binary individuals are hidden, obscured or invalidated. Dev et al. explain this can for example occur when NLP systems predict the gender identities of individuals, but only consider binary gender identities as potential outcomes. If these predictions are then used in downstream applications, non-binary gender identities are entirely obscured. Another example of erasure in a downstream application is a system built on coreference resolution that fails to extract all relevant information about non-binary individuals because neopronouns are not recognised.

Continuing, Lauscher et al. (2022) contend that new pronouns can continuously be introduced. To illustrate this, they analyse the range of pronouns present in a large Reddit corpus by searching for tokens with the suffixes *-selves* or *-self*, and identify thousands of potential pronouns, where most tokens have very few occurrences. Because language models that rely on co-occurrence statistics typically have poor performances for infrequent words, Lauscher et al. argue that pronouns should be treated as an open word class, for language models to handle all recent pronoun phenomena. This requires a different way of dealing with pronouns entirely, and to this end they propose five desiderata:

1. "Refrain from assuming an individual's identity and pronouns."
2. "Allow for the existing set of pronouns as well as for neopronouns."
3. "Allow for novel pronouns at any point in time."
4. "Allow for multiple, alternating and changing pronouns."
5. "Provide an option for individuals to define their sets of pronouns."

Bias evaluations

In this section, I first discuss non-binary gender bias evaluations of language models, followed by a discussion of evaluations of the following downstream tasks: machine translation, abusive language detection, named entitiy recognition and part of speech tagging. Across these evaluations, the majority uses pronouns as a proxy for gender, whereas the way that bias is measured varies between most studies. In the final part of this section, I discuss debiasing efforts for non-binary gender biases.

Language models. Three studies evaluate non-binary gender biases in English language models, while two studies perform multilingual evaluations. Across these studies, all the evaluations use a different method and metric for measuring bias.

Starting with the English evaluations, Dev et al. (2021) perform an extrinsic prediction evaluation, in which they evaluate BERT's (Devlin et al., 2019) ability to correctly predict pronouns. For this task they create templates of two sentences, where a pronoun is visible in the first sentence but masked in the second:

(5) Alex went to the hospital for her appointment. [MASK] was feeling sick.

The performance is lower on gender-neutral pronouns than on gendered pronouns, with a further decline noted for neopronouns. Moreover, they evaluate whether BERT can distinguish between singular and plural pronouns through fine-tuning the model on a binary classification task, which involves predicting whether a masked pronoun is plural or singular. They compare the performance in distinguishing between (a) *he* and plural *they* and (b) singular and plural *they*, and observe a lower performance in the latter case.

Somewhat similarly, Hossain et al. (2023) present an extrinsic prediction evaluation framework for determining whether language models can effectively incorporate an individual's specified preferred pronouns, encompassing gendered, gender-neutral, and neopronouns. Their findings indicate a lower ability in using declared gender-neutral pronouns in comparison to gendered pronouns, with a further decline observed for neopronouns.

Taking a completely different approach, Watson et al. (2023) measure bias by computing the correlation between social attitudes of human subjects and BERT's surprisal for singular *they*, measuring surprisal as $-\log P(they|context)$. They find the strongest correlation of BERT's surprisal with the acceptance scores of participants who show moderate to low acceptance of non-binary individuals.

Moving on to multilingual evaluations, Brandl et al. (2022) evaluate Swedish, Danish and English language models. They again take a very different approach to measuring bias, using sentence perplexity as their bias metric, which they consider as an indicator of

processing difficulty. Across all languages, they create two versions of their test data: one containing the original sentences that include gendered pronouns, and a second version wherein they replace the pronouns by gender-neutral pronouns. For all languages they find significantly higher perplexity scores in the gender-neutral setting.

Finally, Martinková et al. (2023) evaluate Czech, Slovak and Polish language models. They measure bias in terms of toxicity and genderedness in generated sentence completions. They compare the completions across sentences that only differ in the used pronouns, which are either masculine, feminine or gender-neutral. Contrary to the studies described above, they observe the strongest bias for male subjects, as the completions for masculine pronouns lead to the highest toxicity scores.

Machine translation. Moving on to downstream tasks, within the domain of machine translation, Lauscher et al. (2023) investigate the translation of third-person pronouns, including both gender-neutral pronouns and neopronouns, across five languages and English. Additionally, they explore reverse translations from Danish to English. They use translation quality as their bias metric. Their observations reveal that in many instances of translation, gender-neutrality tends to diminish, and the incorporation of gender-neutral pronouns frequently results in grammatical and semantic errors within the translated text.

NER. Continuing, Lassen et al. (2023) evaluate intersectional biases in Danish named entity recognition (NER) systems, using names as proxies for gender and ethnicity. To measure non-binary gender bias, they use unisex name as a proxy. Their results report lower performances for unisex names, in comparison to gender-conforming names, across all systems.

POS-tagging. In part of speech (POS) tagging, Björklund and Devinney (2023) perform an extrinsic performance evaluation of Swedish systems. Particularly, they evaluate system performance on gendered pronouns and a gender-neutral pronoun, observing a lower performance for the latter group.

Abusive language detection. Finally, in abusive language detection systems, Sobhani et al. (2023) perform an extrinsic performance evaluation of gender bias across gender-neutral, female and male groups. They measure bias as the performance difference between the groups. In contrast to the results observed in the evaluations for other downstream tasks, they find the highest bias scores for the female group, while similar bias scores are reported for the gender-neutral and male groups.

Debiasing. To my best knowledge, only two studies consider debiasing NLP systems for non-binary gender bias. Firstly, Hossain et al. (2023) aim to improve language model performance in incorporating the declared preferred pronouns of an individual. They explore the application of few-shot in-context learning using explicit examples and note an enhancement in performance. However, the improvement plateaus rapidly, falling short of achieving comparable accuracy levels to those observed for gendered pronouns.

Second, Björklund and Devinney (2023) aim to improve the POS-tagging performance on the Swedish gender-neutral pronoun *hen*. To do so, they augment the training data with semi-synthetic data that includes the gender-neutral pronoun. Specifically, they create this semi-synthetic data by taking training sentences that include binary gendered

pronouns, and replace those pronouns by gender-neutral ones. This method is similar to the data transformation procedure of the current study (Section 3.3), but differs in two main ways. Firstly, the current study replaces pronouns across documents, rather than across sentences. The reason for this is that, while POS-tagging is a worl-level task, coreference resolution is a document-level task, and therefore pronouns should be consistently replaced throughout documents. Secondly, the data transformation algorithm applied in the current study includes two additional steps, particularly name anonymisation and gendered noun rewriting, which are further explicated in Section 3.3.

Subsequently, to perform debiasing, Björklund and Devinney (2023) fine-tune their models from scratch on the augmented data. Encouragingly, they observe that including the gender-neutral pronoun in 2% of the training sentences is sufficient to remove the performance gap. Because Björklund and Devinney also debias a downstream task, I consider their work to be the most similar to the current project. Moreover, their debiasing method is similar to my application of Counterfactual Data Augmentation. But, besides considering a different task and language than Björklund and Devinney, the current study makes the additional contributions of (1) evaluating a "further fine-tuning" debiasing configuration, besides fine-tuning from scratch; (2) additionally evaluating the delexicalisation debiasing method (see Section 2.4.2); (3) investigating the effect of the debiasing methods on the performance on previously unseen pronouns (Section 5.4).

2.3 Coreference resolution

Coreference resolution was first introduced in the 1970s (Woods, 1972; Winograd, 1972). This task entails deciding whether two referring expressions *corefer*, i.e. whether they refer to the same entity. *Referring expressions* or *mentions* are linguistic expressions that are used to refer to entities. A *cluster* is a set of coreferring expressions. An entity that only has a single mention is called a *singleton*. Figure 2.1 shows an example sentence, where the mentions with the same colour refer to the same entity. Here, the following coreference clusters can be identified:

1. *{je, hun huisgenoot, Thorn}*
2. *{hen, hun, Raven}*
3. *{Tobi}*

In this example, *Tobi* is a singleton. Within a cluster, the mentions that precede a certain mention m are called its *antecedents* while its later mentions are *anaphors* or *anaphoric*.

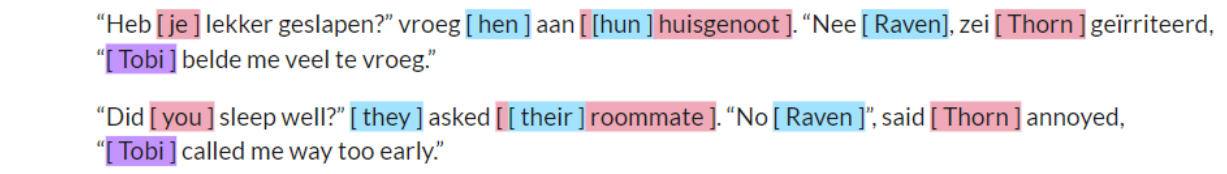


Figure 2.1: Coreference resolution task example, where mentions with the same colour refer to the same entity and thus belong to the same coreference cluster.

For instance, in the second cluster, the mention *hun* has antecedent *hen* and anaphor *Raven*. As can be observed from Figure 2.1, referring expressions can be nested: the referring expression *hun* is also part of the mention *hun huisgenoot*.

Coreference resolution is an important task because it forms the basis for other high-level NLP tasks such as information extraction, text summarization, machine translation and question answering (Ng, 2017). Coreference resolution consists of two subtasks, which modern end-to-end systems perform simultaneously (Lee et al., 2017, 2018): (1) mention detection, i.e. identifying the spans of referring expressions and (2) identifying the coreference links between the mentions. Systems are evaluated by comparing the indentified coreference links with gold coreference annotations.

Referring expressions are typically one of the following word classes:

(i) Names, e.g. *Sam Smith*;
(ii) Pronouns, such as *they*;
(iii) Indefinite noun phrases, like *an English singer*, which typically introduce a new entity;
(iv) Definite noun phrases, e.g. *the English singer*;
(v) Demonstrative pronouns *this* or *that*, which can be used individually or in combination with a noun phrase, as in *this song*.

But not all spans belonging to these word classes are referring expressions, which can be difficult for models to process correctly as they do resemble referring expressions. Such cases include (Jurafsky and Martin, 2021):

- Expletives or pleonasms: such as

  (6) *It* is possible that ..

  (7) *It* rains.

  where in both cases *it* does not refer to an entity.
- Generics: generic references do not refer back to specific entities. Examples include:

  (8) As a citizen *you* should vote.

  where *you* does not refer to an individual, and

  (9) I want to buy some tulips. *They* are blooming now.

  where *they* refers to the class of tulips in general and not to a specific entity.

Besides mention detection, coreference resolution itself is a hard task even once the mentions are identified, as the following following example illustrates:

(10) *Claudia asked Jessica to help her daughter.*

Here, we cannot know from syntax alone whether the pronoun *her* refers to *Claudia* or *Jessica*. While humans might easily resolve this ambiguity from the context or through their world knowledge, such cases are difficult for automated coreference resolution systems. In such situations, gender-neutral pronouns can pose an extra challenge, because

Dataset	Language(s)	Size	Time period of popularity
MUC-6 *MUC-7*	English	50 - 60 documents	1995 - 2004
ACE-1	English	225k words	2004 - 2010
ACE-2	English, Chinese	270K words	2004 - 2010
ACE03 *ACE04*	English, Chinese, Arabic	150 - 350K words	2004 - 2010
OntoNotes 5.0	English, Chinese, Arabic	1M English words, 1M Chinese words, 300K Arabic words	2010 - now

Table 2.2: Overview of popular coreference resolution datasets.

they do not provide gender information, nor number information in some cases: e.g. *they* can be used both in third-person singular and plural. Consider the following sentence:

(11) *Claudia asked Neil to clean their garden.*

Here, *their* could either refer to *Claudia*, *Neil* or to both *Claudia and Neil*.

In the rest of this section, I first give an overview of popular datasets for coreference resolution in Section 2.3.1, describe common methods in Section 2.3.2. and finally discuss coreference resolution evaluation metrics in Section 2.3.3.

2.3.1 Datasets

In this section I firstly describe influential English and multilingual coreference resolution datasets, of which Table 2.2 provides a summary. Secondly, I discuss Dutch coreference resolution datasets.

The first widely used coreference resolution corpora were published with the MUC-6 (1995) and MUC-7 (1998) conferences: the coreference section of the MUC-6 corpus consists of 30 training and 30 test texts, and the MUC-7 corpus contains 30 training and 20 test documents (Ng, 2017). Between 1995 and 2004 most coreference resolution systems were trained and evaluated on the MUC corpora (Ng, 2017).

In the successive period between 2004 and 2010 the four ACE corpora (ACE-1, ACE-2, ACE03 and ACE04) became the most popular. These datasets are much larger than the MUC corpora, e.g. ACE03 contains 100K tokens in the training set alone. They also include more languages: ACE-1 only contains English texts but ACE-2 additionally includes Chinese documents and ACE03 and ACE04 contain Arabic texts on top of that (Doddington et al., 2004). A further difference is that the ACE corpora only annotate entities of certain semantical categories (PERSON, ORGANISATION, GPE, FACILITY or LOCATION), whereas MUC includes all semantic types.

Continuing, the CoNLL 2011 (Pradhan et al., 2011) and 2012 shared tasks (Pradhan et al., 2012) popularised the OntoNotes 5.0 corpus (Hovy et al., 2006). This dataset

Dataset	Genre	Size	Annotated entity types
KNACK-2002	News magazines	122k words	All
COREA	News articles, transcribed spoken language, medical encyclopedia text	200k words	All
SoNaR-1	Mix	1M words	All
NewsReader	Wikinews articles	120 files	All
RiddleCoref	Literary text	33 documents with 4897 words on average	Person and object entities
ENCORE	News text	1115 documents	Event entities

Table 2.3: Overview of Dutch coreference resolution datasets.

contains one million hand annotated words in English and Chinese, and 300,000 words in Arabic. Similar to the MUC corpora, all semantic types are considered in this corpus. But contrary to earlier datasets, singletons are not annotated in OntoNotes, which greatly simplifies the task since singletons constitute between 60 and 70% of all mentions and they can be "hard to distinguish from non-referential NPs" (Jurafsky and Martin, 2021). This corpus remains popular to this day.

Dutch datasets

Table 2.3 gives an overview of Dutch coerference corpora. The first Dutch coreference corpus was KNACK-2002 (Hoste and De Pauw, 2006), which contains 267 Flemish news magazine documents, adding up to 122k words in total. This corpus was also used in the SemEval 2010 Shared Task (Recasens et al., 2010). A follow-up was the COREA project (Bouma et al., 2007; Hendrickx et al., 2008a,b), which produced an annotated corpus collecting more than 200k words. In a continuation of this project, the SoNaR-corpus was created (Reynaert et al., 2010), a 500M word dataset that contains published texts such as books, newsletters and magazines, and digital texts like websites, emails, teletext pages and chat messages. The SoNaR-1 corpus is a 1M word subset of this dataset that is annotated for coreference resolution (Schuurman et al., 2010). This is the largest Dutch coreference resolution corpus to date. Each instance in this corpus was annotated by a single annotator.

Continuing, the multilingual NewsReader corpus (Schoen et al., 2014) contains a Dutch component that includes 120 English Wikinews news article files that were translated to Dutch. This corpus was used in the CLIN26 shared task.[6]

More recently, van Cranenburgh (2019) published the RiddleCoref corpus, the first Dutch coreference corpus for literary texts, which collects 33 annotated documents with texts from contemporary novels in Dutch. Notably, only person and object entities are considered in this corpus, events and actions are excluded.

Finally, the recent ENCORE corpus (De Langhe et al., 2022) focuses on event coreference resolution, considering "[a]ny real, hypothetical or fictional situation that occurs, occupying a space-time and involving a number of participants" (De Langhe et al., 2022)

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

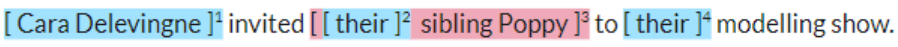


Figure 2.2: Coreference resolution task example, where mentions with the same colour refer to the same entity and thus belong to the same coreference cluster.

as an event. This dataset collects 1,115 Dutch news text in which coreference relations are not only annotated within documents but also across them.

2.3.2 Methods

In this section I describe common methods for coreference resolution. Firstly, I briefly discuss common modelling approaches: rule-based methods, feature engineering based supervised machine learning methods and end-to-end systems that use neural representation learning. For a more complete discussion of these methods, see Ng (2017). I subsequently describe Dutch coreference resolution models.

Modelling approaches

First of all, rule-based systems (e.g. Raghunathan et al., 2010; Lee et al., 2011, 2013; Krug et al., 2015; van Cranenburgh, 2019) typically base their decisions on features extracted from an NLP pipeline, particularly using information from named entity recognition and syntactic parsers. The advantages of rule-based systems are that they are transparent and can make global decisions based on the full document. The disadvantage is that they are knowledge-intensive. Moreover, for English, rule-based systems achieve a much lower performance than neural models (e.g. the rule-based model by Lee et al. (2011) achieves an F-score of 58.3 on the CoNLL 2012 shared task, where the SpanBERT based system by Wu et al. (2020) achieves an F-score of 83.1). For Dutch systems, a similar but less pronounced performance gap between rule-based and neural systems can be observed (see Table 4.5).

Continuing, popular supervised machine learning based systems often apply a *mention-pair architecture* (e.g. Aone and William, 1995; McCarthy and Lehnert, 1995; Soon et al., 1999, 2001) or a *mention-rank architecture* (e.g. Connolly et al., 1997; Versley, 2006; Denis and Baldridge, 2007; Wiseman et al., 2015; Lee et al., 2017). Across both architectures, the task is to find a *mention* m_j that is an antecedent of mention m_k. For example, let us consider the sentence in Figure 2.2, where the mention $m_k = [their]^4$ has the antecedent mentions $[their]^2$ and $[Cara\ Delevinge]^1$. For this mention, the models make predictions over the mention pairs:

$$s1 = \{([their\ sibling\ Poppy]^3, [their]^4), ([their]^2, [their]^4), ([Cara\ Delevinge]^1, [their]^4)\}.$$

The *mention-pair architecture* incorporates a binary classifier that makes local decisions about input pairs: it predicts for each pair whether it is either coreferring (1) or

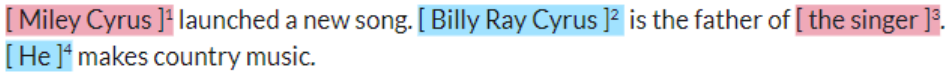


Figure 2.3: Coreference resolution task example, where mentions with the same colour refer to the same entity and thus belong to the same coreference cluster.

not (0), and stops after a positive prediction. A drawback of this method is that it does not directly compare antecedents. Therefore, the model might fail to identify a correct antecedent, in case another candidate antecedent that is considered earlier gets assigned a positive score. Even if this score would be lower than the score(s) for the correct antecedent(s), the model stops after this first positive prediction and thus fails to identify a correct antecedent.

The *mention-rank architecture* directly addresses this issue by simultaneously computing the probability of all candidate antecedent mentions, through applying one softmax function over the full set of candidates. The candidate with the highest probability is then selected. There is an additional dummy mention ϵ included in the set of candidates to indicate that the mention does not have an antecedent. This architecture is particularly popular amongst recent systems (Lee et al., 2017, 2018; Joshi et al., 2020; Dobrovolskii, 2021, e.g.). Both of these approaches can be implemented through a feature engineering based system or an end-to-end architecture.

Feature engineering based systems usually predominantly use features extracted from syntactic parsers and named entity recognition classifiers. Within the full set of features, there are typically features included that provide information about (a) the mention m_k and (b) the candidate antecedent m_j, such as their span length and a representation of the tokens in the span, and (c) the relation between m_k and m_j, e.g. the number of tokens between them and the cosine distance between their embeddings.

The current state of the art models for English are all neural models (Xu and Choi, 2020; Wu et al., 2020; Dobrovolskii, 2021), of which most use an encoder to create span representations. These models are end-to-end systems, and thus perform the tasks of creating span representations, detecting mentions from these spans and identifying coreference links between mentions simultaneously. Lee et al. (2017) for instance use a mention span-ranking architecture with a bidirectional LSTM to encode the spans. Lee et al. (2018) extend this span-ranking architecture with a method they call *higher-order inference*, to allow for conditioning on the entity cluster of the candidate antecedent: this means that when deciding whether *the singer* is the antecedent of *he* in the example sentence in Figure 2.3, the model softly conditions on its earlier prediction that *Miley Cyrus* and *the singer* corefer. Joshi et al. (2020) further improve this model by replacing the LSTM with SpanBERT. Dobrovolskii (2021) propose an alternative method, in which they predict the coreference links between words, rather than between spans. This way, they manage to reduce the complexity compared to the above mentioned models. Wu et al. (2020) adopt a completely different approach, framing coreference resolution as a question answering task. This method manages to achieve a good performance, but is particularly computationally expensive.

Overall, neural models achieve very successful results, but they increase computational

costs and require a higher amount of training data compared to rule-based and feature engineering based methods (Glasmachers, 2017).

Dutch models

The first Dutch coference system was a machine learning based mention-pair system (Hoste, 2005) that was trained and evaluated on the KNACK-2002 dataset and was further advanced in the COREA project (Hendrickx et al., 2008b). More recent Dutch systems include the rule-based dutchcoref (van Cranenburgh, 2019), the hybrid model by van Cranenburgh et al. (2021) and the end-to-end e2e-Dutch.[7] I will discuss each of these systems in more detail.

The dutchcoref system (van Cranenburgh, 2019) is a rule-based model that improves on earlier rule-based Dutch models (van der Goot et al., 2015; Recasens et al., 2010) and is based on the rule-based Stanford system (Lee et al., 2011, 2013). It performs the steps of mention detection, quote attribution and coreference resolution. During mention detection, candidate mentions are identified, and a set of filter rules is subsequently applied to improve the precision. One of these filter rules involves inferring binary gender, number and animacy of pronouns, names and nouns. This step might hinder resolving gender-neutral pronouns because they do not fall into binary gender categories. Next, in the quote attribution step, quoted speech sections are marked and attributed to their speaker. Finally, coreference resolution is performed using entity-centric *sieves* to combine mentions into entities.

Second, the hybrid model by van Cranenburgh et al. (2021) is an extension on dutchcoref, in which the authors experiment with three feed-forward neural classifiers to perform the tasks of mention span detection, mention attribute classification and pronoun resolution. The mention attribute classifier replaces dutchcoref's attribution of binary gender, number and animacy in a multi-label classification setup. This model is trained on the RiddleCoref corpus and compared to e2e-Dutch (described below) and an updated version of dutchcoref. The authors find mixed results: including the mention span and attribute classifiers give the best CoNLL score, while the adapted dutchcoref model gives the best LEA score. The pronoun classifier does not improve the scores of any metric.

