*Trans*forming Dutch: Debiasing Dutch Coreference Resolution Systems for Non-binary Pronouns

Goya van Boven
goya_jackie@live.nl
Utrecht University
Utrecht, the Netherlands

Yupei Du
y.du@uu.nl
Utrecht University
Utrecht, the Netherlands

Dong Nguyen
d.p.nguyen@uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

Gender-neutral pronouns are increasingly being introduced across Western languages. Recent evaluations have however demonstrated that English NLP systems are unable to correctly process gender-neutral pronouns, with the risk of erasing and misgendering non-binary individuals. This paper examines a Dutch coreference resolution system's performance on gender-neutral pronouns, specifically *hen* and *die*. In Dutch, these pronouns were only introduced in 2016, compared to the longstanding existence of singular *they* in English. We additionally compare two debiasing techniques for coreference resolution systems in non-binary contexts: Counterfactual Data Augmentation (CDA) and delexicalisation. Moreover, because pronoun performance can be hard to interpret from a general evaluation metric like LEA, we introduce an innovative evaluation metric, the *pronoun score*, which directly represents the portion of correctly processed pronouns. Our results reveal diminished performance on gender-neutral pronouns compared to gendered counterparts. Nevertheless, although delexicalisation fails to yield improvements, CDA substantially reduces the performance gap between gendered and gender-neutral pronouns. We further show that CDA remains effective in low-resource settings, in which a limited set of debiasing documents is used. This efficacy extends to previously unseen neopronouns, which are currently infrequently used but may gain popularity in the future, underscoring the viability of effective debiasing with minimal resources and low computational costs.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing**; • **Social and professional topics** → *Gender*;

KEYWORDS

coreference resolution, gender bias, gender-neutral pronouns, neopronouns, Dutch, debiasing

ACM Reference Format:
Goya van Boven, Yupei Du, and Dong Nguyen. 2024. *Trans*forming Dutch: Debiasing Dutch Coreference Resolution Systems for Non-binary Pronouns. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3659049

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659049

1 INTRODUCTION

Gender-neutral pronouns are increasingly introduced and popularised across Western languages, as suitable alternatives to traditional gendered pronouns for non-binary individuals. The Swedish gender-neutral pronoun *hen* was politically introduced in 2013 [21], the Dutch *hen/die* was democratically chosen by the Transgender community in 2016 [45] and while English has long known singular *they*, the set of *neopronouns* such as *ze* and *thon* is continuously growing [30]. However, the majority of work in natural language processing (NLP) considers gender as binary and immutable [8, 14], thereby excluding transgender individuals, who do not identify with the gender they were assigned at birth. The term *transgender* includes both people with a binary transgender identity (such as transgender women) and people with a non-binary transgender identity. *Non-binary* individuals do not conform to the traditional Western binary categorisation of male or female, identifying for instance as both female and male, as neither or their gender might fluctuate [40].

Within Western societies, transgender people face various forms of discrimination and marginalisation. They experience high rates of unemployment, homelessness, abuse and poverty, and frequently experience bullying and discrimination at the workplace [26, 44]. Furthermore, transgender people often encounter significant barriers in accessing essential institutions, such as healthcare services and the legal system [53].

Dev et al. [13] point out how NLP models can contribute to the marginalisation of transgender people by perpetuating trans-exclusive practices, highlighting the dangers of *erasing* and *misgendering* non-binary individuals. 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 [1].

Recent works have started to investigate non-binary gender biases in NLP systems. In this work we adopt the definition of *bias* provided by Friedman and Nissenbaum: "computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" [18]. Non-binary gender bias evaluations of NLP systems consider large language models [6, 13, 24, 35, 38, 48], machine translation [10, 20, 31], POS-tagging [4], and coreference resolution [2, 6, 8]. In this study we focus on evaluating and debiasing a Dutch coreference resolution system in processing gender-neutral pronouns. Coreference resolution is the task of identifying expressions that refer to the same entity.

We focus on this fundamental NLP task, because any structural mistakes for non-binary individuals at the coreference resolution level, such as failing to recognise their pronouns — and thereby failing to extract information about these individuals — can lead to their erasure in downstream applications.

While earlier studies have evaluated the performance of English coreference resolution systems on gender-neutral pronouns [2, 8], we are the first to perform such an evaluation for a Dutch system, zooming in on the pronouns *hen* and *die*. The Dutch context differs from the English one because (a) Dutch gender-neutral pronouns are less frequent than English singular *they*; (b) in Dutch many nouns are gender-specific, lacking gender-neutral alternatives (e.g. no term like *sibling* exists); and (c) there are generally fewer NLP resources available for Dutch than for English.