Finally, e2e-Dutch is a neural mention-span ranking end-to-end system based on Lee et al. (2018). This model consists of two main steps: (1) creating span representations and (2) predicting antecedent scores for pairs of spans, i.e. deciding whether span s_j is an antecedent of span s_i. In the first step, spans representations are created by combining pre-trained fastText common crawl embeddings (Grave et al., 2018) and pre-trained contextualised representations from BERTje (de Vries et al., 2019) through a Convolutional Neural Network and a bidirectional LSTM. In the second step, a neural span-ranking model then predicts for each span i its antecedent span j, where an alternative output is dummy antecedent ϵ, which represents (a) that the span is a mention that has no antecedent or (b) that the span is not a mention. E2e-Dutch extends Lee et al.'s (2018)

[7]No paper was published to introduce this model, but its implementation can be found at https://github.com/Filter-Bubble/e2e-Dutch

(a) Reference clusters G (b) Hypothesis clusters H

Figure 2.4: An example of reference clusters (a) and hypothesis clusters (b).

architecture by supporting singletons. Poot and van Cranenburgh (2020) compare this model with dutchcoref and find that dutchcoref outperforms e2e-Dutch on the RiddleCoref corpus, but that e2e-Dutch performs better by a larger margin on the SoNaR-1 corpus.

2.3.3 Evaluation metrics

Coreference resolution systems are evaluated by comparing the set of gold clusters G, which are annotated by humans, with the set of hypothesis clusters H, i.e. the clusters identified by the model. In Figure 2.4 (a) are the gold clusters, which include

$$G = \{g_1 = \{a, b, c\}, g_2 = \{e, f\}, g_3 = \{d\}\}$$

and (b) are hypothesis clusters:

$$H = \{h_1 = \{a, b\}, h_2 = \{c, e, f\}, h_3 = \{d\}\}$$

There are five common metrics for evaluation: mention based B^3 (Bagga and Baldwin, 1998), entity based CEAF (Luo, 2005), link based MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011; Luo et al., 2014) and link based entity aware LEA (Moosavi and Strube, 2016). Additionally, there is the CoNLL score, which is the average of the MUC, B^3 and CEAF F-scores. Each of these metrics has its own recall r and precision score p, and correspondingly computes its F-score by taking the harmonic mean:

$$F = \frac{2pr}{p + r}$$

However, except for LEA, all of these metrics have been criticised and demonstrated to be flawed (Luo, 2005; Luo and Pradhan, 2016; Moosavi and Strube, 2016; Denis and Baldridge, 2009; Stoyanov et al., 2009). For instance, B^3 assigns a perfect recall score when systems classify all mentions as part of the same cluster. LEA was particularly designed to overcome the limitations of its predecessors and I therefore use LEA as the main evaluation metric in this work. The discussion in this section is therefore limited to the LEA score, but refer to Appendix A for a description of the other metrics.

LEA (Moosavi and Strube, 2016) is a link based and entity aware metric, which first of all means that it evaluates the coreference *links* within clusters. In the clusters in Figure

2.4, we can identify the following links:

$$links_G = \{g_1 = \{(ab), (bc), (ac)\}, g_2 = \{(ef)\}\}$$

$$links_H = \{h_1 = \{(ab)\}, h_2 = \{(cf), (ce), (ef)\}\}$$

The clusters g_3 and h_3 are not included in $links_G$ and $links_H$ respectively because they are singletons and thus do not contain links.

Secondly, it takes the relative importance of the cluster the link belongs to into account, to ensure that larger clusters carry more weight towards the final score. The importance measure is adaptable, but the authors use the size of cluster g_i, i.e. $importance(g_i) = |g_i|$.

Continuing they use the following *resolution score* for each cluster g_i, which can be interpreted as the portion of coreference links that are correctly resolved:

$$resolution\_score(g_i) = \sum_{h_j \in H} \frac{link(g_i \cap h_j)}{link(g_i)}$$

In order to deal with singletons, these entities are considered to have self-links: a link that connects a mention to itself. Only singletons have self-links and therefore if g_i is a singleton, $link(g_i \cap h_j) = 1$ only if h_j is a singleton with the same mention as g_i.

The recall is then computed as:

$$r = \frac{\sum importance(g_i) \times resolution\_score(g_i)}{\sum importance(g_i)} = \frac{\sum_{g_i \in G}(|g_i| \cdot \sum_{h_j \in H} \frac{link(g_i \cap h_j)}{link(g_i)})}{\sum_{g_z \in G} |g_z|}$$

When we compute recall resolution and importance score for the entities in G of Figure 2.4, we get the following values:

g_i	Importance score $\lvert g_i \rvert$	Resolution score $\sum_{h_j \in H} \frac{link(g_i \cap h_j)}{link(g_i)}$
g_1	3	$\frac{1}{3}$
g_2	2	1
g_3	1	1

Giving the following outcome :

$$r = \frac{3 \times \frac{1}{3} + 2 \times 1 + 1 \times 1}{3 + 2 + 1} = \frac{4}{6} = \frac{2}{3}$$

And similarly the precision is computed through:

$$p = \frac{\sum importance(h_i) \times resolution\_score(h_i)}{\sum importance(h_i)} = \frac{\sum_{h_i \in H}(|h_i| \cdot \sum_{g_j \in G} \frac{link(h_i \cap g_j)}{link(h_i)})}{\sum_{h_z \in H} |h_z|}$$

Which results in the following precision score:

h_i	Importance score $\lvert h_i \rvert$	Resolution score $\sum_{h_j \in H} \frac{link(h_i \cap g_j)}{link(h_i)}$
h_1	2	1
h_2	3	$\frac{1}{3}$
h_3	1	1

$$p = \frac{2 \times 1 + 3 \times \frac{1}{3} + 1 \times 1}{3 + 2 + 1} = \frac{4}{6} = \frac{2}{3}$$

This finally gives an F1-score of:

$$F1 = \frac{2 \times \frac{2}{3} \times \frac{2}{3}}{\frac{2}{3} + \frac{2}{3}} = \frac{2}{3}$$

2.4 Gender bias in coreference resolution

This section concerns gender bias in coreference resolution systems. I start by discussing datasets for identifying gender bias in these systems (Section 2.4.1). Next, I discuss studies that experiment with debiasing coreference resolution systems (Section 2.4.2).

2.4.1 Bias evaluations

Here I describe existing datasets for identifying gender bias in coreference resolution systems, of which Table 2.4 gives an overview. From these six datasets, three take inspiration from the general setup of the Winograd schema challenge dataset (Levesque et al., 2012). This challenge dataset contains coreference resolution problems that are easily disambiguated by humans but require a deeper understanding of language than superficial pattern matching, for instance because they incorporate common-sense reasoning and world knowledge. The corpus consists of pairs of sentences with coreference questions that differ only in one word, but this word changes the correct resolution. An example of such a pair can be found below, where boldface highlights the word that differs between the two sentences:

(12) a. The dog chased the cat, which ran up a tree. It waited at the **top**. Which waited at the **top**? Answer: The cat.
b. The dog chased the cat, which ran up a tree. It waited at the **bottom**. Which waited at the **bottom**? Answer: The dog.

The structure of this dataset has formed the basis for three of the gender bias detection datasets described below, which likewise consist of sentence pairs that differ only in one word that causes them to have a different correct resolution.

Dataset	Text type	Gender bias type	Pronouns	# Sentences
WinoBias	Templates	Occupational	*he, she*	3160
Winogender	Templates	Occupational	*he, she, they*	720
WinoNB	Templates	Non-binary	Singular / plural *they*	4077
GAP	Naturally occurring Wikipedia sentences	Ambiguous pronoun resolution	*he, she*	4454
MAP	Naturally occurring Wikipedia senstences where gender clues are ablated	Identifying what gender-related information affects performance	Originally: *he, she* After ablation: *they, xey, ze*	549 in 9 ablation settings (total 5490)
GICoref	Naturally occurring gender-related phenomena	Non-binary, misgendering	*He, she, they* and neopronouns	95 documents

Table 2.4: Overview of gender bias datasets for coreference resolution

WinoBias

Firstly, the WinoBias dataset (Zhao et al., 2018) tests for binary gender occupation bias. In this template-based dataset of 3,160 sentences, the sentences contain a gendered pronoun and two occupation words, and the task is to identify which of the two occupation words is the antecedent of the pronoun. Each sentence has a stereotypical version, in which the pronoun gender aligns with the stereotypical gender of the antecendent occupation, and an anti-stereotypical version, in which the pronoun is of the opposite binary gender. The example below illustrates this, where underlined text indicates the correct pronoun resolution, and boldface highlights the difference between the sentences:

(13) a. Anti-stereotypical: The developer argued with the designer and slapped **him** in the face.
b. Pro-stereotypical: The developer argued with the designer and slapped **her** in the face.

The authors evaluate three coreference resolution systems on occupational gender bias using this dataset. They find that all systems suffer from strong gender biases, performing better in the stereotypical conditions.

Winogender

Secondly, Rudinger et al. (2018) introduced the Winogender schemas, which were similarly created to identify occupational gender bias. The dataset consists of 120 hand-crafted templates that each contain an occupation, a participant and a pronoun. Like in the WinoBias dataset, the task here is to find the antecedent of the pronoun. The templates have two versions which differ slightly, thereby changing the correct pronoun resolution: in version (a) the antecedent is the occupation and in version (b) it is the patient. A filled out example of such a template pair is:

(14) a. The technician told the customer that she **had completed the repair**.
b. The technician told the customer that she **could pay with cash**.

Each template pair is filled with three pronouns (*he, she, they*). Additionally, the authors create a version of all sentences in which the patient is replaced by *someone*. So another version of the example above could look as follows:

(15) The technician told **someone** that they could pay with cash.

This creates a total dataset of 720 sentences.

The authors evaluate three systems for gender bias using Winogender, and find all systems to suffer from bias. For instance, all systems are more likely to resolve *he* pronouns with an occupation antecedent than the other pronouns. Additionally, they find *they* pronouns are commonly resolved neither as a participant nor as an occupation, which the authors ascribe to "the number ambiguity of "they/their/them.""(Rudinger et al., 2018).

Hansson et al. (2021) publish a Swedish version of this test set, called SweWinogender. Moreover, Brandl et al. (2022) extend English Winogender to include the neopronoun *xe*. They compare accuracy scores across pronouns and find that while the performance for binary gendered pronouns is above 40%, the accuracy for *they* drops to 28% and that of *xe* is 0%. *Xe* was rarely recognised as a mention and when it was, it was incorrectly resolved.

WinoNB

Thirdly, Baumler and Rudinger (2022) recently introduced the 4077 sentence template-based WinoNB dataset, which was created to test whether systems can disambiguate between singular and plural *they*. To this end, all sentences contain a named individual, an occupational referent which refers to a group of people and the pronoun *they*, as in the following example sentence pair:

(16) a. The paramedics tried to help Riley even though they **knew it was too late**.
b. The paramedics tried to help Riley even though they **were already dead**.

The task here is again to identify the antecedent of the pronoun. But the occupations are included in this dataset for a different reason than in earlier datasets: whereas in WinoBias and Winogender they served to identify occupational gender bias, in this dataset they provide the required semantic information to resolve the pronoun. Because *they* is used as plural in (a) but as singular in (b), such sentence pairs allow for evaluating the resolution abilities of plural and singular *they* by comparing the performance scores across the two settings. Additionally, authors add a version of the sentences where they replace the name with *someone*, to evaluate generic singular they:

(17) The paramedics tried to help someone even though they **were already dead**.

The authors evaluate five models on the WinoNB dataset. They find that the models on average perform over 90 times better for plural *they* than for singular personal *they*.

Continuing, the generic singular *they* case (example 17) is six times more likely to be correctly resolved than personal singular *they* (example 16b). This indicates that *they* as a gender-neutral pronoun remains poorly handled by state-of-the-art coreference models, even despite the fact that other usages of the pronoun are handled correctly.

GAP

Not all gender bias test sets for coreference resolution are based on the Winograd Schemas. Webster et al. (2018) introduce the *Gender Ambiguous Pronoun* (GAP) dataset that includes sentences with ambiguous pronoun resolution, such as the following sentence:

(18) In May, *Fujisawa* joined *Mari Motohashi*'s rink as the team's skip, moving back from Karuizawa to Kitami where she had spent her junior days.

Here the task is again to find the antecedent of the pronoun, which is ambiguous between the two entities, which are of the same binary gender in all sentences. Gender bias can then be tested for by comparing system performance for female and male pronouns. The corpus consists of 4,454 manually annotated Wikipedia sentences, with an equal number of female and male sentences. The authors evaluate system performance of four models and find lower overall scores in the female case for all models.

Kurita et al. (2019) furthermore evaluate a BERT-based coreference resolution system (Tenney et al., 2019) on GAP. Despite the fact that their training data is balanced in terms of gender, their results indicate a better performance for male pronouns, which the authors attribute to bias in the BERT representations.

MAP

Building on GAP, Cao and Daumé III (2020) create the *Maybe Ambiguous Pronoun* (MAP) dataset, a dataset that is similar to GAP but where gender clues are (partially) hidden and where the constraint that antecedents should have the same gender is lifted. They start with Wikipedia sentences and hide gender cues by four rule-based operations:

- Replace third-person pronouns with gender-neutral pronouns *they, xey* and *ze*;
- Replace names with a random first initial and a random last name;
- Replace semantically gendered nouns with gender-neutral nouns (e.g. *sister* → *sibling*);
- Take out terms of address such as *Mrs.*

They experiment with ablating different combinations of these gender cues from the Wikipedia data, to see what effect this information has on the resolution performance of both human annotators and coreference resolution systems. For human annotators, they find that taking out gendered pronouns affects the performance most strongly and that names also have a significant effect. Continuing, they evaluate five coreference resolution systems, which follow the same trends as that of human annotators, with particularly strong performance drops for ablated pronouns.

GICoref

Finally, Cao and Daumé III introduce the GICoref dataset in the same study, which contains naturally occurring data of "gender-related phenomena"(Cao and Daumé III, 2020): it represents, among others, (i) genderfluid individuals who are referred to by varying names and pronouns throughout the texts, (ii) people in queer relationships and (iii) people that are being misgendered. Moreover, the data contains a relatively balanced distribution of the pronouns *he*, *she*, *they* and neopronouns. In total the corpus contains 95 documents from (a) Wikipedia pages about non-binary individuals, (b) LGBTQ periodicle articles and (c) fan fiction stories.

They test five systems on this dataset and find disappointing results: the best LEA F1-score is 34%, while e.g. the neural Stanford model achieves a F1-score of 60% on the CoNLL-12 shared task (Moosavi, 2020). Because the recall scores are particularly low, the authors further analyse whether pronouns are recognised as mentions at all: while 95% of the binary pronouns are detected and 90% of *they* pronouns are found, only 13% of the neopronouns are identified.

Non-English bias evaluations

Brandl et al. (2022) perform an evaluation of a Danish coreference resolution model. To my best knowledge, this is the only gender bias evaluation of a coreference resolution system in a language other than English. They do not introduce a dataset for their evaluation, but instead use a regular Danish coreference dataset (Barrett et al., 2021). Because this corpus only contains gendered pronouns, they extend it by creating a gender-neutral version in which they replace all the gendered pronouns with gender-neutral pronouns and neopronouns. The authors report the overall coreference resolution performance on the original and the gender-neutral data, in terms of the CoNLL score. They report only a small drop in performance for the gender-neutral data, compared to the original dataset. However, it remains unclear whether the gender-neutral pronouns are correctly resolved by the model, since (1) the CoNLL score is the average of three metrics that have all been demonstrated to be flawed (See section 2.3.3) and (2) the authors do not mention what portion of the dataset constitutes pronouns.

2.4.2 Debiasing

To the best of my knowledge, only three studies have explored debiasing coreference resolution systems from gender bias. In this section I describe these studies.

Firstly, Zhao et al. (2018) explore a combination of three techniques for debiasing two coreference resolution systems from binary gender bias. The considered debiasing techniques are (i) anonymising all names in the training data, (ii) using debiased word embeddings (Bolukbasi et al., 2016) and (iii) creating a gender-balanced version of the training data, in which originally 80% of the gendered pronouns are male. They do so through a form of Counterfactual Data Augmentation (CDA) called *gender swapping*: for

every male entity m in sentence s they include a copy of s in which m is replaced with a female entity f, and vice versa. Entities are replaced through a rule-based method, which e.g. includes the rules *he* $\rightarrow$ *she*, and *father* $\rightarrow$ *mother*. They find that using a combination of these three techniques reduces the bias score of 19.25 F1-points down to only 2.2 points.

In a follow-up study Zhao et al. (2019) perform a similar exploration of debiasing a coreference resolution system that partially relies on ELMo embeddings (Peters et al., 2018) from binary gender bias. Specifically, this system (Lee et al., 2018) forms span representations based on a combination of GloVe (Pennington et al., 2014) and ELMo embeddings. They compare two debiasing methods: (1) combining gender swapping and debiasing GloVe word embeddings, following Zhao et al. (2018) and (2) a method called *neutralisation* to mitigate biases introduced by the ELMo embeddings. This latter technique entails combining the contextualised word embeddings of a sentence (e.g. c_m for a male sentence) with the embeddings of the gender-swapped version of this sentence (e.g. c_f, the female version), and using their average as sentence representations ($c = \frac{c_m + c_f}{2}$). Rather than applying the latter method to training instances and retraining the coreference model, they only apply it to the test instances. A benefit of neutralisation over gender swapping therefore is that neutralisation does not require additional training. But, while gender swapping in combination with using debiased embeddings again proves to be effective, neutralisation proves less effective.

While gender swapping appears promising, it requires inserting pronouns into the training data, so it can only be used for debiasing a predefined set of pronouns. However, as Lauscher et al. (2022) point out, new pronouns can and likely will be introduced and popularised in the future. For this reason, Lauscher et al. opt for a very different approach to processing pronouns: they argue that pronouns should be treated as an open word class and should therefore be *delexicalised*, i.e. all pronoun tokens should be replaced by their POS tag, so that the lexical form of the pronoun becomes irrelevant. They evaluate how this strategy affects model performance by training a RoBERTa based model (Dobrovolskii, 2021) on three versions of the OntoNotes 5.0 dataset:

1. the original dataset,
2. a version of the data wherein pronouns are delexicalised in the test split.
3. a version of the data wherein pronouns are delexicalised in all splits.

Performances are evaluated in terms of MUC, B^3 and CEAF. While averaged performance scores drop with 21.2 F1 points when pronouns are delexicalised in the test set alone, this drop reduces to 4.2 points when pronouns are delexicalised in all splits. However, the authors do not evaluate the most realistic scenario in which the pronouns are delexicalised during training but not during testing. This scenario is more realistic because coreference resolution systems will be applied to naturally occurring data in which pronouns appear in their lexical form. One might argue that input sentences could be preprocessed by covering pronouns, but this carries the risk that previously unseen pronouns might not be recognised, which defeats the purpose of delexicalisation entirely. Therefore, a more thorough evaluation is necessary to find out whether delexicalisation increases the system's ability to handle previously unseen and infrequent pronouns. Additionally, it might well be that (in line with the findings by Kurita et al. (2019)) the RoBERTa-based system

still prefers *he* pronouns even after being trained on delexicalised data due do biases in the RoBERTa representations.

2.4.3 Conclusion

Overall, the studies that evaluate coreference resolution systems on processing English gender-neutral pronouns and neopronouns consistently find very poor performances (Cao and Daumé III, 2020; Baumler and Rudinger, 2022; Brandl et al., 2022). So far, Lauscher et al. (2022) are the only ones to explore a gender debiasing approach for gender identities beyond the binary, but the effectiveness of their method requires further inspection. Except for Brandl et al. (2022) who consider Danish, all current evaluations look into English. The current study contributes to this line of work by evaluating Dutch coreference resolution systems. Moreover, I evaluate two debiasing methods: (1) the gender-swapping debiasing method used by Zhao et al. (2018), which I extend to include two gender-neutral pronouns (*hen* and *die*) and (2) delexicalisation (Lauscher et al., 2022). I do not evaluate using debiased embeddings because the main issue with gender-neutral pronouns is *not* that language models pick up on their stereotypical associations, but that they cannot process them at all.

3 Data

In this chapter, I firstly perform an analysis of the data used in this study in Section 3.1. Second, I describe the steps performed to preprocess the data (Section 3.2), in order to use it for training and evaluating the model. Finally, in Section 3.3, I describe the transformation steps for inserting gender-neutral pronouns into corpus.

3.1 Data analysis

In this section I analyse the data used in this study. I start by giving a general overview of the dataset, continue by zooming in on the usage of pronouns in the corpus and finally analyse the presence of gender-neutral and neopronouns in the data.

I use the subset of the SoNaR corpus (Reynaert et al., 2010) that is annotated for coreference resolution, called SoNaR-1. The SoNaR corpus was created as part of the STEVIN-funded SoNaR project, which ran from 2008 to 2011. All the documents in this corpus thus originate from 2011 or earlier, while the first Dutch gender-neutral pronoun was only introduced in 2016 (Transgender Netwerk Nederland, 2016). The 1M-token

Text domain	**Number of tokens**
Autocues	205,040
Books	2,008
Brochures	88,451
E-magazines, E-Newsletters	12,769
Guides, Manuals	28,410
Legal texts	6,468
Magazines	142,840
Minutes	1,655
Newsletters	8,543
Newspapers	81,130
Policy documents	30,021
Press releases	22,261
Proceedings	14,396
Reports	30,751
Speeches	17,320
Websites	47,841
Wikipedia	260,533
Total	1,000,437

Table 3.1: Composition of the SoNaR-1 corpus. Table adapted from Oostdijk et al. (2013).

	Mean	Median	Std	Min	Max
Sentences per document	69.7	31.0	97.3	3	950
Tokens per sentence	16.7	15.0	10.9	1	199
Clusters per document	32.2	16.0	42.9	0	439
Referents per cluster	4.1	2.0	7.2	2	307
Referents per sentence	1.9	2.0	1.7	0	18

Table 3.2: Core statistics of the SoNaR-1 corpus. The reported cluster statistics do not include singleton clusters. In total, the corprus contains 27,724 non-singleton clusters.

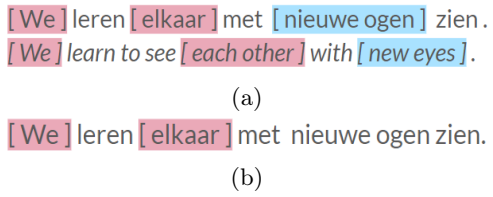


Figure 3.1: Example sentence in the SoNaR-1 corpus before (a) and after (b) removing singleton clusters. Here, the mention `[nieuwe ogen]` (*new eyes*) is a singleton, so this cluster is removed in (b), while the coreferring mentions `[we]` (*we*) and `[elkaar]` (*each other*) are maintained.

SoNaR-1 subset consists of 861 documents, which together comprise 59,960 sentences. SoNaR-1 gathers documents from various domains, e.g. magazines, Wikipedia articles, brochures, websites, legal texts, autocues and press releases. Table 3.1 provides an overview of the composition of the corpus in terms of domains. For the full documentation of this subset, refer to Oostdijk et al. (2013). The coreferential relations were manually annotated, using the COREA guidelines (Hendrickx et al., 2008a) as the basis for the annotations. The corpus additionally contains manually checked annotations for syntactic dependency trees, spatio-temporal relations, semantic roles and named entities.

An overview of the main statistics of the SoNaR-1 corpus can be found in Table 3.2. Moreover, the table presents details on the number of clusters and referents within the dataset. Notably, singleton clusters, defined as clusters comprising only one mention, have been excluded from this analysis. An illustrative example of a singleton cluster (`[nieuwe ogen]`) in the SoNaR-1 corpus is depicted in Figure 3.1a. While singletons are annotated in the SoNaR-1 corpus, they are not annotated in the OntoNotes corpus. Consequently, the wl-coref model (Dobrovolskii, 2021), originally trained on OntoNotes, is not inherently equipped to handle singletons. This same challenge was addressed in the Dutch e2e-Dutch model by modifying the English base model to predict singletons. However, given that pronouns typically refer to proper nouns and thus seldom exist as singletons, recognising singletons appears to be of limited relevance for the current study. Furthermore, coreference resolution systems are frequently trained and evaluated on corpora without singleton annotations (Lee et al., 2018; Joshi et al., 2020; Liu et al., 2022; Bohnet et al., 2023). Therefore, I decide to maintain the model's configuration in this regard.

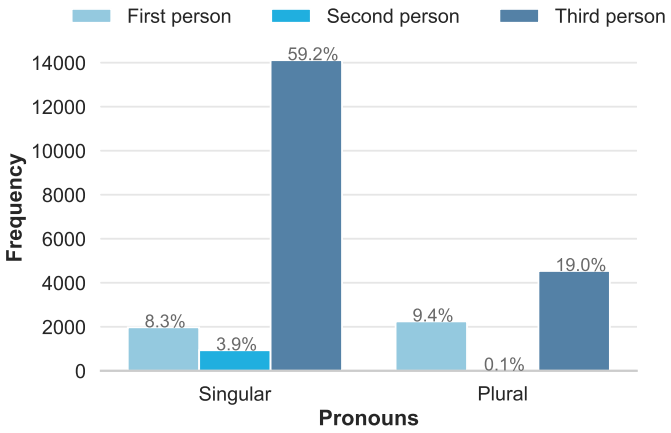


Figure 3.2: Distribution of personal pronouns based on person and number: the prevalence of third-person singular pronouns strongly outweighs other categories, comprising 59.2% of the entire set of personal pronouns.

To mitigate any potential impact of singleton annotations on the model's performance scores, I remove the singleton annotations from the corpus. As I remove the singletons for training and testing, I make sure to also remove them during the data analysis. Therefore, the reported statistics in Table 3.2 do not include singletons.

Subsequently, I turn my attention to the focal point of this investigation: pronouns. To get an idea of prevalence of pronouns within the corpus, I compute the percentage of tokens that is labeled with the POS `PRON`-label, which encompasses all types of pronouns. Within the full dataset, 4.5% of the tokens are identified as pronouns. The corpus provides more detailed POS-tag annotations, differentiating among others between various types of pronouns, their grammatical function, number, and gender where applicable. Personal and possessive pronouns, central to this study's objectives, collectively constitute 2.9% of all tokens, amounting to 29,083 words. Furthermore, 10,187 sentences (17.0%) encompass at least one possessive or personal pronoun.

Figure 3.2 shows the composition of personal and possessive pronouns, specifically focusing on person and number. The figure shows that third-person singular pronouns dominate, constituting 59.2% (14,113 tokens in total). This observation makes sense considering the corpus' sources (Table 3.1), which predominantly feature Wikipedia and published media texts. In such contexts, a higher prevalence of third-person singular pronouns can be expected, especially compared to e.g. direct communication (where the second person is more common) or social media texts (where people often speak in the first person).

Moving forward, Figure 3.3 shows the the distribution of third-person singular pronouns[1] in terms of gender and grammatical function. It can be observed that direct objects are the least common, while subjects marginally surpass possessives in frequency. Notably, the plot highlights a striking gender imbalance in the corpus, with 79.1% of all third-person pronouns identified as male.

[1]The third-person pronoun *het* was excluded from this analysis, as it is only used for inanimate objects.

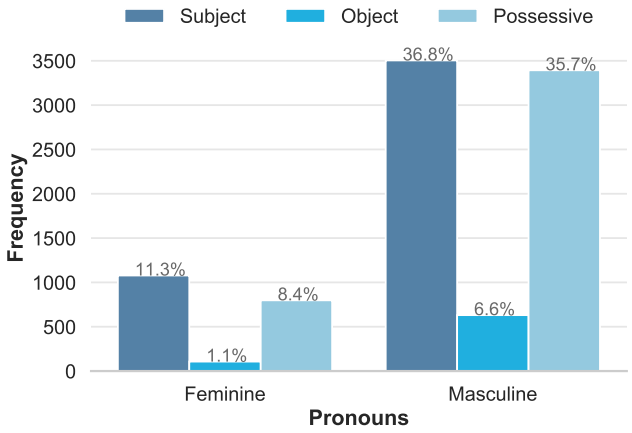


Figure 3.3: Distribution of third-person pronouns with respect to gender and grammatical function: the distribution is highly skewed towards masculine pronouns, which constitute 79.1% of the third-person pronouns.

Concluding the analysis, I examine the frequency of gender-neutral pronouns and neopronouns in the corpus. The gender-neutral pronoun *die* appears a total of 7,282 times in the data, with 1,995 instances functioning as a demonstrative pronoun, 5,268 instances as a relative pronoun, and 19 occurrences with an alternative label. Notably, *die* does not manifest as a personal pronoun. Its possessive form, *diens*, is identified 35 times in the corpus, consistently as a demonstrative pronoun.

Furthermore, the gender-neutral pronoun *hen* is identified 290 times, exclusively functioning as a third-person plural object personal pronoun. Its possessive counterpart, *hun*, is observed 1,865 times as a third-person plural possessive pronoun, with neither *hen* nor *hun* being used as third-person singular form throughout the corpus. Among the neopronouns explored in this study, *vij* appears once and *zeer* (also connoting *very* or *sore*) occurs 295 times as an adverb. The remaining neopronouns considered in the Unseen Pronouns Experiment (Section 5.4), namely *dee, dem, dijr, dij, dem, dijr, nij, ner, nijr, vijn, vijns, zhij, zhaar* and *zem*, do not feature in the dataset. Consequently, none of the neopronouns considered in this study appear as pronouns within the corpus.

3.2 Data preprocessing

The original format of the SoNaR-1 corpus is MMAX. However, the model that I use in this study, wl-coref, necessitates data in jsonlines format. Poot and van Cranenburgh (2020) have published code to transform the SoNaR-1 MMAX data into the standard CoNLL-2012 format, and Dobrovolskii (2021) provide code to rewrite CoNLL-2012 formatted data into jsonlines format. I consecutively apply these transformation steps to obtain data formatted in jsonlines. Moreover, I use the genre-balanced 70/15/15 division by Poot and van Cranenburgh (2020), to divide the documents over train/dev/test splits.[2]

[2] https://gist.github.com/CorbenPoot/ee1c97209cb9c5fc50f9528c7fdcdc93

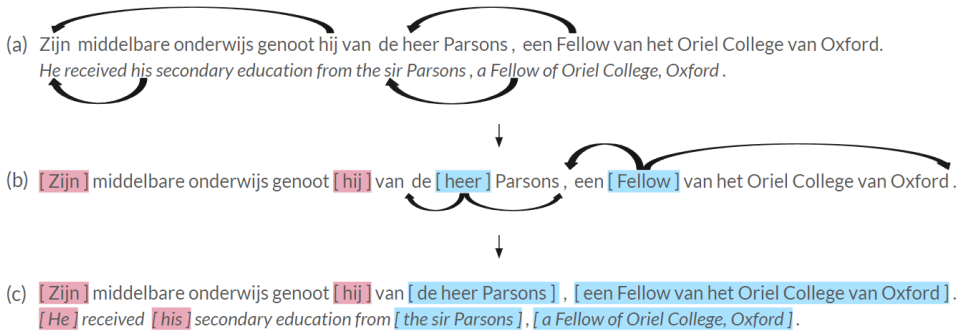


Figure 3.4: Illustration of the wl-coref model steps for making coreference predictions, using an example sentence. In the initial phase (a), the model predicts, for each word in the corpus, its antecedent or assigns a dummy variable if the word lacks an antecedent. Notably, the training data exclusively incorporates links between span heads, thereby training the model to predict antecedents solely for tokens serving as the head of their respective mentions. Subsequently, in the second step (b), the model, having already discerned coreference relations between mention heads, is tasked with identifying the boundaries of each mention span. This phase utilises the complete span boundaries as training data. The ultimate phase (c) depicts the output generated for the specified example sentence.

As an integral component of the CoNLL-2012 to jsonlines conversion process, the coreference data is enriched with syntactic details, including the syntactic head and dependency relation for each word. This information is used in the subsequent preprocessing stage to (1) identify the head of each mention span, and (2) to separately store the coreference links between heads. The head of a mention span is defined as the only word in the span with a head *outside* of the span, or as the root of the sentence. For instance, in the mention span *a Fellow of Oriel College, Oxford* the head is defined as *Fellow*. These preprocessing procedures are imperative to the training paradigm of the wl-coref model, which involves a two-step coreference resolution process that is illustrated by Figure 3.4. In the first step, the model is trained to recognise the antecedents of span heads exclusively, utilising the coreference links between span heads as training data. Subsequently, in the second step, the model predicts the boundaries of each mention span, determining which other words in the sentence are part of this particular mention span. This is achieved by employing the span boundaries of the original mentions as training data. More information about the training procedure can be found in Section 4.

The syntactic information that is added to the data can be acquired by a parser, such as Alpino (van Noord, 2006). For the SoNaR-1 corpus, this syntactic information is available within the dataset in the form of manually checked Alpino annotations. So, I directly extract these annotations and append them to the jsonlines data instances. However, it is noteworthy that for 36 documents, discrepancies arose between the syntactic and coreference data. This discrepancy primarily resulted from instances where the Alpino parser encountered difficulties parsing specific sentences within the document. To address this challenge, I conducted a manual review of these files, and I systematically removed any sentence that did not appear in both datasets.

This project is executed on a 25G Quadro RTX 6000 GPU. However, the GPU encounters memory constraints when processing the longest documents in the corpus. To address this issue, all documents exceeding 3,500 tokens are partitioned into files of uniform sizes, each containing fewer than 3,500 tokens. In total, 54 out of 861 documents require partitioning, with 43 documents divided into two segments, seven documents divided into three segments, and four documents divided into four or more segments.

For some mention-antecedent pairs, the partitioning of documents results in a division of the pair over two separate documents. This likely affects the LEA performance scores, as some of the clusters have changed. Moreover, some mentions may become singletons, if they are the only mention left of their cluster in a document. As described in Section 3.1, all singletons are removed from the corpus. Consequently, the now singleton cluster is removed from the data, and thus no longer affects the performance scores. But, as the data is partitioned in exactly the same way across all the models considered within this study, all models are affected by the partitioning equally. This means that, although the exact LEA-scores are likely somewhat different compared to the same analysis without data-partitioning, the observed trends between models are likely to be the same.

3.3 Data transformation

In order to conduct the experiments, I generate several versions of the dataset. In this section, I explain the various transformed versions, outline the algorithm used to generate them, and then assess the quality of the transformed data. Table 3.3 summarises the various transformed versions of the data, and examples from each transformed set can be found in Table 3.4.

In pursuit of addressing *SQ1* I create *pronoun-specific* versions of the corpus, wherein all instances of third-person pronouns are systematically replaced with pronouns of specific types. Four distinct datasets are thereby created, featuring the following pronouns:

1. with *hij/hem/zijn* pronouns,
2. with *zij/haar/haar* pronouns,
3. with *hen/hen/hun* pronouns,
4. with *die/hen/diens* pronouns.

The creation of these datasets enables a direct evaluation of the model's performance

Version	**Pronouns**	**Split**	**Sub-RQ**
Pronoun-specific test set	Per version, all third-person pronouns are substituted by a particular pronoun, namely, *hij*, *zij*, *hen* or *die*	Test	*SQ1*
Gender-neutral training set	All third-person pronouns are substituted by either *hen* (in 50% of the documents) or *die* pronouns (in the remaining 50%)	Train/dev	*SQ2*
Delexicalised training set	All pronouns are replaced by their POS-tag	Train/dev	*SQ3*
Unseen test set	All pronouns are replaced by previously unseen neopronouns	Test	*SQ4* & *SQ5*

Table 3.3: An overview of the various transformed versions of the SoNaR-1 corpus.

Dataset	Sentence
Original	**Hij** stierf toen **Ensor** 27 jaar was en op het toppunt van **zijn** creatieve periode. ***He** died when **Ensor** was 27 years old and at the peak of **his** creative period.*
Pronoun-specific	1. **Hij** stierf toen **ANON_1** 27 jaar was en op het toppunt van **zijn** creatieve periode. 2. **Zij** stierf toen **ANON_1** 27 jaar was en op het toppunt van **haar** creatieve periode. 3. **Hen** stierf toen **ANON_1** 27 jaar was en op het toppunt van **hun** creatieve periode. 4. **Die** stierf toen **ANON_1** 27 jaar was en op het toppunt van **diens** creatieve periode
Gender-neutral	**Hen** stierf toen **ANON_1** 27 jaar was en op het toppunt van **hun** creatieve periode.*
Delexicalised	<**SUBJ**> stierf toen **ANON_1** 27 jaar was en op het toppunt van <**POSS**> creatieve periode.
Unseen	**Nij** stierf toen **ANON_1** 27 jaar was en op het toppunt van **vijns** creatieve periode.
Original	Na **zijn** herstel vindt **hij zijn vrouw** en **zijn moeder** terug in Folkestone. *After **his** recovery **he** finds **his wife** and **his mother** back in Folkestone.*
Pronoun-specific	1. Na **zijn** herstel vindt **hij zijn persoon** en **zijn ouder** terug in Folkestone. *After **his** recovery **he** finds **his person** and **his parent** back in Folkestone.* 2. Na **haar** herstel vindt **zij haar persoon** en **haar ouder** terug in Folkestone. 3. Na **hun** herstel vindt **hen hun persoon** en **hun ouder** terug in Folkestone. 4. Na **diens** herstel vindt **die diens persoon** en **diens ouder** terug in Folkestone.
Gender-neutral	Na **diens** herstel vindt **die diens persoon** en **diens ouder** terug in Folkestone.*
Delexicalised	Na <**POSS**> herstel vindt <**SUBJ**> <**POSS**> **persoon** en <**POSS**> **ouder** terug in Folkestone.
Unseen	Na **Dijr** herstel vindt **vij vijns persoon** en **vijns ouder** terug in Folkestone.

Table 3.4: Examples of two sentences in the SoNaR-1 dataset, before and after transforming them into different settings. Words that are changed between the versions are marked in bold.
* indicates that in the *gender-neutral* dataset, the usage of *hen* and *die* pronouns is alternated between documents.

on each of the *pronoun-specific* test set, thereby allowing for direct comparisons between the four pronouns.

Continuing, for *SQ2* a *gender-neutral* training set is constructed, in which I include both types of gender-neutral pronouns. I use this dataset to debias the wl-coref model through the application of Counterfactual Data Augmentation (CDA). In this procedure, the model is trained on a *gender-neutral* training set, in order to familiarise it with gender-neutral pronouns. Because there are two gender-neutral pronouns of interest, I decide to adopt an alternation strategy between documents. Specifically, *die* is employed in 50% of the documents, while *hen* is utilised in the remaining 50%. This way, I aim to ensure equal exposure to both pronouns across the corpus. In the event that this approach proves inadequate for achieving satisfactory debiasing outcomes, an alternative approach involves generating an additional version of the dataset in which each document could be duplicated, incorporating instances for both pronouns to enhance exposure and potentially augment the effectiveness of the debiasing process.

Continuing with *SQ3*, which concerns delexicalisation, I create a *delexicalised* training set. In this version of the data, all pronouns are replaced by a syntactical tag: `<SUBJ>` is employed for subjects, `<OBJ>` for objects and `<POSS>` for possessive pronouns. Here, I made a slight modification from the original tags introduced by Lauscher et al., who use the POS-tag `PRP` for personal pronouns and `PRP$` for possessive pronouns. This adaptation is made to distinguish between various grammatical functions, given that the lexical forms adopted by subjects and objects in the Dutch language differ.

Finally, as I will evaluate the performance of the original and debiased systems on previously unseen pronouns in the experiment that concerns *SQ4*, I create an *unseen* test set. Within this set, all pronouns are systematically substituted by a randomly selected neopronoun p, which has not been previously encountered by the model. The selection is made from a set of six Dutch neopronouns:[3] $p \in \{dee/dem/dijr, dij/dem/dijr, nij/ner/nijr, vij/vijn/vijns, zhij/zhaar/zhaar, zem/zeer/zeer\}$.

In order to create these new versions of the data, I create a rule-based rewriting algorithm based on Zhao et al. (2018). The algorithm consists of three principal steps. Across the different data versions, step 2 and 3 are always performed in the same way, while the pronouns used for replacement in step 1 differ between the versions.

1. **Swapping pronouns:** This step involves the identification of third-person personal and possessive pronouns. Pronouns are recognised by their POS-tag, rather than their lexical form, because several Dutch pronouns have identical lexical forms but distinct grammatical functions, such as possessive or personal object *haar* and third-person singular or plural *zij*. Because the POS-tags include information related to number and grammatical function, recognising pronouns based on their POS-tags allows for distinguishing between these otherwise ambiguous tokens. The pronouns are subsequently rewritten according to the rules stipulated for the targeted dataset version (e.g., replacing *hij* with `<SUBJ>` for the delexicalised training set).
2. **Name anonymisation:** I anonymise names across all data versions, because names often convey gender information in Dutch (e.g., the name *Jan* typically denotes males while *Janneke* is typically used for females). The motivation for this step is three-fold. Firstly, this step helps to obscure the overrepresentation of male entities the corpus (Section 3.1). Second, Zhao et al. show name anonymisation to benefit debiasing. Third, name anonymisation makes sure that the performance on the different *pronoun-specific* test sets can fairly be compared. Consider the following sentence:

   Jan is op vrijdag vrij omdat zij dan voetbalt
   (*Jan is free on Friday because she plays football*)

   Here, the model might fail to link the typically male associated name *Jan* to feminine pronoun *zij*, because it expects this name to refer to a male entity. This would result in a lower performance on feminine pronouns, while this is not due to an inherent difficulty in processing feminine pronouns. Through anonymising names we can be sure that name-based gender associations do not cause confounding effects.
   Following Zhao et al.'s method, I recognise names in the data using named entity annotations, considering all tokens with a `PER` (person) tag. The SoNaR-1 corpus includes manually checked named entity annotations, so I can use these annotations directly. Subsequently, all names are replaced with a standardised tag `ANON_`x, with $x \in \mathbb{N}$, where the same value of x always replaces the same string. For example, the sentence

   Jan Jansen is op vrijdag vrij omdat Jan dan voetbalt
   (*Jan Jansen is free on Friday because Jan plays football*)

[3] Pronouns were extracted from the list on https://nl.pronouns.page/voornaamwoorden

becomes

ANON_0 ANON_1 *is op vrijdag vrij omdat* ANON_0 *dan voetbalt*
(ANON_0 ANON_1 *is free on Friday because* ANON_0 *plays football.*)

This step may introduce some processing difficulty to the wl-coref model, because the model is not trained on data with anonymised names. But, as this step is performed across all data versions, it will affect all test sets in the same way.

3. **Replacing gendered nouns:** This step includes replacing gendered nouns by gender-neutral nouns. The motivation for including this step is similar as that for step 2: (1) it hides the overrepresentation of male entities, (2) Zhao et al. show this step to benefit debiasing and (3) it avoids confounding gender effects in case the gender associations of a noun and pronoun do not match.
In order to perform this step, I create a list of gendered nouns, such as e.g. *moeder* (*mother*), and gender-neutral replacements of these nouns, e.g. *ouder* (*parent*). I use this list to replace all the gendered words in the corpus by their gender-neutral counterpart, in order to remove gender clues. The full list of noun rewriting rules can be found in Appendix B. The frequency distribution of the originally gendered nouns in the SoNaR-1 corpus is presented in Figure 3.5. Out of these nouns, 59.0% were originally categorised as male and 41.0% were classified as female. To create this Dutch list, I used the English list by Zhao et al. (2018) as a basis. I iteratively reviewed this list on five occasions, with each iteration incorporating new words and their respective replacements. To ensure the quality of translations, a panel of six individuals participated in reviewing this list and contributed their suggestions. Among these individuals, three use *she/her* pronouns, two use *he/him* pronouns, and one uses *he/they* pronouns. The recruitment of these individuals was facilitated through my personal network.

The applied rewriting algorithm is subject to certain limitations. In some cases, the task of identifying gender-neutral translations for nouns proved challenging, as numerous Dutch words exclusively have gendered forms. For instance, the Dutch term *nicht* (*niece*) lacks a gender-neutral alternative akin to *cousin*. In such instances, we settled for a gender-neutral hypernym of the term of interest, such as *familielid* (*family member*). While retaining some resemblance of the gendered word's meaning, these alternatives inherently sacrifice a substantial portion of the meaning of the original term. Such replacement instances are marked in the rewriting list in Appendix B.

In other cases, the original gendered terms exhibited multiple meanings, necessitating the selection of a singular interpretation for the gender-neutral replacement. Per illustration, the term *vrouw* means both *woman* and *wife*. We decided to use the word *persoon* (*person*) as its gender-neutral translation, rather than the word *echtgenoot* (*spouse*), because we expected the meaning *woman* to be more prevalent. However, this decision introduces potential inaccuracies, exemplified by the second sentence in Table 3.4, where *zijn vrouw* (*his wife*) is rewritten as *zijn persoon* (*his person*). But, considering that the word *vrouw* only occurs around 150 times in a 1 million token dataset, the effect of this limitation on the model's performance will be limited.

Finally, Dutch has two determiners, *de* and *het*. In certain instances, modifying a

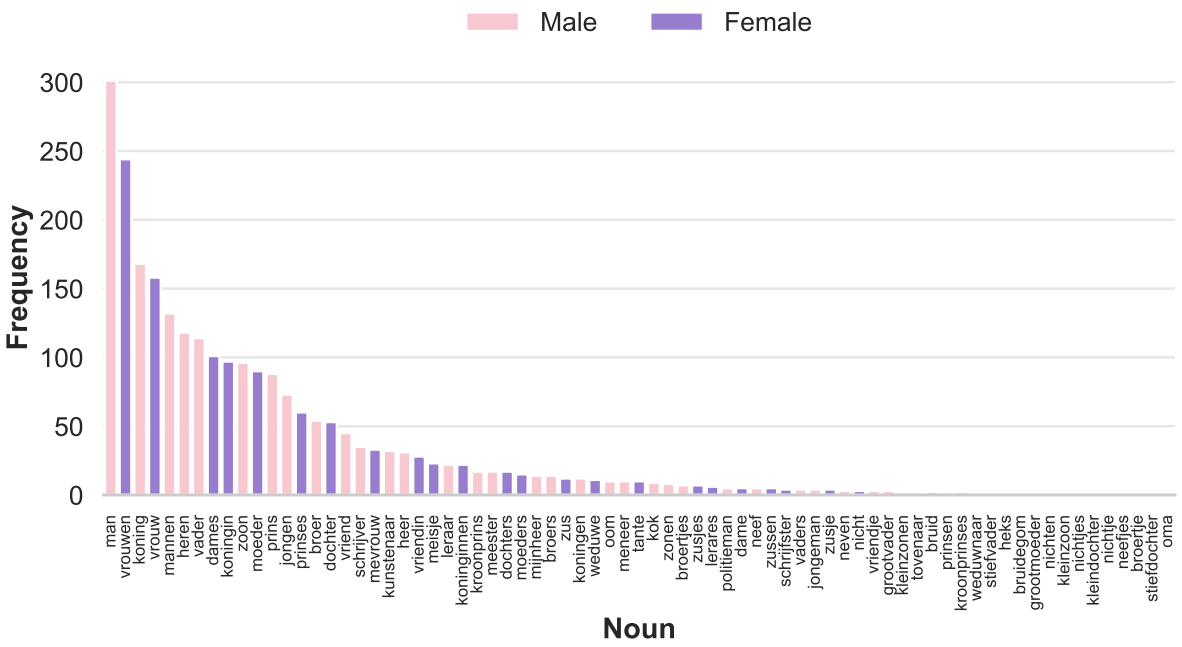


Figure 3.5: Frequencies of gendered nouns in the SoNaR-1 corpus, that are rewritten as gender-neutral nouns in the transformed version of the data. The full list of rewriting rules can be found in Appendix B.

noun may necessitate a corresponding change in the determiner, such as for *de dochter* (*the daughter*) → *het kind* (*the child*). However, due to time constraints, extending the algorithm to accommodate determiner changes was unfeasible. Consequently, certain nouns exhibit an incongruent determiner usage, e.g. *het dochter*. However, I do not expect this flaw to have severe consequences on the modelling performance, because I expect that a coreference resolution system can aptly discern that *het* and *dochter* belong to the same span, given that Dutch only has two determiners. To empirically assess this assumption, I perform an evaluation of the model's performance on the SoNaR-1 data, subjected to noun rewriting only (Section 5.2). I indeed observe only a marginal negative effect originating from this rewriting step.

To more comprehensively evaluate the quality of the transformed data, I manually review a subset of transformed documents comprising 12,584 tokens, and keep track of any mistakes. This subset is balanced over the different data transformations and data splits. A list of the documents included in this subset can be found in Appendix C. The identified errors are systematically documented and summarised in Table 3.5. Notably, a mere total of seventeen mistakes is discerned, while the subset contains 1,111 replacements. This gives an error rate of $\frac{17}{1111} \cdot 100\% = 1.53\%$. I therefore consider the transformed data to be of good quality.

Mistake	**Count**	**Example**
The wrong determiner is used after rewriting a gendered noun	3	de dochter → de kind (instead of *het kind*) *the daughter → the child*
Gender-neutral replacement of a gendered noun does not make sense	6	dames en heren → personen en personen *ladies and gentlemen → persons and persons*
A name was not anonymised, because the name is part of e.g. an organisation, and therefore has a different named entity tag than `PER`	3	stichter van de religieuze orden de "Kinderen van Marie" en de "Personen van Maria" *founder of the religious orders the "Children of Mary" and the "Persons of Mary"*
After anonymisation it is no longer clear that a name is in a possessive form, because the original [NAME]+s is rewritten as `ANON_x`	3	Darwins evolutietheorie → `ANON_12` evolutietheorie *Darwin's evolution theory →* `ANON_12` *evolution theory*
A gendered noun is not rewritten, because it was not on the rewriting list	2	de Nederlandse; echtgenote *the female Dutch person; female spouse*
Total	17	

Table 3.5: Overview of the identified mistakes a 12k-token subset of the transformed data, which I manually assess to evaluate the quality of the data transformation algorithm. Only seventeen errors are identified. Six errors pertain to gendered nouns that underwent rewriting, resulting in a partial loss of meaning, as exemplified by the aforementioned *vrouw* → *persoon* example. Two gendered nouns were not on the gendered nouns list and were thus not rewritten. Three names were not anonymised, because they were part of the name of an organisation, and therefore did not have a `PER` named entity tag.

4 Model

As described in Section 2.3.2, three Dutch coreference resolution systems currently exist: the neural e2e-Dutch, rule-based dutchcoref and the hybrid model by van Cranenburgh et al. (2021). Prior work on debiasing coreference resolution systems only debias neural systems, and leave rule-based systems out of their considerations (e.g. Zhao et al., 2018). This makes sense, given the inherent differences with rule-based systems, which would accordingly require different types of debiasing techniques. In alignment with prior studies I also decide to not include the rule-based and hybrid model in my experiments, but instead to focus solely on a neural model. While the e2e-Dutch model would have been the logical choice here, I encountered installation challenges during its setup. Despite reaching out to the creators for assistance, the issues persisted. Therefore, I decided to develop a new neural Dutch model, by training an existing English model on Dutch data.

I decide to select the wl-coref (Dobrovolskii, 2021) model and train it on Dutch data. Originally trained on English data, I choose this model because (a) it achieves a competitive performance, obtaining a CoNLL F1-score of 80.75 on the OntoNotes corpus, (b) its base models are available in Dutch and (c) the code was well-documented and easy to adapt.

In this chapter, the wl-coref model is further explained. I first give an account of its architecture in Section 4.1. What follows is a description of how this model is trained on Dutch data, in Section 4.2. In this section, I additionally perform an hyperparameter search and compare the model's performance to that of other Dutch models.

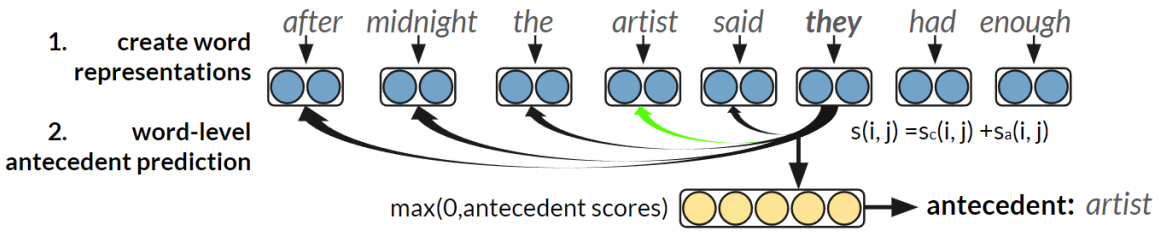


Figure 4.1: Step 1 and 2 of the wl-coref architecture. In Step 1, word representations are created. In Step 2, antecedent predictions are made for each individual word, employing a dual-step approach comprising a coarse step (s_c) and a fine step (s_a). During antecedent prediction, the model is trained to only identify coreference links between the *heads* of mention spans. Per illustration, in the current example, the complete mention span of the antecedent of *they* is *the artist*; however, the model is trained to identify its head, *artist*, as the antecedent.

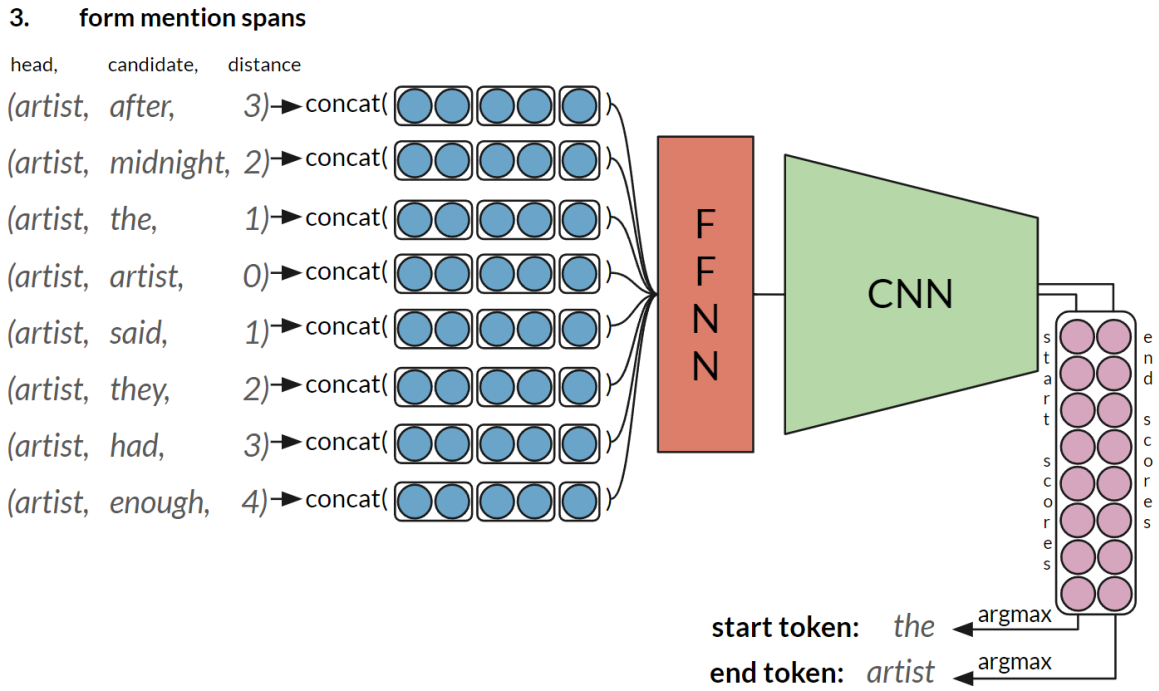


Figure 4.2: Step 3 of the wl-coref architecture involves predicting span boundaries, for all head words that are identified as antecedents or anaphors during Step 2. This prediction is performed through the combination of a feed-forward neural network and a convolutional neutral network. The inputs are a concatenation of (a) the head word representation, (b) the candidate boundary word representation and (c) the distance between the two words. The model produces outputs comprising both start and end scores for each candidate word. Subsequently, the determination of span boundaries is achieved by selecting the argmax over the generated start and end scores.

4.1 Architecture

I use the architecture of the wl-coref model (Dobrovolskii, 2021), which was originally trained on English data, to create a Dutch end-to-end coreference model. This architecture consecutively performs the three following steps, as illustrated by Figures 4.1 and 4.2:

1. Creating word representations.
2. Predicting the antecedent for each word individually, or predicting that the word does not have an antecedent. Notably, during this phase, the model exclusively focuses on identifying antecedents for the *heads* of mention spans. The span's head is defined as the only word in the span with a head outside of the span, or as the root of the sentence. For example, as illustrated in Figure 4.1, the head of the mention *the artist* is *artist*, and consequently, the model learns to predict *artist* as the antecedent for *they* in this step.
3. Predict the full mention span boundaries from the mention heads, ultimately culminating in the final coreference predictions for complete mentions.

This model diverges from many other coreference models that initially form spans and

subsequently make antecedent predictions at a span-level (Lee et al., 2017, 2018; Joshi et al., 2020; Xu and Choi, 2020). I will now explain each of these steps in more detail.

In order to create word representations, the first step is to extract the contextual representations of all the subtokens in the document from a base model, such as BERT (Devlin et al., 2019) or Longformer (Beltagy et al., 2020). Next, the representation of a word is obtained by taking the weighted sum of its subtoken representations. More precisely, the word representations T are formed by:

$$T = W_t \cdot X$$

where X are the subtoken representations and the weights W_t are acquired by taking the softmax over the subtoken representations.

Continuing, antecedent prediction is performed. Particularly, the links between the *head* of a mention and the *head* of its antecedent mention are identified during this step. From these heads, the full mention spans are formed in step 3. Antecedent prediction is performed through the combination of a coarse and a fine step. Firstly, a coarse bilinear scoring function s_c is computed. This coarse function is more efficient to calculate, but less precise than the fine function, and is therefore applied to quickly get a broad sense of which words j are likely candidate antecedents for a target word i. Moreover, this function is used to narrow down the number of possible antecedents to k. The standard value for k is $k = 50$. This function s_c looks as follows:

$$s_c(i, j) = T_i \cdot W_c \cdot T_j^\top$$

where W_c is a learnable matrix of weights, matrix T_i contains the word representations of target words i and matrix T_j contains the representations of the antecedent candidates j. Next, the output matrix of this operation is pruned to only keep the best k candidates for each target word i.

Next, the fine score is computed for the top k candidates by using a feed-forward neural network over an $n \times k$ matrix that contains information about a target-candidate pair (i, j). This matrix contains a concatenation of:

1. The word representations T_i and T_j.
2. Their element-wise product $T_i \odot T_j$.
3. A feature embedding ϕ that contains information about (a) the distance between words i and j in the text, (b) the genre of the document, and (c) whether or not the words i and j were uttered by the same speaker.

Taken together, the fine score s_a looks as follows:

$$s_a(i, j) = FFNN_a([T_i, T_j, T_i \odot T_j, \phi])$$

After performing the two steps consecutively, the fine and coarse scores are combined to compute the final coreference score $s(i, j)$ for a target-candidate pair:

$$s(i, j) = s_c(i, j) + s_a(i, j)$$

This module applies negative log marginal likelihood, with binary cross-entropy as an additional regularization factor. This is computed as follows:

$$L_{COREF} = L_{NLML} + \alpha L_{BCE}$$

with $\alpha = 0.5$. The predicted antecedent for a target word i is the candidate antecedent j with the highest positive score $s(i, j)$. In case no candidate has a positive score, the target is predicted not to have an antecedent.

As the last step, the mention span boundaries are predicted for the words identified as mention heads in step 2. A spans is formed by predicting its start and end words, which are required to be within the same sentence as the head. For this step, a combination of a feed-forward neural network and a convolutional block are used, applying cross-entropy loss. The model's inputs consist of a concatenation of (1) the head word, (2) a candidate boundary word and (3) the distance between them. The full set of candidate words comprises all words in the sentence of the head. The convolutional module has a kernel size of three, and two output channels: one for the start scores and one for the end scores. Concretely, the model thus assigns a score to each candidate word, representing its likelihood of being the start and end token of the span. The precise span boundaries are subsequently obtained by taking the argmax over the start and end scores attributed to all candidate words.

The antecedent prediction module and the span prediction module are trained jointly by taking the sum of their losses.

4.2 Training

I use the wl-coref architecture to train a Dutch model, without making any changes to the core modules. Dobrovolskii (2021) train their models with the *large* versions of their base models. Unfortunately, the memory requirements of these models exceed the memory limits of the 25G GPU that I use. For this reason I am not able to reproduce the original models of Dobrovolskii (2021) nor can I evaluate my installation by checking whether I achieve the same results as in their study. Instead, I directly train the Dutch models, for which I use smaller base models that adhere to the memory constraints. Following Dobrovolskii (2021), I train the models for twenty epochs and report the performance of the epoch that performs best on the development data. Although I do not change the main architecture, I do make three small adjustments to the setup:

1. I change the evaluation metric from the CoNLL score to the LEA score, as the CoNLL metric has been demonstrated to be flawed (see section 2.3.3).
2. While wl-coref uses speaker and genre information in the fine antecedent score, the SoNaR-1 corpus does not contain speaker information. Therefore, I use the same speaker value for all instances, using a value of zero. I also experiment with taking out the speaker component entirely, but this does not improve the performance.
3. While genre information is available for SoNaR-1, I follow Poot and van Cranenburgh (2020) in always using the same genre value. I thus leave exploring the effect of including genre information to future work.

Base model	Dev F1-score
robBERT (Delobelle et al., 2020)	45.5
mBERT-base (Devlin et al., 2019)	47.0
XLM-RoBERTa-base (Conneau et al., 2020)	**52.4**

Table 4.1: Coreference resolution performances of the wl-coref model on the development set of the SoNaR-1 corpus using three different base models. Models were trained for twenty epochs on the SoNaR-1 train set, using the same hyperparameters as Dobrovolskii (2021). Scores are in terms of the LEA metirc.

Learning rate	Best epoch	Precision	Recall	F1
8e-4	17	52.9	56.4	54.6
6e-4	20	50.6	**59.1**	54.5
5e-4	18	51.2	58.6	**54.7**
4e-4	15	52.8	56.0	54.3
3e-4 (original value)	18	52.6	54.7	53.6
1e-4	19	54.3	49.8	52.0
5e-5	19	58.2	37.7	45.8
4e-5	19	58.4	35.1	43.8
3e-5	19	60.1	24.8	35.1
2e-5	20	60.0	18.7	28.5
1e-5	18	**63.9**	5.7	10.5
5e-6	18	52.8	9.9	16.7

Table 4.2: Learning rate tuning of the wl-coref model, using XLM-RoBERTa (Conneau et al., 2020) as its base model. LEA performance scores on the SoNaR-1 development set. Models were trained for twenty epochs, keeping all other hyperparameters the same as Dobrovolskii (2021).

I report the span-level performance scores in terms of the LEA metric. I prioritise recall over precision, because, for non-binary individuals, false negatives can be more detrimental than false positives in coreference resolution, as false negatives can lead to erasure (see section 2.2.3).

I use the Hugging Face Transformers (Wolf et al., 2020) implementations to compare three base models: the Dutch RoBERTa-based (Liu et al., 2019) robBERT model (Delobelle et al., 2020), and the multilingual mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) models, both in their *base* versions. I also tried to use the BERT-based (Devlin et al., 2019) BERTje model (de Vries et al., 2019), which serves as the base model for e2e-Dutch, but I did not manage to get this model to work with wl-coref. Table 4.1 shows the performances using the three base models on the SoNaR-1 dev set, using equivalent hyperparameters to Dobrovolskii (2021). As XLM-RoBERTa obtains the best performance (F1-score = 52.4), I continue to use this base model.

Furthermore, I execute a hyperparameter search for two hyperparameters:

- The learning rate (original value = $3e-4$).
- The BERT learning rate (original value = $1e-5$).

keeping all other settings and hyperparameters the same. I also experiment with lowering k and using a different learning rate schedule than the standard linear one, but these adaptations do not increase the performance.

BERT learning rate	Best epoch	Precision	Recall	sl F1
1e-4	17	50.2	58.9	54.2
1e-5 (original value)	18	51.2	58.6	54.7
2e-5	15	53.2	57.3	55.2
3e-5	15	52.0	**59.5**	**55.5**
4e-5	15	52.0	57.9	54.8
5e-5	16	**54.1**	56.6	55.3
5e-6	18	51.2	56.6	53.8

Table 4.3: BERT learning rate tuning of the wl-coref model, using XLM-RoBERTa (Conneau et al., 2020) as its base model. LEA performance scores on the SoNaR-1 development set. Models were trained for twenty epochs, using a learning rate $= 5e-4$, and keeping all other hyperparameters the same as Dobrovolskii (2021).

	Data split	Precision	Recall	F1
Mean	Development	53.04	57.87	55.25
Standard deviation		2.30	3.08	0.16
Mean	Test	55.48	55.83	55.57
Standard deviation		2.33	2.69	0.46

Table 4.4: Average LEA performance scores on the SoNaR-1 development and test set, using five random seeds. Models were trained for twenty epochs, using a learning rate $= 5e-4$, BERT learning rate $= 3e-5$ and keeping all other hyperparameters the same as Dobrovolskii (2021).

I experiment with twelve different values for the learning rate. The outcomes are presented in Table 4.2. The best F1-scores are observed for a learning rate of $5e-4$ (F1-score $= 54.7$), which I subsequently continue to use. Following this, I optimise the BERT learning rate, considering seven different values. These results are outlined in Table 4.3. The best recall and F1-scores are obtained using a value of $3e-5$ (F1-score $= 55.5$), and I therefore continue to use this value.

Finally, I report the development and test performance scores in Table 4.4, as the average of five random seeds. The model obtains a test F1-score of 55.57. The precision and recall scores exhibit considerable variability, specifically $\sigma = 2.33$ for test precision and $\sigma = 2.69$ for test recall. The standard deviation for the test F1-score is notably lower ($\sigma = 0.46$). This difference can be explained by the precision/recall trade off. Throughout the rest of this study, I report the main results as the average of five random seeds.

In Table 4.5, the performance of the wl-coref model is compared to that of other Dutch coreference resolution models, particularly e2e-Dutch and dutchcoref. e2e-Dutch achieves a test F1-score of 61.6, meaning that wl-coref scores 6.03 points lower. Despite the fact that the performance of the wl-coref model is lower than that of e2e-Dutch, it is noteworthy that the gap in precision score (-7.0), is larger than the recall gap (-4.9), which is considered as more important for this study. Continuing, an improvement over e2e-Dutch is that the difference between the development and test set performance is smaller for wl-coref (F1 $\Delta = -3.7$ for e2e-Dutch and F1 $\Delta = +0.3$ for wl-coref). Wl-coref even has a slightly better performance on the test data than on the development

Model	Data	Precision	Recall	F1
dutchcoref	SoNaR-1 development	52.2	38.0	44.0
e2e-Dutch	SoNaR-1 development	65.6	65.0	65.3
wl-coref	SoNaR-1 development	53.0	57.9	55.3
dutchcoref	SoNaR-1 test	52.6	37.9	44.0
e2e-Dutch	SoNaR-1 test	62.5	60.7	61.6
wl-coref	SoNaR-1 test	55.5	55.8	55.6

Table 4.5: Comparison between three Dutch coreference resolution models on the SoNaR-1 development and test set: rule-based dutchcoref, neural e2e-Dutch and neural wl-coref. Performance scores of dutchcoref and e2e-Dutch are as reported by Poot and van Cranenburgh (2020). Scores are reported in terms of the LEA metric. Performance scores of the wl-coref model are the mean of 5 random seeds.

data, suggesting minimal overfitting. Finally, the wl-coref model notably outperforms the rule-based dutchcoref model (+11.6 in test F1-score).

5 Experiments

In this chapter, I give a description of the experiments. Table 5.1 summarises the experimental setup. Prior to discussing the main experiments, I introduce a novel evaluation metric, called a *pronoun score*, in Section 5.1, which is used as a main evaluation metric throughout the experiments, in addition to LEA. Next, I continue to describe the experiments. For each experiment, I begin with a description of the method, followed by a reporting of the results. In Section 5.2 I describe the gender-neutral pronoun evaluation experiment, in which I compare between the wl-coref model's performance on gendered pronouns and gender-neutral pronouns. Subsequently, I discuss the debiasing experiment in Section 5.3. In this experiment, I evaluate the effectiveness of two debiasing techniques: Counterfactual Data Augmentation and delexicalisation. Next, I perform the unseen pronouns experiment in Section 5.4, in which I evaluate the performance of the original wl-coref model and the debiased models on previously unseen pronouns. Finally, I create a test suite for pronoun-related behaviour in Section 5.5, which I use to conduct a more in-depth evaluation of the original and a debiased model's abilities to handle gender-neutral pronouns.

5.1 Pronoun score

Because standard performance measures such as precision and recall reflect the overall performance on *all* clusters in the corpus, they do not directly reflect the model's ability to resolve pronouns. However, because the *pronoun-specific* test sets only differ in third-person pronouns, any difference in performance between these sets can be attributed to the predictions for third-person pronouns. Therefore, I compare the LEA scores for the different *pronoun-specific* test sets in order to answer the research question.

However, if the LEA F1-score is e.g. one point lower for the gender-neutral *hen* pronoun set than for the masculine *hij* pronoun set, it is not directly clear how many more *hen* pronouns are incorrectly resolved than *hij* pronouns (see Section 2.3.3 for an explanation of the LEA metric). To get a more direct insight into the model's ability to process

	Gender-neutral pronoun evaluation experiment	**Debiasing experiment (CDA)**	**Debiasing experiment (delexicalisation)**	**Unseen pronouns experiment**
Sub-RQ	*SQ1*	*SQ2*	*SQ3*	*SQ4* & *SQ5*
Purpose	Evaluation	Debiasing	Debiasing	Evaluation
Debiasing method	-	CDA	Delexicalisation	-
Training data	Regular SoNaR-1 training set	*Gender-neutral* training set	*Delexicalised* training set	-
Evaluation data	*Pronoun-specific* test set	*Pronoun-specific* test set	*Pronoun-specific* test set	*Unseen* test set

Table 5.1: Experimental setup overview.

[Raven] entered the kitchen. "Did [you] sleep well?", [they] asked [[their] roommate] "No [Raven]", said [Thorn] annoyed, "[Tobi] called me way too early"

(a)

[Raven] entered the kitchen. "Did [you] sleep well?", they asked [[their] roommate] "No [Raven]", said [Thorn] annoyed, "[Tobi] called me way too early"

(b)

Figure 5.1: An example sentence, with its gold annotations in (a), and example predictions in (b). Mentions are indicated with brackets. Mentions with the same colour belong to the same cluster.

pronouns in particular, I introduce a **pronoun score**, which I use as complementary to the LEA score, that computes *the percentage of third-person pronouns for which at least one correct antecedent is identified*. I compute this score for each of the four pronouns, using the *pronoun-specific* test sets. I will now illustrate how this metric is computed and will subsequently motivate my choice of computing the pronoun score based on its antecedents.

Mathematically, this score is computed as follows:

$$pronoun\_score = \frac{\sum_{p \in pronouns}[(gold\_ants(p) \cap predicted\_ants(p) > 1]}{|pronouns|} \cdot 100\%$$

To illustrate the usage of this score, let us consider the example in Figure 5.1, for which the gold annotations can be found in Figure 5.1a and an example prediction is presented in 5.1b. The example predictions are correct, except for the fact that the mention `[they]` is not recognised as a mention, and is therefore not considered as part of the *Raven* cluster. In the example, we can identify the following gold and predicted antecedents for the two third-person pronouns *they* and *their*:

$$\begin{array}{ll} gold\_ants(they) & = \{Raven\} \\ gold\_ants(their) & = \{they, Raven\} \\ predicted\_ants(they) & = \{\} \\ predicted\_ants(their) & = \{Raven\} \end{array}$$

Then the pronoun score is computed as follows:

$$\begin{array}{lll} gold\_ants(they) \cap predicted\_ants(they) > 1 & = \{Raven\} \cap \{\} > 1 & = 0 > 1 = 0 \\ gold\_ants(their) \cap predicted\_ants(their) > 1 & = \{they, Raven\} \cap \{Raven\} > 1 & = 1 > 1 = 1 \end{array}$$

$$pronoun\_score = \frac{0+1}{2} = \frac{1}{2}$$

The reason for computing the pronoun score by considering the pronoun's antecedents is because the wl-coref model, like most coreference resolution models, directly predicts the antecedent for each word. It then continues to predict clusters by combining all words that share antecedent links into a cluster. By evaluating pronoun antecedent I thus directly evaluate whether the model makes correct predictions for third-person pronouns.

A potential objection against a metric that considers antecedents is that the first mention of a cluster is always excluded from the evaluation, because the first mention of a cluster does not *have* an antecendent. However, pronouns will hardly ever be used as the first mention of a cluster, because they are typically used to replace names or proper nouns, which have been introduced earlier. Therefore this objection does not appear to be relevant for the evaluation of pronouns.

Additionally, I decide to consider all the pronoun antecedents in the predicted cluster in the evaluation, rather than only the one antecedent that is *directly* predicted by the model. The reason for this is that earlier studies show very poor performances on gender-neutral pronouns (Baumler and Rudinger, 2022; Cao and Daumé III, 2021), indicating that task of correctly processing these pronouns is difficult. Therefore, I prefer a more lenient configuration of the evaluation metric, which considers at least one correct antecedent in the prediction cluster to be sufficient. However, this metric is highly adaptable, and can easily be made more strict to fit more straightforward tasks.

Finally, I make the decision to consider one correct pronoun antecedent to be sufficient to consider this pronoun as correctly resolved. I prefer this option over an alternative such as requiring all of the pronoun's antecedents to be correct. The reason for this is that otherwise, the model might be punished double for a single mistake, as illustrated by the example in Figure 5.1. In the example predictions, the third-person pronoun *they* is missed as a mention. Therefore, the *their* also misses one of its correct antecedents in the predictions, as *they* is an antecedent of *their*. This means that requiring all the correct antecedent to be found, results in a pronoun score of zero for this example, despite the fact that the pronoun *their* is predicted to be part of the correct cluster. I consider this outcome undesirable. The LEA score already provides a holistic view of the model's performance by evaluating the full cluster predictions, and this way the pronoun score can complement the LEA score by zooming in on the pronoun alone, without considering the quality of the rest of the cluster.

5.2 Gender-neutral pronoun evaluation experiment

In this section, I first describe the setup of the gender-neutral pronoun evaluation experiment in Section 5.2.1. Next, discuss the results and relate them back to the research question in Section 5.2.2.

5.2.1 Setup

In this experiment, I aim to answer *SQ1*: *How good is an existing Dutch coreference resolution system at processing gender-neutral pronouns compared to gendered pronouns?* To answer this question, I create four *pronoun-specific* versions of the test set (see section 3.3), which contain the regular SoNaR-1 test set data, but with all third-person pronouns replaced by either *hij, zij, hen* or *die* pronouns. I then answer the research question by comparing the average performance on the gendered *pronoun-specific* test sets (*hij* and *zij*), with the average performance on the gender-neutral *pronoun-specific* test sets. My

	Precision	Recall	F1	Δ F1 regular data
Regular data	55.48 (σ=2.33)	55.83 (σ=2.69)	55.57 (σ=0.46)	-
Anonymisation only	53.94 (σ=2.44)	49.77 (σ=2.51)	51.68 (σ=0.25)	-3.89
Noun rewriting only	54.99 (σ=2.40)	55.25 (σ=2.63)	55.03 (σ=0.40)	-0.54
Anonymisation + noun rewriting	53.58 (σ=2.24)	49.59 (σ=2.62)	51.41 (σ=0.44)	-4.16

Table 5.2: Model performance scores on the test set, before and after transforming the data to obscure gender clues. The full data transformations are described in Section 3.3. Reported scores are the average of five random seeds.

Data	Precision	Recall	F1	Δ F1 baseline
Baseline	53.58 (σ=2.24)	49.59 (σ=2.62)	51.41 (σ=0.44)	-
Hij pronouns (masculine)	52.23 (σ=2.30)	49.66 (σ=2.76)	51.29 (σ=0.42)	-0.12
Zij pronouns (feminine)	53.18 (σ=2.33)	48.73 (σ=2.56)	50.77 (σ=0.37)	-0.64
Hen pronouns (gender-neutral)	53.29 (σ=2.56)	45.82 (σ=3.16)	49.14 (σ=0.68)	-2.27
Die pronouns (gender-neutral)	52.55 (σ=1.46)	44.94 (σ=2.24)	48.36 (σ=0.44)	-3.05

Table 5.3: Model performance on the *pronoun-specific* test sets. The reported scores are the average of five random seeds. The reported *baseline* performance refers to the model performance on the version of the test set in which gender clues are removed, but the pronouns remain unchanged, as reported in Table 5.2.

hypothesis is that the existing Dutch coreference resolution system performs worse on gender-neutral pronouns than on gendered pronouns.

Besides changing pronouns, the data transformation that is performed in order to create the *pronoun-specific* test sets (see section 3.3) involves obscuring gender clues through (a) replacing gendered nouns by gender-neutral nouns, and (b) by anonymising names. Prior to performing the experiment, I investigate how these transformations, in isolation and combined, affect the model's performance. This serves as a baseline, to be able to isolate the impact of changing the pronouns in the *pronoun-specific* data sets.

I do so by evaluating the model on variations of the test set that include these transformations, but exclude the changing of pronouns. Table 5.2 reports the results. As can be seen here, anonymising names has a negative effect on the performance, as the F1-score drops with 3.89 points. This suggests that the model does rely on name information, and the gender clues they reveal, in making its predictions. Continuing, rewriting gendered nouns also has a small negative effect on the performance (-0.54). The small size of this effect makes sense considering the low frequencies of rewritten nouns in the corpus (see Figure 3.5). Finally, combining this transformation with anonymisation results in the strongest negative effect, reducing the F1-score with 4.16 points, roughly equalling the effect of the isolated transformations combined.

5.2.2 Results

Table 5.3 reports the performance scores on the *pronoun-specific* test sets in terms of the LEA metric, and Table 5.4 reports the pronoun scores. I now discuss the main observations.

Pronouns	Pronoun score	Standard deviation	Δ with *hij*
Hij/hem/zijn (masculine)	88.36%	0.89	-
Zij/haar/haar (feminine)	86.65%	1.23	-1.71
Hen/hen/hun (gender-neutral)	75.85%	2.93	-12.51
Die/hen/diens (gender-neutral)	57.49%	6.55	-30.87

Table 5.4: Pronoun scores on the *pronoun-specific* test sets. Scores are computed as the average of five random seeds.

[ANON_0] praat met [[zijn] grote bruine ogen] .
[ANON_0] talks with [[his] big brown eyes].

(a)

[ANON_0] praat met [diens grote bruine ogen] .
[ANON_0] talks with [their big brown eyes].

(b)

Figure 5.2: Example of a sentence in which the model correctly resolves the masculine pronoun *zijn* (a), but fails to find a correct antecedent for gender-neutral pronoun *diens* (b). English translations are provided in italics.

The best performance is achieved on *hij* pronouns (pronoun score = 88.36%; F1=51.29). This is according to expectations, as masculine pronouns constitute 79,1% of the third-person pronouns in the training data (see Section 3.1).

The performances on *hij* and *zij* pronouns are similar (-0.64 in F1-score and -1.71 percentage points in pronoun score on *zij* compared to *hij* pronouns). This is surprising, because earlier studies found English coreference resolution models to perform better on masculine than on feminine pronouns (Webster et al., 2018; Kurita et al., 2019; Rudinger et al., 2018). A difference between English and Dutch that might play a role here is that in Dutch, the feminine third-person singular pronoun *zij* is also used as a third-person plural pronoun. In the corpus, the type *zij* occurs 250 times as a third-person singular pronoun, but 481 times as a third-person plural pronoun. This might increase the model's familiarity with the type, and thereby boost it's recognition as a pronoun.

The performance scores are lower for the gender-neutral pronouns. The *hen* pronoun set loses 2.27 points in F1-score compared to the *hij* pronouns, and for *die* this gap increases to 3.05 points. Similarly, the pronoun score drops with 12.51 percentage points for the *hen* pronouns and decreases even further with 30.87 percentage points for *die* pronouns. These results indicate that the model performs worse on gender-neutral than on gendered pronouns.

Per illustration, Figure 5.2 provides an example sentence in which the model correctly resolves the masculine pronoun *zijn* in 5.2a, but fails to identify a correct antecedent for the gender-neutral pronouns *diens* in 5.2b. An additional example can be found in Figure 5.3. Here, the model correctly resolves *zij* as referring to *ANON_4*, whereas gender-neutral *hen* is identified to refer to the plural *foundry workers*.

***Hen* pronouns are better resolved than *die* pronouns.** The high standard deviations for *die* additionally indicate an unstable resolution. A potential reason for this

Een punt van [kritiek] is dat [ANON_4] veel vertrouwde op de professionaliteit van de gieters. [Zij] nam wel de leiding maar maakte de mallen niet zelf.
One point of [criticism] is that [ANON_4] relied a lot on the professionalism of the foundry workers. [She] took the lead but did not make the molds themself.

(a)

Een punt van [kritiek] is dat [ANON_4] veel vertrouwde op de professionaliteit van [de gieters]. [Hen] nam wel de leiding maar maakte de mallen niet zelf.
One point of [criticism] is that [ANON_4] relied a lot on the professionalism of the [foundry workers]. [They] took the lead but did not make the molds themself.

(b)

Figure 5.3: Example of a sentence where feminine pronouns *zij* is correctly resolved, but the gender-neutral pronoun *hen* (b) is not. English translations are provided in italics. Note that despite of the fact that the resolution of *hen* (*they*) appears ambiguous in the English translation, this is not the case in Dutch because the verb conjugation used here is only employed for singular subjects

is that *hen* is always used as a personal or possessive pronoun in Dutch (be it as a plural pronoun), whereas this is not the case for *die*. In the data, *hen* has a frequency of 290, invariably occurring as a third-person plural pronoun, functioning as an object. *Die*, on the other hand, occurs 5,268 times as a relative pronoun, 1,995 times as a demonstrative pronoun, and does not appear as a personal pronoun at all. This likely makes it harder for the model to recognise and resolve the usage of *die* as a personal pronoun.

Conclusion. All taken together, this experiment provides two results that contribute to addressing SQ1. These results are outlined as follows:

1. The F1-scores on the *pronoun-specific* test tests for the gender-neutral pronouns are on average 1.92 points lower than on the *pronoun-specific* test sets for the gendered pronouns.
2. The pronoun scores for the gender-neutral pronouns are on average 13.08% lower than for the gendered pronouns.

Based on this evidence I conclude that the wl-coref model performs worse on gender-neutral pronouns than on gendered pronouns.

5.3 Debiasing experiment

In this section I discuss the debiasing experiment. I first describe the setup (Section 5.3.1) and then continue to discuss the results (Section 5.3.2). Finally, I perform an additional experiment with a reduced amount of debiasing documents in Section 5.3.3.

5.3.1 Setup

The debiasing experiment aims to address *SQ2* and *SQ3* : *Can the debiasing method Counterfactual Data Augmentation / delexicalisation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns?* I experiment with two

debiasing methods. The first debiasing method is Counterfactual Data Augmentation (CDA). To apply this method, I train the wl-coref model on the *gender-neutral* version of the data. In this altered dataset, all third-person singular pronouns are substituted by gender-neutral pronouns: *hen* in 50% of the documents, and *die* in the remaining 50% (refer to Section 3.3 for details). The rationale behind this methodology is that inserting gender-neutral pronouns into the training data is expected to improve the model's processing of these pronouns.

The second debiasing method is delexicalisation. This method is applied by training the model on the *delexicalised* version of the data, wherein all third-person pronouns are replaced by their corresponding syntactic tag (see Section 3.3). The fundamental concept behind this methodology is that by systematically removing all lexical variations associated with third-person singular pronouns, the model will develop the capability to identify any token in this grammatical position as a pronoun, irrespective of its lexical form.

I evaluate the efficacy of both debiasing methods in two conditions:

1. Fine-tuning the model from scratch on the respective debiasing dataset.
2. Further fine-tuning the original wl-coref model, initially trained on the regular SoNaR-1 data, with the respective debiasing dataset.

For consistency, the same hyperparameters are employed as for the regular model. Similarly, in accordance with the regular model, debiased models trained from scratch are trained for 20 epochs, while the fine-tuned models are trained for 10 epochs. In both experimental conditions, the weights kept for evaluation are from the epoch that shows the highest performance on the development set aligning with the training data – i.e., the *gender-neutral* and *delexicalised* development sets. The models are subsequently evaluated on the *pronoun-specific* test sets. The debiasing performance is measured as the difference between the average performance on gender-neutral pronouns by the debiased model and the regular wl-coref model, measured through the LEA F1-score and the pronoun score.

In order to evaluate whether any information is lost through debiasing and to assess the broader impact of debiasing on overall model performance, an evaluation of the debiased models is also conducted on the original SoNaR-1 test set.

Given the favorable outcomes demonstrated by CDA in mitigating binary-gender bias within coreference resolution systems (Zhao et al., 2018, 2019), I expect this method will similarly demonstrate to be effective in introducing gender-neutral pronouns to the model. In contrast, the efficacy of delexicalisation has not previously been tested in a similar setup. Lauscher et al. (2022), who introduce this methodology, conducted tests in a reversed configuration, training a model on regular data and evaluating its performance on delexicalised data. In this context, the model did not exhibited a good performance. Additionally, Lauscher et al. conducted an experiment in which both training and testing was performed on delexicalised data, resulting in a satisfactory performance. However, this experimental design does not faithfully simulate a realistic scenario in which a model is debiased through delexicalised data, but subsequently deployed on naturally occurring data, which includes pronouns in their lexical forms. Therefore, empirical inquiry is required to ascertain whether this more realistic setup proves equally effective.

Model	Hij	Zij	Hen	Die
Original model	51.29 (σ=0.42)	50.77 (σ=0.37)	49.14 (σ=0.68)	48.36 (σ=0.44)
Fine-tuning the wl-coref model from scratch				
Delexicalisation	53.04 (σ=0.70)	53.31 (σ=0.53)	50.67 (σ=0.79)	50.69 (σ=0.64)
CDA	54.44 (σ=0.41)	54.47 (σ=0.49)	54.40 (σ=0.33)	54.33 (σ=0.41)
Further fine-tuning the wl-coref model				
Delexicalisation	53.74 (σ=0.78)	53.53 (σ=0.78)	50.51 (σ=1.05)	50.07 (σ=0.90)
CDA	54.57 (σ=0.59)	54.48 (σ=0.63)	54.50 (σ=0.58)	54.36 (σ=0.59)

(a)

Model	Hij	Zij	Hen	Die
Original model	88.36% (σ=0.89)	86.65% (σ=1.23)	75.85% (σ=2.93)	57.49% (σ=6.55)
Fine-tuning the wl-coref model from scratch				
Delexicalisation	76.50% (σ=4.56)	82.79% (σ=2.42)	71.55% (σ=4.94)	61.89% (σ=5.53)
CDA	86.88% (σ=1.64)	89.08% (σ=0.93)	88.02% (σ=0.74)	89.37% (σ=0.57)
Further fine-tuning the wl-coref model				
Delexicalisation	89.29% (σ=1.17)	88.76% (σ=0.98)	72.91% (σ=2.80)	57.17% (σ=1.95)
CDA	90.52% (σ=0.44)	90.60% (σ=0.33)	90.16% (σ=0.51)	89.60% (σ=0.50)

(b)

Table 5.5: LEA F1-scores (a) and pronoun scores (b) on the pronoun-specific test sets after debiasing. The reported scores represent the average across five random seeds.

Given that further fine-tuning is computationally less demanding compared to fine-tuning from scratch,[1] it would be a preferable debiasing approach, provided it achieves a satisfactory performance. However, there potentially exists a trade-off between computational efficiency and debiasing performance. I expect that fine-tuning might not be sufficient to achieve an acceptable debiasing performance with delexicalisation, because the core idea behind this technique involves using a unified representation for *all* pronouns, but the pre-trained wl-coref model will already have acquired distinct representations for gendered pronouns. Consequently, the effectiveness of the syntactic-tag representation may be reduced in this context. In contrast, CDA inserts novel pronouns into the data, which will have their own representation, separate from the representation of familiar pronouns. Therefore, fine-tuning might be less problematic for this method and the heightened exposure to gender-neutral pronouns during further fine-tuning might suffice to enhance this model's performance on these pronouns.

5.3.2 Results

Table 5.5 displays the LEA F1-scores (5.5a) and the pronoun scores (5.5b) for the *pronoun-specific* test sets after (a) fine-tuning from scratch and (b) further fine-tuning the original wl-coref model, using the two debiasing techniques. Here I discuss the main observations.

[1]Here, I do not take the computational costs of fine-tuning the original wl-coref model into account, because I consider this an off-the-shelf model and I want to isolate the costs of debiasing an existing model.

Model	F1 performance regular test set	Δ original model
Original model	55.57 ($\sigma = 0.46$)	-
Delexicalisation full	53.04 ($\sigma = 0.58$)	-2.53
Delexicalisation fine	54.38 ($\sigma = 0.86$)	-1.37
CDA full	54.48 ($\sigma = 0.51$)	-1.27
CDA fine	55.17 ($\sigma = 0.47$)	-0.58

Table 5.6: Average LEA F1-scores achieved by the debiased models on the regular SoNaR-1 test data. This evaluation aims to assess potential losses in abilities through the debiasing process. The reported scores represent the average across five random seeds.

Delexicalisation does not appear to successfully debias the model. Despite some improvement in the F1-scores for the *hen* and *die* test sets after fine-tuning from scratch (+1.55 and +2.33 respectively), the pronoun scores remain low. The pronoun score for *die* improves slightly with 4.4 percentage points, but there is a corresponding decrease for *hen* of 4.3 percentage points. Similarly, further fine-tuning on delexicalised data fails to improve the performance. Instead, the pronoun scores for the gender-neutral pronouns even deteriorate (-4.7 percentage points for *hen* and -11.21 percentage points for *die*). This suggests that the removal of lexical information alone proves insufficient to effectively improve the model's performance on gender-neutral pronouns.

The application of CDA fine-tuning from scratch shows substantial improvements on gender-neutral pronouns. After fine-tuning from scratch, the F1-scores surpass 53.60 for all pronouns, while the best F1-score of the original model, on the *hij*-test set, was only 51.29. Furthermore, the pronoun scores exceed 86% for all pronouns, representing an improvement of 31.88 percentage points for *die* and 12.17 percentage points for *hen*. Noteworthy reductions in standard deviations are also observed for the gender-neutral pronouns (-5.89 for *die* and -2.19 for *hen*).

An additional observation is that both of the fine-tuned from scratch models sustain a high performance for *hij* and *zij*, despite not encountering these pronouns during training. This implies that the base model's pre-training already imparts sufficient familiarity with these pronouns. The performance for *zij* surpasses that of *hij* for both models, possibly due to the continued occurrence of *zij* in the corpus as a third-person plural pronoun, while *hij* ceases to appear altogether, lacking an alternative meaning in Dutch.

Further fine-tuning with CDA results in the best debiasing outcomes. The F1-scores across pronoun-specific test sets surpass 54.0 (an improvement of +6.00 for *die* and +5.35 for *hen*, but also an improvement of +3.28 for *hij* and +3.71 for *zij*). Moreover, all pronoun scores exceed 89.5%, achieving even slightly higher scores than the fine-tuned from scratch CDA model. These results are encouraging, particularly considering that further fine-tuning already was the preferred method, due to its computational efficiency.

Lastly, Table 5.6 shows that CDA through further fine-tuning also obtains the best performance on the original SoNaR-1 test data. I evaluate the impact of debiasing on the performance on the original dataset, in order to investigate whether any knowledge is lost through the debiasing process. While all debiased models show a small performance drop in comparison to the original model, this decline is most pronounced

for the delexicalised models. In contrast, this decrease is smaller for the CDA models, with a decrease smaller than 1 point (-0.58) in the further fine-tuned setting.

Conclusion. In conclusion, delexicalisation does not appear to be a successful debiasing method to learn a coreference resolution model to correctly process gender-neutral pronouns. On the other hand, CDA does show good results, improving the average pronoun scores for gender-neutral pronouns with more than 22 percentage points. Therefore I conclude that the debiasing method CDA can improve the ability of the wl-coref model to process gender-neutral pronouns.

5.3.3 Low-resource debiasing exploration

As a subsequent investigation, I explore whether the best debiasing method can also be effectively applied in a scenario with limited data availability. As the best debiasing method, I select CDA through further fine-tuning, because this method (a) obtains the best results across all categories, (b) diminishes the performance gap between gendered and gender-neutral pronouns to less than 1% and (c) this model employs further fine-tuning instead of fine-tuning from scratch, the more computationally efficient approach.

The reason for this follow-up experiment is that large corpora for debiasing may not always be available. Moreover, debiasing with a smaller corpus reduces the computational costs. Consequently, my objective for this experiment is to explore the effectiveness of debiasing in a low-resource context.

Setup

To investigate this, I implement CDA in the same experimental set-up as above. But, for this exploration, I only employ fractions of the complete *gender-neutral* training set, specifically 10%, 5%, 2.5%, and 1.25%, corresponding to 62, 30, fifteen, and seven documents respectively. It is important to note that in the *gender-neutral* training set, the usage of the pronouns *hen* and *die* alternates between documents, with each pronoun featured in only 50% of the documents. Thus, when debiasing with, for instance, 30 documents, each pronoun is present in only fifteen documents. Furthermore, the documents in the training set exhibit considerable variation in terms of length and the number of included pronouns, as discussed in Section 3.1. In light of these variations, five partitions are employed for each training size fraction, of which the average scores are reported. This is done to ensure, at least to a certain extent, that the results are a representative approximation of the complete corpus characteristics. In the interest of computational efficiency, only one seed is used for this experiment.

Results

Table 5.7 presents the outcomes in terms of F1-score (5.7a) and pronoun score (5.7b) on the pronoun-specific test sets.

Percentage	# Train documents	Hij	Zij	Hen	Die
100%	625	54.64	54.18	54.65	54.53
10%	62	52.88 (σ=0.38)	52.64 (σ=0.40)	52.29 (σ=0.40)	52.13 (σ=0.17)
5%	31	52.74 (σ=0.46)	52.41 (σ=0.37)	52.03 (σ=0.31)	51.87 (σ=0.46)
2.5%	15	51.93 (σ=0.20)	51.80 (σ=0.26)	51.21 (σ=0.33)	50.88 (σ=0.31)
1.25%	7	51.65 (σ=0.25)	51.52 (σ=0.32)	50.90 (σ=0.37)	50.38 (σ=0.34)
Original model	0	51.12	50.56	49.29	49.10

(a)

Percentage	# Train documents	Hij	Zij	Hen	Die
100%	625	90.76%	90.60%	89.94%	89.67%
10%	62	92.41% (σ=0.19)	91.26% (σ=0.41)	88.64% (σ=0.79)	85.42% (σ=0.94)
5%	30	92.02% (σ=0.48)	90.66% (σ=0.43)	87.32% (σ=0.90)	83.65% (σ=1.08)
2.5%	15	91.40% (σ=0.64)	89.96% (σ=0.54)	85.09% (σ=0.89)	79.48% (σ=0.94)
1.25%	7	91.36% (σ=0.62)	90.25% (σ=0.58)	85.12% (σ=1.06)	78.44% (σ=1.81)
Original model	0	88.19%	86.66%	78.79%	65.77%

(b)

Table 5.7: F1-scores (a) and pronoun scores (b) after further fine-tuning the wl-coref model using the debiasing technique CDA with various fractions of the full *gender-neutral* training set. The reported scores represent average across five data partitions.

The pronoun scores improve after debiasing with just a few debiasing documents. With a tiny dataset of seven documents (equivalent to 3-4 documents per pronoun), a substantial improvements of 12.67 percentage points for *die* and 6.33 percentage points for *hen* can be observed. Furthermore, by using 5% of the documents, i.e. fifteen debiasing documents per pronoun, the pronoun scores already surpass 80% for the gender-neutral pronouns. This is in line with Björklund and Devinney (2023), who observe that including gender-neutral pronouns in 2% of the training instances leads to a satisfying POS-tagging performance on these pronouns.[2] The gap between debiasing with 5% of the documents and employing the complete debiasing training set of 625 documents, is only 2.62 percentage points for *hen* and 6.02 percentage points for *die*. These impressive results show that effective debiasing can be achieved with reduced access to resources.

The improvements in F1-scores are less pronounced compared to the improvements achieved with the full debiasing set. The disparity between pronoun scores and F1-scores suggests that, while the model rapidly adapts to accurately process gender-neutral pronouns, it requires additional exposure in order to adapt to the other differences between the original and debiasing set: name anonymisation and replacing gendered nouns by gender-neutral nouns. To illustrate this, I compute the F1-scores on the *hen* test set, but with the original names and gendered nouns unchanged. I evaluate the performance of the 5%-debiased model on this test set, and indeed observe an improvement in F1-

[2]Because two pronouns are simultaneously debiased in the current study, the 5% setting here corresponds to Björklund and Devinney's 2% setting, as the debiasing documents alternate between the usage of *hen* (2.5%) and *die* (2.5%).

score to 53.37 (+1.34). Note that this additional removal of gender clues is performed to support the debiasing of pronouns. This is achieved, among others, by concealing the strong overrepresentation of male entities in the corpus (see Section 3.1). However, real-world data does contain names and gendered nouns. Therefore, the model's comparatively slower adaptation to obscured gender clues does not impact the main findings. Future work could perform a more in-depth evaluation of these debiased models on naturally occurring data, but this is not done in the current study because the only available naturally occurring data is skewed in terms of gender and does not specifically contain gender-neutral names.

5.4 Unseen pronouns experiment

To address SQ4 and SQ5, namely, *can the debiasing methods Counterfactual Data Augmentation / delexicalisation improve system performance on previously unseen neopronouns?*, the final experiment evaluates the ability of all models (the original model, and fine-tuned from scratch and further fine-tuned delexicalisation and CDA models) to process pronouns that have not previously been encountered by the model. The reason for executing this evaluation is that novel (neo-)pronouns may be popularised in the future (Lauscher et al., 2022). Creating systems that will correctly process these pronouns or can easily adapt is preferred over debiasing strategies that are tailored exclusively to specific pronouns, as the latter require recurrent debiasing efforts each time a new pronoun gains popularity. Consequently, a debiasing method designed to address both current and prospective pronouns is favoured, in order to design future-proof systems.

In this section, I first describe the setup of the experiment (Section 5.4.1), I continue to discuss the results (Section 5.4.2) and finally perform an additional debiasing experiment with neopronouns in Section 5.4.3.

5.4.1 Setup

This experiment is conducted by evaluating the original and the debiased models on the *unseen test set*: a version of the regular data wherein all third-person pronouns are substituted by a neopronoun p, which is randomly selected from a set of six Dutch neopronouns: $p \in \{$ *dee/dem/dijr, dij/dem/dijr, nij/ner/nijr, vij/vijn/vijns, zhij/zhaar/zhaar, zem/zeer/zeer* $\}$. Notably, prior studies that involve debiasing coreference resolution systems (Lauscher et al., 2022; Zhao et al., 2018, 2019) have not included evaluations that focus on previously unseen pronouns. This will thus be the first evaluation of its kind.

Nonetheless, the delexicalisation method was specifically designed to enable the model to process pronouns of diverse lexical forms. However, as observed in the debiasing experiment, the debiasing abilities of this method proved unsatisfactory for gender-neutral pronouns (see Section 5.3). Consequently, I expect that this model will similarly fall short on effectively debiasing unseen pronouns. Conversely, CDA relies on instructing the model to process a particular pronoun by exposing it directly to the pronoun's lexical

	Precision	Recall	F1	Pronoun score
Original model	52.83 (σ=2.35)	44.37 (σ=2.92)	48.12 (σ=0.66)	46.68% (σ=2.31)
Fine-tuning the wl-coref model from scratch				
Delex	52.17 (σ=1.94)	49.55 (σ=1.95)	50.77 (σ=0.46)	48.03% (σ=2.01)
CDA	53.46 (σ=2.56)	49.66 (σ=3.23)	51.36 (σ=0.61)	51.72% (σ=2.90)
Further fine-tuning the wl-coref model				
Delex	50.98 (σ=0.83)	51.03 (σ=1.66)	50.99 (σ=0.72)	49.56% (σ=2.07)
CDA	53.01 (σ=0.73)	50.22 (σ=1.05)	51.57 (σ=0.61)	53.37% (σ=3.55)

Table 5.8: Model performances, in terms of the LEA metric and the pronoun score, on the *unseen* test set. This test set includes neopronouns that were not previously encountered by the models. The models include the original wl-coref model, alongside four models that were debiased through fine-tuning from scratch or further fine-tuning, using delexicalisation or CDA. The reported scores are the average of five random seeds.

form. Consequently, it is also expected that CDA will not enhance performance on unseen pronouns, as the model lacks exposure to these specific pronouns. Therefore, I expect neither delexicalisation nor CDA to enhance system performance on unseen pronouns.

5.4.2 Results

Table 5.8 reports the results for this experiment, both in terms of the LEA metric and the pronoun score.

Neither of the debiasing methods improves the performance on unseen pronouns. The original model has an unsatisfactory performance on unseen pronouns, with an F1-score of 48.12 and a pronoun score of only 46.68%. But, none of the the debiased models increases the performance on unseen pronouns to an acceptable level, the highest sores being for the *further fine-tuned CDA* model, which achieves an F1-score of 51.57 (+3.45 compared to the original model) and a pronoun score of 53.37% (+6.69% compared to the original model). This pronoun score is still 36.3 percent points lower than the performance of the same model on *die* pronouns (Table 5.5b). So, it appears that neither of the debiasing methods can effectively improve the model's performance on previously unseen neopronouns.

5.4.3 Neopronouns debiasing

In light of the disappointing abilities of the considered debiasing techniques to process previously unseen pronouns, I perform an additional investigation. In this additional experiment I assess the applicability of CDA fine-tuning, the most successful method identified in the preceding debiasing experiment, to the domain of neopronouns. For this experiment, I employ a small dataset, because the exploration in Section 5.3.3 showed satisfactory debiasing results for gender-neutral pronouns with a small set of debiasing documents. Such a setup is desirable, because it requires minimal resources and computational costs, still making the debiasing method somewhat future-proof.

Setup

The experimental procedure for this experiment aligns with that of the low-resource debiasing exploration performed for CDA fine-tuning in Section 5.3.3. However, in this experiment, I debias each pronoun individually, to be able to compare differences between neopronouns. For each pronoun, I create a debiasing and a test set with this particular pronoun inserted. So, rather than using the full *unseen* test set, I debias and evaluate the model's performance on each specific neopronoun set. I vary the amount of debiasing documents between three (0.625%), seven (1.25%), fifteen (2.5%) and 62 (10%) documents, always using five data partitions. I report the average performances across the five partitions. Additionally, for comparison, I evaluate the model performance on each *neopronoun-specific* test set (1) before debiasing and (2) after debiasing with the full training set (625 documents).

Debiasing for neopronouns is different from debiasing gender-neutral pronouns, because the former types do not yet exist in the language at all. This distinction introduces potential advantages and challenges to the debiasing process. On the one hand, a higher amount of debiasing data may be required to familiarise the model with these new types. But on the other hand, the absence of pre-existing usage patterns may alleviate ambiguity. For example, the gender-neutral pronoun *hen* may exhibit ambiguity in certain sentence structures as it can refer to either third-person singular or plural. In contrast, a neopronoun consistently maintains a singular and unambiguous referent, potentially facilitating the debiasing process.

Results

The outcomes of the debiasing process, in terms of pronoun scores, are presented in Figure 5.4.

The performance on the different pronouns varies a lot before debiasing. A relationship can be identified between the initial performance and the resemblance of the neopronoun to pronouns already familiar to the model. For instance, *zhij*, which is formed by combining the known gendered pronouns *hij* and *zij*, demonstrates an impressive initial performance of 84.55%. In contrast, *zem*, the neopronoun least resembling known pronouns, exhibits the lowest initial score of 35.66%.

A small number of debiasing documents can already improve the performance. For example, employing merely three training documents results in a substantial average performance improvement from 49.7% to 67.3% (+17.6 percentage points). However, with such a limited number of debiasing documents, high standard deviations between the data partitions are observed for the individual pronouns, as depicted in Figure 5.5. For instance, *dee* exhibits a standard deviation of 14.5 after debiasing with three documents. This is sensible, considering the significant variation in length and pronoun frequency across training documents (refer to Section 3). As the number of debiasing documents increases, the performances show improvements and standard deviations generally decrease.

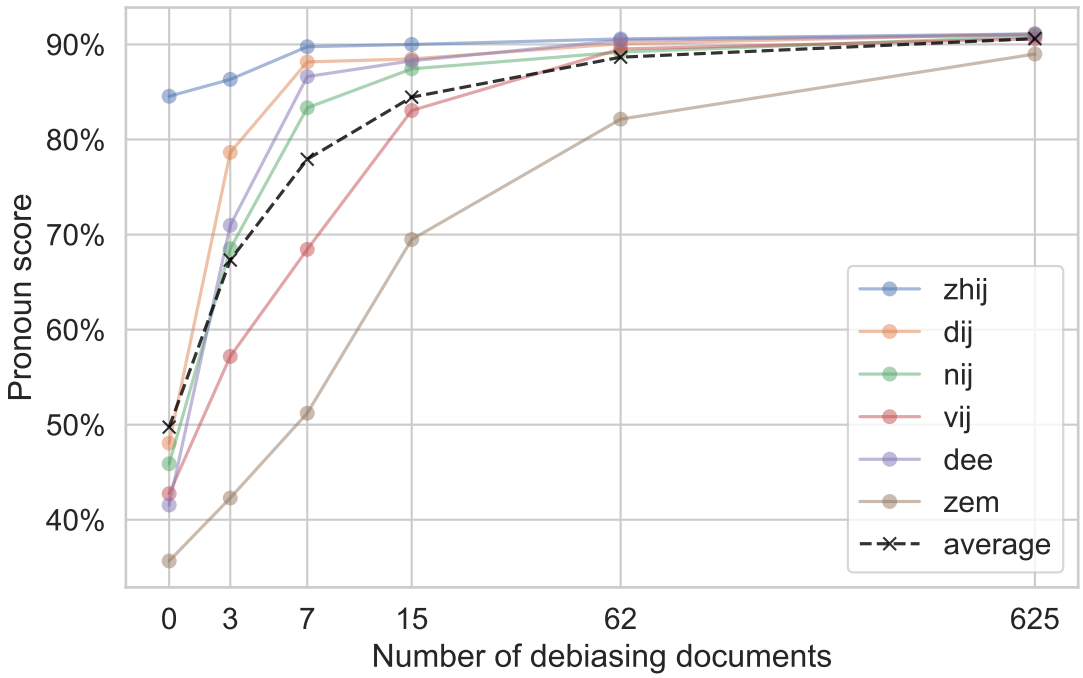


Figure 5.4: Pronoun score performance across six neopronouns as a function of the number of debiasing documents included in CDA fine-tuning. The black dotted line indicates the average pronoun scores across the different neopronouns. Reported scores are the average of five data partitions.

With the inclusion of 15 debiasing documents, the average performance on neopronouns already improves to 84.5%. For certain pronouns (*zhij, dij, nij*), performance begins to plateau with additional debiasing documents, whereas particularly *zem* continues to benefit from further debiasing. The precise quantity of debiasing documents required appears to be contingent on the specific pronoun. Nevertheless, a consistent observation across all pronouns is that even a small number of debiasing documents can substantially enhance performance on the respective pronoun.

Conclusion. Taken together, the outcomes of this experiment are promising, even despite the fact that debiasing with gender-neutral pronouns does not improve the performance on previously unseen neopronouns. Namely, satisfactory results are achieved across various sets of neopronouns through the application of CDA fine-tuning with a limited dataset. These findings underscore the feasibility of future-proof gender-inclusive debiasing with minimal resource requirements and low computational costs.

5.5 A test suite for pronoun-related behaviour

In this section, I create a test suite to more thoroughly evaluate the models' ability to process gender-neutral pronouns, with the purpose of identifying potential systematic

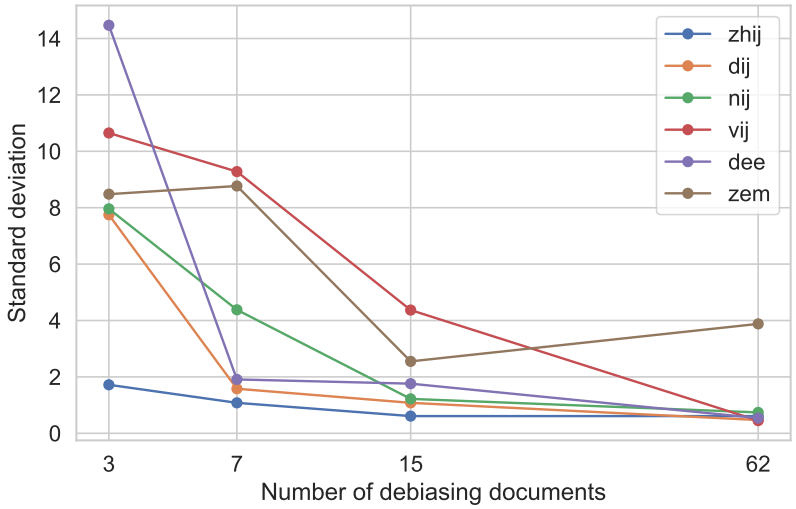


Figure 5.5: Standard deviations of the pronoun score performances across five different data partitions, for six neopronouns, as a function of the number of debiasing documents included in CDA fine-tuning. The zero and 625 training documents settings are excluded from this figure, because these two settings only use one data partition (an empty set or the full set).

errors. The template-based test suite is created through the CheckList framework (Ribeiro et al., 2020). This framework facilitates the creation of a large number of test cases through the usage of templates. The created templates will function as minimum functionality tests: simple tests that evaluate whether a model can perform a specific functionality.

Some of the templates take inspiration from the WinoNB dataset (Baumler and Rudinger, 2022) (Section 2.4.1), but are adapted to fit the Dutch context. Across the tests, a comparative analysis is conducted between the original wl-coref model and the most efficacious debiased model, *CDA through fine-tuning*. I consider this the best debiased model because it obtains the best results across all evaluations and it reduces the performance gap between gendered and gender-neutral pronouns to less than 1%. I evaluate both models using five seeds, and report the average scores and standard deviations.

Using this 1800-sentences evaluation set, I investigate four core capabilities related to the processing of gender-neutral pronouns. Table 5.9 provides an overview of the number of sentences per test. The first test (Section 5.5.1) assesses the models' ability to correctly resolve pronouns to co-refer with three distinct categories of names: gender-neutral names, typically male associated names and typically female associated names. Subsequently, in the second test (Section 5.5.2) I investigate the model's ability to handle sentences in which an individual uses multiple different pronouns. The third evaluation (Section 5.5.3) investigates the models' capacity to differentiate between the singular and plural usage of the gender-neutral pronoun *hen*. Finally, in Section 5.5.4, I investigate whether the model can distinguish between the usage of *die* as a personal pronoun and its usage as a relative pronoun.

Test	# of templates	# of settings	# of sentences
Pronoun-name links	1	3 name settings; 2 gender settings	600
Multiple pronouns per entity	2	3 gender settings	600
Singular - plural disambiguation	2	2 gender settings	400
Recognising different functions of die	1	2 gender settings	200
Total			1800

Table 5.9: The number of templates, settings and sentences per test in the test suite.

Gender setting	Example sentence
Gendered	*Robin gaat naar **zijn/haar** buren.* *(Robin goes to **his/her** neighbours)*
Gender-neutral	*River gaat naar **hun/diens** dokter.* *(River goes to **their** doctor)*
Place	$\in$ {*werk (work), dokter (doctor), afspraak (appointment), buren (neighbours), ouders (parents), interview (interview)*}

Table 5.10: Example of the *pronoun-name link* template, in the two gender settings. Boldface marks the pronouns of interest and the underlined words highlight words varied between the sentences. The slots (i.e. the name slot, the pronoun slot and the place slot) are filled randomly for each instance. This means that the distribution of *zij*/*hij* pronouns in the gendered setting is roughly 50/50, and similarly the distribution of *die*/*hen* pronouns in the gender-neutral setting is roughly 50/50.

5.5.1 Pronoun-name links

In the first test, I evaluate whether pronouns can correctly be resolved to corefer with names. This should be a very simple coreference task, and the objective of this evaluation therefore is to ascertain the models' proficiency in executing this fundamental capability. For this reason, the template for this test is intentionally made as simple as possible, to eliminate confounding factors that could potentially complicate the basic task, and thereby hinder the interpretation of the results.

I separately test two gender settings : a gendered setting, using gendered pronouns (*zij*, *hij*), and a gender-neutral setting, incorporating gender-neutral pronouns (*die*, *hen*). I test these sets of pronouns separately, to inspect if the usage of gender-neutral pronouns poses an additional challenge to the models in performing this core task. Table 5.10 shows the template used in this test across the two settings, where the underlined words are varied throughout the sentences, and the bold words mark the pronouns of interest.

For both of the two pronoun settings, three distinct groups of names are compared: gender-neutral names, typically male associated names and typically female associated names. By considering these different groups of names, I aim to discern whether the model incorporates prevalent gender associations of names in its decision-making processes. CheckList has numerous lists of names integrated into its framework, which are extracted from Wikipedia, and divided over separate lists per binary gender and country. I use the lists of female and male names from the Netherlands. Moreover, the list

	Original wl-coref model mistakes		Debiased mistakes	
	Gender setting		Gender setting	
Name setting	*Gendered*	*Gender-neutral*	*Gendered*	*Gender-neutral*
Gender-neutral	0.8% ($\sigma = 0.8$)	41.0% ($\sigma = 12.0$)	0.8% ($\sigma = 0.5$)	0.0% ($\sigma = 0.0$)
Male	0.8% ($\sigma = 0.8$)	48.1% ($\sigma = 13.1$)	0.4% ($\sigma = 0.5$)	0.4% ($\sigma = 0.4$)
Female	4.4% ($\sigma = 4.4$)	37.8% ($\sigma = 11.9$)	0.4% ($\sigma = 0.9$)	0.0% ($\sigma = 0.0$)

Table 5.11: Percentage of mistakes made by the original wl-coref model and the debiased model on the *pronoun-name link* template sentences, across the name and gender settings. Results are the average of five random model seeds.

of gender-neutral names is hand-crafted for this particular task, compiled by searching online for gender-neutral names popular in the Dutch context. This list can be found in Appendix D.

Per setting, 100 test instances are generated. As there are three name settings, and two gender settings, there are six distinct settings in total, resulting in the creation of 600 sentences. For each sentence, a word is randomly selected from the full list of options to fill each designated slot (in this case the name slot, the pronoun slot and the place slot).

Table 5.11 shows the portion of sentences incorrectly resolved by the models across the settings. Given that each setting comprises 100 sentences, the percentage of incorrectly resolved sentences corresponds to the average number of sentences in which an error occurred. Across all names settings, the original model hardly makes any mistakes for the sentences that include gendered pronouns. Notably, the model demonstrates no problems with linking male pronouns to typically female associated names and vice versa. However, the number of mistakes increases to around 40% for the sentences with gender-neutral pronouns, with the highest number of mistakes for male names. In contrast, the debiased model demonstrates a near perfect performance, indicating a substantial improvement over the original model.

5.5.2 Multiple pronouns per entity

In the second test, I assess the model's competence in correctly resolving an individual's concurrent usage of multiple sets of pronouns. Lauscher et al. (2022) shed light on the importance of this ability, as they point out that non-binary individuals often identify with multiple sets of pronouns. I test this through two templates.

In the first template, illustrated in Table 5.12, I investigate the coreference relation between two pronouns, specifically testing whether the model correctly identifies the two distinct pronouns as belonging to the same cluster. I again compare the usage of gendered pronouns to that of gender-neutral pronouns, to ascertain any potential influence of the pronoun type on the models' proficiency in executing this core task. Additionally, I test a mixed setting wherein any pronoun combination is possible.

The percentage of mistakes per setting can be found in the Table 5.13. Again, each gender setting contains 100 sentences, so the percentage of incorrectly resolved sentences equals the average number of mistakes per setting. The original model makes a low number

Gender setting	Example sentence
Gendered	***Hij/zij*** *gaat naar* ***zijn/haar*** *afspraak.* *(****He/she*** *goes to* ***his/her*** *appointment).*
Gender-neutral	***Hen/die*** *gaat naar* ***hun/diens*** *ouders.* *(****They*** *go to* ***their*** *parents).*
Mixed	***Hij/zij/hen/die*** *gaat naar* ***zijn/haar/hun/diens*** *werk.* *(****He/she/they*** *goes/go to* ***his/her/their*** *work).*
Place	∈{*werk (work), dokter (doctor), afspraak (appointment),* *buren (neighbours), ouders (parents), interview (interview)* }

Table 5.12: Example of the first *multiple pronouns per entity* template, in the three gender settings. Note that because each slot is filled randomly, the sentence set also includes instances where the pronouns match (e.g. using *zij* and *haar*).

Original wl-coref model mistakes			**Debiased model mistakes**		
Gender setting			Gender setting		
Gendered	*Gender-neutral*	*Mixed*	*Gendered*	*Gender-neutral*	*Mixed*
9.8% ($\sigma = 9.5$)	63.0% ($\sigma = 11.8$)	39.6% ($\sigma = 15.1$)	0.8% ($\sigma = 1.1$)	0.0% ($\sigma = 0.0$)	0.8% ($\sigma = 1.1$)

Table 5.13: Percentage of mistakes on the first *multiple pronouns per entity* template sentences, across the gender settings. Results are the average of five random model seeds.

of mistakes ($< 10\%$) in combining gendered pronouns. This is an interesting observation, indicating that the original model does not necessarily have problems with linking different pronouns together for an individual. Conversely, the percentage of mistakes increases in the other settings: to 63% in the gender-neutral setting, and to nearly 40% in the mixed setting. Notably, within the latter category, most sentences for which an error occurs involve at least one gender-neutral pronoun. This implies a challenge for the model in accurately processing gender-neutral pronouns, as opposed to an inherent difficulty in the fundamental capacity of combining different pronoun types within a single cluster. The debiased model makes close to no mistakes for this template. Interestingly, the performance on in the gendered setting also improves. This is in line with the fact that the pronoun scores for *hij* and *zij* also improve after debiasing (Table 5.5): exposure to additional training also appears to benefit the performance on familiar pronouns.

In the second template a name is additionally included, which is selected from the list of gender-neutral names. Again, I test whether the model correctly identifies the name and the two pronouns as corefering, using the same three gender settings as before. The templates are presented in Table 5.14. Again, each gender setting contains 100 sentences, so I test 300 sentences in total. The template slots are again filled randomly, so the sentences are not *exactly* the same as in the first template set.

Table 5.15 presents the percentage of mistakes by both models across the different settings. The original model has a nearly perfect performance in combining gendered pronouns, but fails to combine gender-neutral pronouns in over half of the sentences. These scores are somewhat better than for the first template, but the standard deviation in the gender-neutral setting is higher. In the mixed setting, the performance is similar to that of the first template. On the other hand, the debiased model does not make any

Gender setting	Example sentence
Gendered	*Billie heeft **zijn/haar** afspraak bij de sportschool om 3 uur. **Hij/zij** gaat zo heen.* *(Billie has **his/her** appointment at the gym at 3 o'clock. **He/she** is going there soon).*
Gender-neutral	*Moos heeft **hun/diens** afspraak bij de buren om 3 uur. **hen/die** gaat zo heen.* *(Moos has **their** appointment at the neighbours at 3 o'clock. **they** are going there soon).*
Mixed	*Kit heeft **zijn/haar/hun/diens** afspraak bij de winkel om 3 uur. **Hij/zij/hen/die** gaat zo heen.* *(Kit has **his/her/their** appointment at the store at 3 o'clock. **He/she/they** is/are going there soon).*
Place	$\in$ { *dokters (docters), huisartsenpost (general practice centre), gemeente (municipality), buren (neighbours), winkel (store), sportschool (gym)* }

Table 5.14: Example of the second *multiple pronouns per entity* template, in the three gender settings.

Original wl-coref model mistakes			**Debiased model mistakes**		
Gender setting			Gender setting		
Gendered	*Gender-neutral*	*Mixed*	*Gendered*	*Gender-neutral*	*Mixed*
0.6% ($\sigma = 1.3$)	51.0% ($\sigma = 22.3$)	36.4% ($\sigma = 17.7$)	0.0% ($\sigma = 0.0$)	0.0% ($\sigma = 0.0$)	0.0% ($\sigma = 0.0$)

Table 5.15: Percentage of mistakes on the second *multiple pronouns per entity* template sentences, across the gender settings. Results are the average of five random model seeds.

mistakes. This result is encouraging, as it shows that the model can correctly identify the usage of multiple pronouns for an individual, at least within a single sentence.

5.5.3 Singular - plural disambiguation

Thirdly, I test if the model can distinguish between the singular and plural usage of *hen*. This test is inspired by the WinoNB dataset (Baumler and Rudinger, 2022), in which each sentence includes an individual and a group, together with a *they* pronoun, which refers to either of the two. The context of the sentence indicates whether the usage of *they* is singular or plural, and the evaluation set tests whether the singular and plural usages can be resolved equally well. I create two templates that are based on this structure, adapted to the Dutch context.

In the first template, presented in Table 5.16, two pronouns are included. The first pronoun is a third-person subject pronoun (*hij, zij* or *hen*), that refers to an individual. The second pronoun is the third-person plural object *hen*. The template again includes a gendered and a gender-neutral version. While the English translation shows an ambiguous resolution of the pronouns in the gender-neutral setting, this is not the case in Dutch, where the verb conjugation unequivocally indicates that the initial instance of *hen* is singular, and the subsequent one is plural. Through the gendered template, I aim to test whether the model can correctly resolve the second pronoun as plural. In the gender-neutral setting, the pronoun *hen* appears twice. By comparing this gender-neutral setting with the gendered setting, I evaluate whether different resolutions of the same type poses an extra challenge to the model. I again use names from the list of gender-neutral names,

Gender setting	Example sentence
Gendered	*Madu gaat met het hockeyteam op vakantie.* ***Hij/zij*** *is al vaker met* ***hen*** *weggeweest.* (*Madu goes on holiday with the hockey team.* ***He/she*** *has been away with* ***them*** *[plural] before*).
Gender-neutral	*Lyric gaat met het voetbalteam op vakantie.* ***Hen*** *is al vaker met* ***hen*** *weggeweest.* (*Lyric goes on holiday with the football team.* ***They*** *[singular] have been away with* ***them*** *[plural] before*).
Sports	∈ { *voetbal (football), volleybal (volleyball), hockey (hockey), handbal (handball), softbal (softball), honkbal (baseball), basketbal (basketball)* }

Table 5.16: Example of the first *singular - plural disambiguation* template.

Original wl-coref model mistakes		**Debiased model mistakes**	
Gender setting		Gender setting	
Gendered	*Gender-neutral*	*Gendered*	*Gender-neutral*
1.4% ($\sigma = 2.6$)	60.2% ($\sigma = 34.8$)	76.2% ($\sigma = 24.8$)	91.0% ($\sigma = 10.0$)

Table 5.17: Percentage of mistakes on the first *singular - plural disambiguation* template sentences, across the gender settings. Results are the average of five random model seeds.

and create 100 test sentences per gender setting, testing 200 sentences in total.

The percentage of mistakes is reported in Table 5.17. The original model can correctly identify the plural resolution of *hen* in the gendered setting, with an error rate of 1.4%. However, when gender-neutral pronouns are used, this model fails in 60% of the instances, predicting both instances of *hen* to refer to the (plural) sports team. So, the plural instance of *hen* is correctly resolved, but the singular instance of *hen* is not.

The debiased model makes different mistakes. Using gendered pronouns, it incorrectly resolves the pronouns in 76% of the sentences. Here, it predicts both pronouns to refer to the individual. Moreover, in the gender-neutral setting the model fails in 91% of the cases, by predicting both pronouns to refer to the individual as well. The debiasing thus appears to result in an overcorrection: the model seems to forget the plural usage of *hen* in this context. This is not unexpected considering that plural *hen* only appears 290 times in the original corpus, while its singular sense is inserted 2058 times in the training data through debiasing.

The second template, presented in Table 5.18, also includes two pronouns. The first pronoun is the third-person plural subject *zij* and the second pronoun is a third-person singular object *hem, haar* or *hen*. In this template, the *hen* pronoun (only used in the gender-neutral setting) should thus be resolved as being singular. The main task here is to test the model's ability to identify that the *hen* pronoun is singular and does not corefer with the plural *zij* pronoun. The gendered pronouns are included for comparison, to see whether the test is more straightforward if the singular pronoun lacks multiple usages. But, notably, the type *zij*, which is used as a third-person plural pronoun, is also used for third-person singular female subjects, potentially posing an additional challenge for the model. Consequently, it is of interest to determine whether the model accurately identifies the plural usage of this pronoun.

Gender setting	Example sentence
Gendered	*Bobby gaat met het basketbalteam op vakantie.* ***Zij*** *zijn al vaker met* ***hem/haar*** *weggeweest.* (*Bobby goes on holiday with the basketball team.* ***They*** *have been away with* ***him/her*** *before*).
Gender-neutral	*Ash gaat met het honkbalteam op vakantie.* ***Zij*** *zijn al vaker met* ***hen*** *weggeweest.* (*Bobby goes on holiday with the baseball team.* ***They*** *[plural] have been away with* ***them*** *[singular] before*).
Sports	$\in$ { *voetbal (football), volleybal (volleyball), hockey (hockey), handbal (handball), softbal (softball), honkbal (baseball), basketbal (basketball)* }

Table 5.18: Example of the second *singular - plural disambiguation* template.

Original wl-coref model mistakes		**Debiased model mistakes**	
Gender setting		Gender setting	
Gendered	*Gender-neutral*	*Gendered*	*Gender-neutral*
38.4% ($\sigma = 24.7$)	100.0% ($\sigma = 0.0$)	27.8% ($\sigma = 36.7$)	49.0% ($\sigma = 29.5$)

Table 5.19: Percentage of mistakes on the second *singular - plural disambiguation* template sentences, across the gender settings. Results are the average of five random model seeds.

Similar to the first template, names are drawn from the list of gender-neutral names. Moreover, the Dutch grammar again disambiguates the resolution of the pronouns in the gender-neutral setting, even though the English translation does not reflect this.

The results are presented in Table 5.19. The original model makes mistakes for 38% of the sentences in which gendered pronouns are used, with a high standard deviation. In most of these cases, the plural pronoun *zij* is considered as singular, corefering with *haar* or *hem* and the name. The model thus exhibits difficulty in distinguishing between the different usages of *zij*. When gender-neutral pronouns are used, the original model always fails to resolve the sentences, consistently interpreting *hen* as plural and corefering with the sports team.

The debiased model makes fewer mistakes. In the gendered settings, it gets 28% of the sentences wrong, but with an extremely high standard variation. In the gender-neutral setting, the model makes wrong predictions for around half of the sentences, also with a high standard deviation. Across the different model seeds, the model makes different types of mistakes. Two of the seeds make mistakes for *zij* (e.g. predicting it to be singular). The other three seeds sometimes predict all pronouns to refer the the sports team. This indicates that the model does not always forget the plural usage of *hen* entirely, although the behaviour for this template is not very stable.

5.5.4 Recognising different functions of die

In the last test, I investigate whether the model can distinguish between the usage of *die* as (1) a personal pronoun and (2) a relative pronoun. To this end, I use the templates in Table 5.20. In the gender-neutral setting, the first occurrence of *die* is as a relative pronoun, while the second occurrence is as a personal pronoun. As a comparison, I also

Gender setting	Example sentence
Gendered	*Bo* ***die*** *hier net werkt is te laat omdat* ***hij/zij*** *een lekke band had.* (*Bo* ***who*** *just started working here is late because* ***he/she*** *had a flat tire.*)
Gender-neutral	*Jamie* ***die*** *hier net werkt is te laat omdat* ***die*** *een lekke band had.* (*Jamie* ***who*** *just started working here is late because* ***they*** *had a flat tire.*)

Table 5.20: Example of the *die* template, in the two gender settings.

Original wl-coref model mistakes		**Debiased model mistakes**	
Gender setting		Gender setting	
Gendered	*Gender-neutral*	*Gendered*	*Gender-neutral*
6.4% ($\sigma = 10.3$)	9.4% ($\sigma = 9.2$)	2.0% ($\sigma = 1.4$)	4.2% ($\sigma = 4.3$)

Table 5.21: Percentage of mistakes on the *die* template sentences. Results are the average of five random model seeds.

evaluate the performance in the gendered setting, wherein the personal pronoun is either *hij* or *zij*, rather than *die*, and there is thus no pronoun repetition in the sentence.

The results are presented in Table 5.21. Both models get a correct result in over 90% of the sentences, across both settings. The performance of the debiased model is slightly better, with notably lower standard deviations.

5.5.5 Conclusion

To conclude this evaluation, the debiased model achieves a score of over 90% in three out of four tests, while the original model only achieves such a performance for the last test. The only task on which the debiased model does not achieve a good performance concerns distinguishing between the singular and plural usage of *hen*. Specifically, the model struggles to identify the plural usage of *hen* in specific contexts. Given that the original model performs better in this context, the debiasing process appears to result in an overcorrection, leading to an inaccurate processing of the plural usage of *hen*. This outcome is not unexpected, considering the relatively low frequency of plural instances of *hen* in the training data. Subsequent research efforts may explore approaches to debiasing that preserve the plural usage of *hen*, potentially by incorporating additional instances of plural *hen* in the debiasing set.

Moreover, a limitation of the presented test suite is that it only contains one or two templates per task. Future investigations could expand the scope of templates to provide a more comprehensive understanding of the model's capabilities.

6 Discussion and conclusion

In this final chapter, I line out the limitations of the current study and point to directions for future research in Section 6.1. Following this, I present my conclusions together with a discussion of the findings in Section 6.2.

6.1 Limitations and future work

This study has several limitations. First of all, I zoom in on gender-neutral pronouns alone, and discard any other dimension in which the language of non-binary individuals may differ from that of people with a binary gender identity, such as vocabulary[1] and style. Despite the fact that I observe that debiasing through CDA improves the performance on gender-neutral pronouns, this does not imply that the performance of the coreference resolution system would also improve on real-world data from non-binary individuals, because the data considered in this study still stems from binary-gendered contexts. Therefore, an important direction for future work would be to test, and if necessary debias, model performance on Dutch data from transgender individuals, for instance through creating a Dutch equivalent of the GICoref corpus (Cao and Daumé III, 2020).

Secondly, this study has not been able to actively involve non-binary and transgender individuals in the designing, debiasing and evaluation process. Despite having asked advice at multiple stages from a transgender individual in my personal network, this study is for the main part conducted by cisgender individuals in a binary gendered environment. I recognise that this may lead to having overlooked important barriers, risks or opportunities relating to the emancipation of non-binary individuals. Personally, I believe the current study, which exclusively looks at gender-neutral pronouns, can be considered a small step, at the beginning stages of achieving emancipation and a fair treatment of non-binary individuals in Dutch language technologies. But, in the steps that follow towards achieving this goal, the active involvement of non-binary individuals, for instance through participatory design initiatives (Caselli et al., 2021), is essential (Devinney et al., 2022).

Thirdly, the current study only considers a single model in its evaluation. Subsequent investigations could extend the scope by conducting a more comprehensive comparison, exploring potential trends across various models. Moreover, a compelling avenue for future research involves assessing the applicability of the current setup to other languages. Notably, for languages like Italian or French, in which gender is more intricately woven into grammatical structures, the debiasing task may prove more complex.

Another compelling direction for future exploration involves extending a similar investigation to diverse NLP tasks, such as machine translation or question answering. This also prompts the question of how the proposed methodology performs when applied to

[1] For example, non-binary individuals may use *neo*nouns such as *brus* (*sibling*): a contraction of *broer* (*brother*) and *zus* (*sister*).

distinct types of data sources. Notably, the prevalence of third-person pronouns can be anticipated to vary across different text genres. For instance, direct interactions tend to incorporate a higher proportion of second-person pronouns, while news text and Wikipedia articles, the primary data genres in the present study, are anticipated to have a higher frequency of third-person pronouns. Consequently, the debiasing of systems that rely on direct conversational data, such as chat bots, might necessitate additional debiasing data or more specific data tailored to debiasing.

Finally, a noteworthy observation in the test suite results revealed that debiasing can lead to the partial forgetting of the plural usage of *hen*, which gives rise to several follow-up questions. An compelling direction for exploration involves investigating whether debiasing can be executed in a manner that prevents the model from forgetting the plural usage, potentially through additionally enhancing the frequency of this pronoun's plural usage in the debiasing data. Furthermore, an interesting future direction is to examine analogous issues in other languages, such as English. Specifically, exploring whether debiasing efforts on English singular *they* could impact the model's performance on plural instances of *they* could be an interesting investigation. Looking into these questions could provide novel insights into the effectiveness and potential nuances of debiasing methodologies in non-binary contexts.

6.2 Discussion of results

In this study, the efficacy of existing debiasing techniques, specifically delexicalization and Counterfactual Data Augmentation (CDA), in enhancing the performance of an existing Dutch coreference resolution model on gender-neutral pronouns is examined. In this section I present the final answers to the main research questions.

SQ1: How good is an existing Dutch coreference resolution system at processing gender-neutral pronouns compared to gendered pronouns?
In the gender neutral pronoun evaluation experiment (Section 5.2), I identify that the wl-coref model exhibits a diminished proficiency in processing gender-neutral third-person pronouns compared to its handling of gendered counterparts. Particularly, the pronoun scores on gender-neutral pronouns are on average 13.1% lower than for the gendered pronouns. These findings lead me to conclude that the evaluated coreference resolution system, wl-coref, performs worse on gender-neutral pronouns than on gendered pronouns.

SQ2: Can the debiasing method Counterfactual Data Augmentation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns?
In the debiasing experiment (Section 5.3), the debiasing results show that applying CDA manifests a considerable improvement, almost entirely closing the performance gap between gendered and gender-neutral pronouns. This method proves effective not only in the context of fine-tuning from scratch but also when employed to further fine-tune an existing model, a setup that reduces the computational costs. These observations lead me to conclude that CDA can indeed improve the model's performance on gender-neutral pronouns.

Moreover, when applying CDA through further fine-tuning with just a handful of documents, as detailed in Section 5.3.3, this method remains effective, notably reducing the performance gap between gendered and gender-neutral pronouns. This outcome aligns with the finding of Björklund and Devinney (2023), that effective debiasing of gender-neutral pronouns can be achieved with a low number of debiasing instances; and more generally it underscores the feasibility of debiasing in non-binary contexts with minimal resources and low computational costs. This result is particularly noteworthy given the absence of gender-neutral pronouns in the original training data and the general novelty of gender-neutral pronouns in the Dutch language.

SQ3: Can the debiasing method delexicalisation improve the ability of a Dutch coreference resolution system to process gender-neutral pronouns?
The outcomes pertaining to delexicalisation do not indicate efficacy in enhancing performance with regard to gender-neutral pronouns. Whether in the context of fine-tuning from scratch or further fine-tuning the original model, the performance scores for gender-neutral pronouns do not improve through the application of delexicalisation.

SQ4 & SQ5: Can the debiasing methods Counterfactual Data Augmentation / delexicalisation improve system performance on previously unseen neopronouns?
In the unseen pronouns experiment (Section 5.4), I evaluate the ability of the original and the debiased models to process neopronouns that are previously unseen by the model. The findings indicate that none of the models demonstrates a satisfactory ability to process unseen pronouns, and that neither of the debiasing techniques improves the model's ability to process them.

However, when applying CDA with neopronouns inserted into the debiasing data (Section 5.4.3), the performance on these pronouns readily improves, even when just a few debiasing documents are used. This observation highlights the effectiveness and adaptability of CDA. Despite the necessity of additional debiasing efforts for each potentially novel pronoun, the method can still be considered future-proof in the sense that it requires only minimal intervention to facilitate the model in processing emergent pronouns.

The results observed in this study furthermore suggest that there exist an opportunity for NLP technologies to be at the forefront of emancipation movements, by enabling systems to adeptly process emerging languages structures, which are embraced by pioneers but are not yet prevalent throughout broader societies. The Dutch gender-neutral pronouns and neopronouns serve as illustrative instances of such emergent linguistic constructs. Notably, the implementation of NLP technologies in this context holds a potential of facilitating the wider adoption of these innovative structures within societies, by showing people an example of how to correctly use these structures.

By addressing the challenges posed by Dutch gender-neutral pronouns and neopronouns, this research contributes to the development of more inclusive AI systems. It opens avenues for similar approaches in other languages and encourages the integration of novel linguistic constructs, potentially fostering societal acceptance and adoption of these linguistic innovations. On top of that, this study makes a significant contribution towards the overarching objective of mitigating the adverse impacts that language technologies may pose to non-binary and transgender individuals, such as misgendering and erasure.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica*, pages 139–159.

Y Gavriel Ansara and Peter Hegarty. 2014. Methodologies of misgendering: Recommendations for reducing cisgenderism in psychological research. *Feminism & Psychology*, 24(2):259–270.

Chinatsu Aone and Scott William. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Maria Barrett, Hieu Lam, Martin Wu, Ophélie Lacroix, Barbara Plank, and Anders Søgaard. 2021. Resources and evaluations for Danish entity resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 63–69, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

Sander Becker. 2020. Is het Nederlands klaar voor het genderneutrale 'Hen loopt'? *Trouw*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Henrik Björklund and Hannah Devinney. 2023. Computer, enhence: POS-tagging improvements for nonbinary pronoun use in Swedish. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 54–61, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rodrigo Borba. 2017. Ex-centric textualities and rehearsed narratives at a gender identity clinic in brazil: challenging discursive colonization. *Journal of Sociolinguistics*, 21(3):320–347.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Gosse Bouma, Walter Daelemans, Iris Hendrickx, Véronique Hoste, and A Mineur. 2007. The COREA-project, manual for the annotation of coreference in dutch texts. *University Groningen*.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dennis Connolly, John D Burger, and David S Day. 1997. A machine learning approach to anaphoric reference. In *New methods in language processing*, pages 133–144.

Kirby Conrod. 2019. *Pronouns raising and emerging*. Ph.D. thesis. University of Washington.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2022. Constructing a cross-document event coreference corpus for Dutch. *Language Resources and Evaluation*, pages 1–30.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2007. A ranking approach to pronoun resolution. In *International Joint Conference on Artificial Intelligence*.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del lenguaje natural*, 42.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Doddington, 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 the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

EditieNL. 2021. Lij of vij: is er een nieuw non-binair persoonlijk voornaamwoord nodig? *RTLnieuws*.

Fatma Elsafoury and Gavin Abercrombie. 2023. On the origins of bias in nlp through the lens of the jim code. *arXiv preprint arXiv:2305.09281*.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1263–1274, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Europees Parlement. 2018. Genderneutraal Taalgebruik in het Europees Parlement. https://www.europarl.europa.eu/cmsdata/187106/GNL_Guidelines_NL-original.pdf. Accessed: 2022-09-28.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347.

Spencer Garrison. 2018. ON THE LIMITS OF "TRANS ENOUGH": Authenticating Trans Identity Narratives. *Gender and Society*, 32(5):613–637.

Maartje Geels. 2022. Politieagent of -persoon? Taalunie zoekt uitweg in gender-taalworsteling. *NOS Nieuws*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336.

Marinel Gerritsen. 2002. Language and gender in Netherlands Dutch: Towards a more gender-fair usage. *Gender Across Languages*, 2.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware adversarial training for name regularity bias in named entity recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 901–912, New York, NY, USA. Association for Computing Machinery.

Tobias Glasmachers. 2017. Limits of end-to-end learning. In *Proceedings of the Ninth Asian Conference on Machine Learning*, volume 77 of *Proceedings of Machine Learning Research*, pages 17–32, Yonsei University, Seoul, Republic of Korea. PMLR.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marie Gustafsson Sendén, Emma A Bäck, and Anna Lindqvist. 2015. Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in Psychology*, 6.

Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist. 2021. Pronouns beyond the binary: The change of attitudes and use over time. *Gender & Society*, 35(4):588–615.

Arno Haijtema. 2021. De portretten van Zanele Muholi zijn een oproep je te bevrijden van betekenis. *De Volkskrant*.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. 2022. A Systematic Study of Bias Amplification. *arXiv preprint arXiv:2201.11706*.

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 452–459, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Annemarie Haverkamp. 2021. Hoe genderneutraal is genderneutrale taal? *Vox*.

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 the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Iris Hendrickx, Veronique Hoste, and Walter Daelemans. 2008b. Semantic and syntactic features for Dutch coreference resolution. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, page 351–361, Berlin, Heidelberg. Springer-Verlag.

Het Neutrale Taal collectief. Lijst van populaire voornaamwoorden. `https://nl.pronouns.page/voornaamwoorden`. Accessed: 2022-10-14.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

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 the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Eduard Hovy, 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.

Robin Mattias Hurkens. 2021. Genderneutraal: geen 'hij', 'zij', 'hen', 'dij', maar 'ij'. *De Volkskrant*.

Sandy James, Jody Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anaf. 2016. The Report of the 2015 U.S. Transgender Survey. *National Center for Transgender Equality*.

Kelly Johnson, Colette Auerswald, Allen J LeBlanc, and Walter O Bockting. 2019. 7. invalidation experiences and protective factors among non-binary adolescents. *Journal of Adolescent Health*, 64(2):S4.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dan Jurafsky and James H. Martin. 2021. *Speech and Language Processing (3rd ed. draft)*.

Lotte Kamphuis and Roelien Akse. 2021. Van overdreven en te correct tot een stuk inclusiever: mensen verdeeld over genderneutraal taalgebruik. *EenVandaag*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Lex Konnelly. 2021. Nuance and normativity in trans linguistic research. *Journal of Language and Sexuality*, 10(1):71–82.

Lex Konnelly and Elizabeth Cowper. 2020. Gender diversity and morphosyntax: An account of singular they. *Glossa: a journal of general linguistics*, 5(1).

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Anne C Kroon, Damian Trilling, Toni GLA van der Meer, and Jeroen GF Jonkman. 2020. Clouded reality: News representations of culturally close and distant ethnic outgroups. *Communications*, 45(s1):744–764.

Anne C Kroon and Toni GLA van der Meer. 2021. Who's to fear? Implicit sexual threat pre and post the "refugee crisis". *Journalism Practice*, 17(2):319–335.

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

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Ida Marie S. Lassen, Mina Almasi, Kenneth Enevoldsen, and Ross Deans Kristensen-McLachlan. 2023. Detecting intersectionality in NER models: A data-driven approach. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 116–127, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

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 the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach . *arXiv preprint arXiv:1907.11692*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xiaoqiang Luo and Sameer Pradhan. 2016. Evaluation metrics. In Massimo Poesio, Roland Stuckardt, and Versley Yannick, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, chapter 10, pages 141–163. Springer, Berlin Heidelberg.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph F. McCarthy and Wendy G. Lehnert. 1995. Using decision trees for conference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1050–1055, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.

Sebastian McGaughey. 2020. Understanding Neopronouns. *The Gay & Lesbian Review Worldwide*.

Kevin A McLemore. 2015. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74.

Mey. 2014. It's Time For People to Stop Using the Social Construct of "Biological Sex" to Defend Their Transmisogyny. *Autostraddle*.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Nafise Sadat Moosavi. 2020. *Robustness in Coreference Resolution*. Ph.D. thesis. University of Heidelberg.

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 the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Vincent Ng. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on web science conference*, pages 405–406.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. SoNaR User Documentation. 1(4).

Hadas Orgad and Yonatan Belinkov. 2022. Choose your lenses: Flaws in gender bias evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 151–167, Seattle, Washington. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Lindsay Poirier. 2018. *Knowledge Representation in Scruffy Worlds an Ethnography of Semiotic Infrastructure Design Work*. Rensselaer Polytechnic Institute.

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 79–90, Barcelona, Spain (online). Association for Computational Linguistics.

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, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA. Association for Computational Linguistics.

Micah Rajunov and A Scott Duane. 2019. *Nonbinary: Memoirs of gender and identity*. Columbia University Press.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural language engineering*, 17(4):485–510.

Marta Recasens, 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 the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

Martin Reynaert, Nelleke Oostdijk, Orphée De Clercq, Henk van den Heuvel, and Franciska de Jong. 2010. Balancing SoNaR: IPR versus processing issues in a 500-million-word written Dutch reference corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Julie Roberts. 2020. 2019 Word of the Year is "(My) Pronouns," Word of the Decade is Singular "They" as voted by American Dialect Society. Press Release, American Dialect Society.

Brian A Rood, Sari L Reisner, Francisco I Surace, Jae A Puckett, Meredith R Maroney, and David W Pantalone. 2016. Expecting rejection: Understanding the minority stress experiences of transgender and gender-nonconforming individuals. *Transgender Health*, 1(1):151–164.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Linda Schlief. 2021. Hoe gebruik je de non-binaire voornaamwoorden hen/hun en die/diens? *Tekstbureau Linda Schlief.*

Anneleen Schoen, Chantal van Son, Marieke van Erp, and Hennie van Vliet. 2014. Newsreader document-level annotation guidelines: Dutch. Technical report, VU University.

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 the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Nasim Sobhani, Kinshuk Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in natural language processing: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1121–1131.

Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim. 1999. Corpus-based learning for noun phrase coreference resolution. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational linguistics*, 27(4):521–544.

Susan A Speer and Richard Green. 2007. On passing: The interactional organization of appearance attributions in the psychiatric assessment of transsexual patients. *Out in psychology: Lesbian, gay, bisexual, trans and queer perspectives*, pages 335–368.

Lydia Gabriela Speyer and Erik Schleef. 2019. Processing 'gender-neutral'pronouns: A self-paced reading study of learners of English. *Applied Linguistics*, 40(5):793–815.

Michael Spivak. 1990. *The Joy of TeX, a Gourmet Guide to Typesetting with the AmSTeX Macro Package*. American Mathematical Soc.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore. Association for Computational Linguistics.

Susan Stryker. 2017. *Transgender history: The roots of today's revolution*. Hachette UK.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Freya Terpstra, Lis Dekkers, Sophie Schers, and Sander Dekker. 2021. Trans, intersekse en non-binaire mensen aan het werk in nederland: Een nationaal rapport. *Transgender Netwerk Nederland (TNN)*.

Trans Student Educational Resources. The Gender Unicorn. `http://www.transstudent.org/gender`. Accessed: 2022-09-29.

Transgender Netwerk Nederland. 2016. ZO MAAK JE NA TOILETTEN OOK TAAL GENDERNEUTRAAL. *Transgender Netwerk Nederland Nieuws*.

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 the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Van Dale. 2021. Welk genderneutraal persoonlijk voornaamwoord gebruik je? *Van Dale*.

Rob van der Goot, Hessel Haagsma, and Dicke Oele. 2015. Groref: Rule-based coreference resolution for dutch. `https://kyoto.let.vu.nl/clin26_presentations/paper73.pdf`.

Gertjan van Noord. 2006. At last parsing is now operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42, Leuven, Belgique. ATALA.

Hellen P Vergoossen, Philip Pärnamets, Emma A Renström, and Marie Gustafsson Sendén. 2020a. Are New Gender-Neutral Pronouns Difficult to Process in Reading? The Case of Hen in SWEDISH. *Frontiers in psychology*, 11:2967.

Hellen Petronella Vergoossen, Emma Aurora Renström, Anna Lindqvist, and Marie Gustafsson Sendén. 2020b. Four dimensions of criticism against gender-fair language. *Sex Roles*, 83(5):328–337.

Yannick Versley. 2006. A constraint-based approach to noun phrase coreference resolution in german newspaper text. In *Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Sprache)*, pages 143–150. Deutsche Nationalbibliothek.

Marc Vilain, 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*.

Storm Vogel. 2021. Waarom het woord 'hen' niet moeilijk is. *Lilith*.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10882–10893. PMLR.

Julia Watson, Barend Beekhuizen, and Suzanne Stevenson. 2023. What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6790–6809, Toronto, Canada. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in Dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 92–97, Florence, Italy. Association for Computational Linguistics.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Mandy Woelkens and Thorn de Vries. 2021. *FAQ Gender*. Blossom Books Bold, Vleuten, The Netherlands.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

William Woods. 1972. The lunar sciences natural language information system. *BBN report*.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Lal Zimman. 2018. Transgender Language, Transgender Moment: Toward a Trans Linguistics. In *The Oxford Handbook of Language and Sexuality*. Oxford University Press.

A Coreference resolution evaluation metrics

In this section I describe the common evaluation metrics for coreference resolution. Coreference resolution systems are evaluated by comparing the set of gold clusters G, which are annotated by humans, with the set of hypothesis clusters H, identified by a model. In Figure A.1 (a) are the reference clusters, which include

$$G = \{g_1 = \{a, b, c\}, g_2 = \{e, f\}, g_3 = \{d\}\}$$

and (b) are hypothesis clusters:

$$H1 = \{h_1 = \{a, b\}, h_2 = \{c, e, f\}, h_3 = \{d\}\}$$

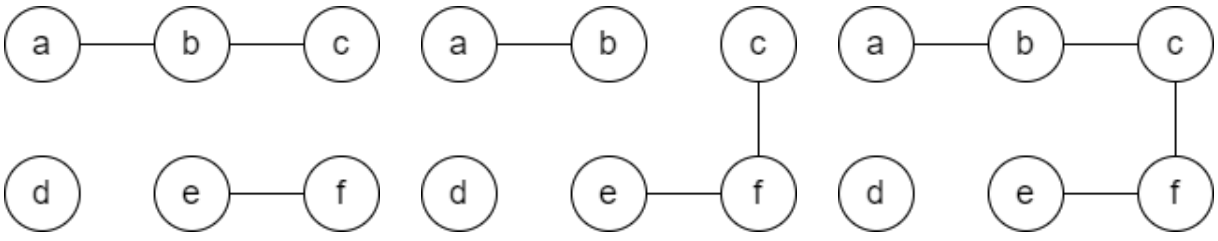
There are five common metrics for evaluation: mention based B^3 (Bagga and Baldwin, 1998), entity based CEAF (Luo, 2005), link based MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011; Luo et al., 2014) and link based entity aware LEA (Moosavi and Strube, 2016). Additionally, the CoNLL score is the average of the MUC, B^3 and CEAF F-scores. Each of these metrics has its own recall r and precision score p, and correspondingly computes its F-score by taking the harmonic mean:

$$F = \frac{2pr}{p + r}$$

I will now describe each of these methods in more detail.

The **MUC score** (Vilain et al., 1995) is based on coreference mention pairs, or *links*. The recall score is the number of links in G that can be found in H divided by the total number of links in G. For hypothesis cluster set (b) in Figure A.1 compared to gold cluster set (a), this gives a recall score of $\frac{2}{3}$, and for hypothesis cluster set (c) the recall is 1. The precision score is the number of links in H that are present in G, divided by the number of links in H. For (b) this gives a precision score of $\frac{2}{3}$, and $\frac{3}{4}$ for (c).

This measure has several drawbacks. Since the MUC score only evaluates mention pairs, it ignores singletons. Moreover, this measure prefers systems with large cluster



(a) Reference clusters R (b) Hypothesis clusters $H1$ (c) Hypothesis clusters $H2$

Figure A.1: An example of reference clusters (a) and hypothesis clusters (b).

outputs, which produce fewer entities. Continuing, it does not sufficiently penalise merging separate entities, because merging two entities only requires one wrong link (Luo, 2005). This problem is illustrated by the example clusters in Figure A.1: although (b) is intuitively better than (c) as it does not confuse two entities as being one, (c) gets higher MUC scores.

B^3 (Bagga and Baldwin, 1998) is a mention based metric: for each individual mention the precision and recall are computed, and the final B^3 score is a weighted average of these scores. Let G_i be a gold cluster, H_j be a hypothesis cluster and w_i a weight value for G_i. The recall is then defined as follows:

$$r = \sum_{i,j} \frac{|G_i \cap H_j|^2}{|G_i| w_i}$$

The precision can be computed the same way, by swapping G and H.

In contrast to MUC, singleton values do contribute to the final scores of this metric. Furthermore contrasting MUC, B^3 does take the sizes of hypothesis clusters into account. However, this metric can assign high scores to arbitrary outputs in some cases: for example a system that classifies all mentions as being part of a single entity cluster will be awarded a perfect recall score (Luo, 2005). An even more serious drawback of B^3 is that duplicate mentions in the hypothesis clusters can lead to arbitrarily large recall scores (Luo and Pradhan, 2016). This is caused by the fact that when the recall score is computed, the intersection with all hypothesis clusters is taken: by repeating a mention over various hypothesis clusters this mention would thus be credited as many times as it is repeated.

A further serious limitation is identified by Moosavi and Strube (2016): after adding incorrect entities to a system's output they find that B^3 scores significantly improve, which they call the *mention identification effect*. This same problem is identified by the CEAF and BLANC scores. They argue that this problem for B^3 is caused by the fact that it evaluates mentions rather than coreference links: any mention in a hypothesis entity is considered to be resolved, whether this mention actually has a coreference relation with the entity or not.

CEAF (Luo, 2005) was introduced to fix the above mentioned problem with B^3 that the recall can become unbounded due to duplicate mentions. To do so CEAF (1) requires a one-to-one entity alignment between H and G and (2) only takes aligned entities into account when computing the final score. This metric uses a similarity measure ϕ to measure the similarity between entities. Continuing, it finds the best one-to-one mapping g^* between the gold and hypothesis entity clusters, using the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1957). Let G^* be the set of response entities in the optimal mapping. Recall is then computes as follows:

$$r = \frac{\sum_{g_i \in G^*} \phi(g_i, g^*(g_i))}{\sum_{g_i \in G} \phi(g_i, g_i)}$$

The precision is computed similarly:

$$p = \frac{\sum_{g_i \in G^*} \phi(g_i, g^*(g_i))}{\sum_{h_i \in H} \phi(h_i, h_i)}$$

There are two variations of CEAF, which differ in their similarity measure. Firstly, in mention-based CEAF, CEAF_m, the similarity is measured as the number of mentions that G and H have in common:

$$\phi(G, H) = |G \cap H|$$

In the second variation, entity-based CEAF_e, the similarity score is measured as the F-score between two entities:

$$\phi(G, H) = \frac{2|G \cap H|}{|G| + |H|}$$

A limitation of CEAF is that all predictions that fall outside the alignment are ignored, including correct predictions (Denis and Baldridge, 2009). A second problem is that entities are not weighted by their size: a system that fails to recognise a singleton is punished the same way as one that misses a large cluster (Stoyanov et al., 2009). Thirdly, as mentioned above, CEAF also suffers from the mention identification problem. Moosavi and Strube (2016) explain this is caused by the similarity measures that, similar to B^3, only evaluate whether entities R_i and H_j have common mentions, not whether these mentions actually form coreference relations with the entity.

BLANC (Recasens and Hovy, 2011; Luo et al., 2014) is a link-based metric, which was designed to overcome MUC's limitation of not evaluating singletons. BLANC solves this problem by evaluating both links within entities and links outside of entities. Let C_g and C_h be the coreference links in the gold and hypothesis entities respectively. For example, take (a) in Figure A.1 as G and (b) as H. This gives:

$$C_g = \{(ab), (ac), (bc), (ef)\}$$

$$C_h = \{(ab), (ce), (cf), (ef)\}$$

Continuing, let N_g and N_h be the non-reference link in these entities:

$$N_g = \{(ad), (ae), (af), (bd), (be), (bf), (cd), (ce), (cf), (de), (df)\}$$

$$N_h = \{(ac), (ad), (ae), (af), (bc), (bd), (be), (bf), (cd), (de), (df)\}$$

The performance scores are then computed for the coreference and non-coreference links separately, after which their averages are taken for the final score. So the precision and recall scores for the coreference links can be computed as:

$$R_c = \frac{|C_g \cap C_h|}{|C_g|}, P_c = \frac{|C_g \cap C_h|}{|C_h|}$$

And the non-coreference link precision and recall can be calculated similarly by swapping the sets in the computations. Then the F-scores for both sets (F_c and F_n) are computed

through the harmonic mean and the final BLANC score can be computed through:

$$BLANC = \frac{F_c + F_n}{2}$$

This measure is most strongly affected by the mention identification effect (Moosavi and Strube, 2016), because it considers non-coreferent relations: if the number of gold mentions in the hypothesis entities increases, the number of identified non-coreference links automatically increases as well, resulting in a higher performance score, whether these gold mentions are correctly resolved or not.

LEA (Moosavi and Strube, 2016) is a link-based and entity aware metric, designed to fix the limitations of its predecessors. This metric considers the importance of each entity through an adaptable measure, for which the authors use the size of the enitity e, i.e. $importance(e) = |e|$. Continuing, they use the following *resolution score*, which can be interpreted as the portion of coreference links that are correctly resolved:

$$resolution - score(G_i) = \sum_{H_j \in H} \frac{link(G_i \cap H_j)}{link(G_i)}$$

In order to deal with singletons, these entities are considered to have self-links: a link that connects a mention to itself. Only singletons have self-links and therefore if G_i is a singleton, $link(G_i \cap H_j) = 1$ only if H_j is a singleton with the same mention as G_i. The recall is computed as:

$$r = \frac{\sum_{G_i \in G}(|G_i| \cdot \sum_{H_j \in H} \frac{link(G_i \cap H_j)}{link(G_i)})}{\sum_{G_z \in G} |G_z|}$$

And similarly the precision is computed through:

$$p = \frac{\sum_{H_i \in H}(|H_i| \cdot \sum_{G_j \in G} \frac{link(H_i \cap G_j)}{link(H_i)})}{\sum_{H_z \in H} |H_z|}$$

LEA does not suffer from the mention identification effect, since it (i) does not consider non-coreferent links and (ii) considers resolved coreference links rather than resolved mentions. It further improves over CEAF by (a) considering all coreference relations, rather than only those within the alignment, and (b) considering the importance of additional or missing entities.

CoNLL SCORE Finally, for the CoNLL 2012 shared task (Pradhan et al., 2012), the CoNLL SCORE was introduced, which is the average of the MUC, B^3 and CEAF F-scores. Moosavi and Strube (2016) reevaluate the final CoNLL 2012 ranking of the shared task with LEA and find a different ranking using their measure, which they explain to potentially be caused by the mention identification effect. They further argue that using a single reliable measure is preferable over an average score because it allows for significance testing and provides meaningful recall and precision scores.

B Gendered nouns rewriting rules

Gendered noun	Gender-neutral noun
tante	familielid*
oom	familielid*
jongen	kind
meisje	kind
man	persoon
vrouw	persoon
mannen	personen
vrouwen	personen
broer	familielid*
zus	familielid*
broertje	familielid*
zusje	familielid*
broertjes	familieleden*
zusjes	familieleden*
broers	familieleden*
zussen	familieleden*
meid	persoon
vader	ouder
moeder	ouder
vaders	ouders
moeders	ouders
zoon	kind
zonen	kinderen
dochter	kind
dochters	kinderen
nicht	familielid*
nichtje	familielid*
nichtjes	familieleden*
nichten	familieleden*
neef	familielid*
neefje	familielid*
neefjes	familieleden*
kleindochter	kleinkind
kleinzoon	kleinkind
kleindochters	kleinkinderen
kleinzonen	kleinkinderen
oma	grootouder
opa	grootouder
grootmoeder	grootouder
grootvader	grootouder
dame	persoon
heer	persoon
dames	personen
heren	personen
koning	staatshoofd
koningin	staatshoofd

Table B.1: Rewriting rules for gendered Dutch nouns to a gender-neutral version of this word. Not all Dutch words have a gender-neutral alternative however. * marks difficult cases, for which some meaning is lost in translation.

Gendered noun	Gender-neutral noun
koningen	staatshoofden
koninginnen	staatshoofden
mevrouw	persoon*
meneer	persoon*
jongedame	jongere*
jongeman	jongere*
politieman	politieagent
politievrouw	politieagent
brandweerman	brandweermens
brandweervrouw	brandweermens
prinses	edele*
prins	edele*
prinsessen	edelen*
prinsen	edelen*
kroonprins	troonopvolger
kroonprinses	troonopvolger
schrijver	auteur
schrijfster	auteur
juf	leerkracht
meester	leerkracht
leraar	leerkracht
lerares	leerkracht
bruid	jonggehuwde
bruidegom	jonggehuwde
tovenaar	magiër
heks	magiër
stiefvader	stiefouder
stiefmoeder	stiefouder
stiefzoon	stiefkind
stiefdochter	stiefkind
weduwe	nabestaande*
weduwnaar	nabestaande*
kok	chef
kokkin	chef
kunstenaar	artiest
kunstenaares	artiest
vriend	maat*
vriendin	maat*
vriendje	partner*
vriendinnetje	partner*

Table B.2: Rewriting rules for gendered Dutch nouns to a gender-neutral version of this word. Not all Dutch words have a gender-neutral alternative however. * marks difficult cases, for which some meaning is lost in translation.

C Data transformation quality check subset

Filename	Data version	Split	Number of words
dpc-ind-001650-nl-sen	zij	test	1,454
wiki-859	die	test	515
WR-P-E-H-0000000052_0	hij	dev	2,483
wiki-7355	hij	dev	540
dpc-bal-001237-nl-sen	hen	train	968
wiki-295	hen	train	1,388
wiki-572	delex	train	355
WS-U-E-A-0000000038	delex	train	1,600
dpc-med-000677-nl-sen	gender-neutral	train	2,2326
dpc-cam-001020-nl-sen	gender-neutral	train	1,411
Total			**12,584**

Table C.1: Overview of the documents that are included in the subset that I use for a quality check of the data transformation algorithm in Section 3.3.

D Gender-neutral names list

Moos
Bo
Lou
Charlie
Jackie
Noa
Sam
Robin
Lux
Nicky
Charly
Jules
Yaniek
Sydney
Pascal
Jos
Marijn
Ocean
Sky
Skye
River
Rowan
René
Renée
Mickey
Jip
Jaimy
Jamie
Luca
Bobby
Dominic
Dominique
Harper
Sasha
Sascha
Revi
Sil
Rho
Phlox
Ihme

Madu
Zilver
Camille
Harley
Jazz
Bailey
Alex
Nova
Noé
Jayden
Roan
Ezra
Novi
Luka
Teddy
Izzy
Riv
Micha
Juda
Eden
Jona
Billie
Parker
Hunter
Ash
Arbor
Everest
Jett
Moss
Oakley
Phoenix
Bowie
Haven
Kit
London
Lyric
Reese