We make the following contributions: 1) We systematically evaluate a Dutch coreference resolution system on its performance on gender-neutral pronouns. Our results show that the model currently performs worse on gender-neutral pronouns, compared to gendered counterparts. 2) We address the limitations of existing evaluation metrics, as these do not provide sufficient insight into the handling of pronouns. To better evaluate systems we propose a new metric, called the *pronoun score*, which directly reflects the percentage of correctly resolved pronouns. 3) We experiment with two different debiasing methods, and find that while delexicalisation [30] does not improve the performance on gender-neutral pronouns in our setup, Counterfactual Data Augmentation does substantially improve the performance on these pronouns. Importantly, our successful debiasing results extend to (a) previously unseen neopronouns, and (b) low-resource conditions, which use just a handful of debiasing documents.

This paper is structured as follows: we provide a theoretical background in Section 2 and describe related work in Section 3. We continue to describe the data (Section 4), model (Section 5) and pronoun score (Section 6). We present our experiments in Section 7 and finally provide a discussion in Section 8. The code used for this project can be found at https://github.com/gvanboven/Transforming_Dutch.

2 BACKGROUND

2.1 Gender in Dutch

Traditional gender manifestation. In Dutch, gender is expressed in nouns and third-person singular pronouns. There is no gendered verb agreement or case inflection. Table 1 summarises the traditional third-person pronouns for animate entities: singular pronouns distinguish between feminine and masculine while there is no gender distinction in plural. Nouns referring to occupations and family members are usually gendered. For many occupations, the masculine form is the root (e.g. *eigenaar* (*owner*), *schrijver* (*writer*)) and the feminine form adds a suffix (e.g. *eigenares*, *schrijfster*) [19]. In other cases the male form is used for all genders (e.g. *professor* (*professor*)). For some occupational terms gender-neutral alternatives exist, e.g. replacing *lerares* (*female teacher*) and *leraar* (*male teacher*) with *leerkracht* (*teacher*), but for many occupations this is not the case. Similarly, most words describing relatives only have a feminine and masculine version: e.g. no term like *cousin* exists in Dutch, only providing the options of male *neef* and female *nicht*.

Table 1: Overview of the traditional Dutch third-person pronouns for animate entities. The neuter singular form *het* is excluded because this form is only used for inanimate entities.

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

This is a problem for non-binary people, since there is no alternative that matches their gender identity.

Gender-neutral pronouns. In 2016, Transgender Netwerk Nederland organised a vote to determine the Dutch gender-neutral pronouns, in which 500 community members participated. Here, *hen/hen/hun* was favoured over *die/die/diens* and neopronouns *dee/dem/dijr* [45]. However, as of 2024, both *hen* and *die* are increasingly being adopted by non-binary people [3, 17]. Additionally, a broader set of neopronouns has been proposed, including *zhij* and *ij* [23, 25], 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 prepositions. 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: "Noa geeft *hun/diens* studieboeken weg omdat *hen/die* **is** afgestudeerd" (Noa gives *their* study books away because *they* **have** graduated); note that singular **is** is used in Dutch while plural **have** is used in English.

2.2 Coreference resolution

Coreference resolution 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*. Within a cluster, the mentions that precede a certain mention are called its *antecedents* while its later mentions are *anaphors* or *anaphoric*. The following sentence shows an example, in which mentions are marked in brackets, and mentions with the same colour refer to the same entity: "[Sam Smith] is a famous singer. [They] collaborated with [Kim Petras] recently." Coreference resolution consists of two subtasks, which end-to-end systems perform simultaneously [32, 33]: (1) mention detection, i.e. identifying the spans of referring expressions and (2) identifying the coreference links between the mentions.

3 RELATED WORK

Non-binary gender bias evaluations. Recent works evaluate non-binary gender bias in English language models [13, 24, 38, 48], and in multi-lingual evaluations [6, 35]. In downstream tasks, non-binary

gender bias evaluations are performed for machine translation [31], NER [29], POS-tagging [4] and abusive language detection [43] systems. In coreference resolution, the WinoNB dataset [2] is designed to test systems' ability to disambiguate singular and plural *they*. Moreover, the GICoref dataset [7] contains naturally occurring data of "gender-related phenomena"[7], and a relatively balanced distribution of *he*, *she*, *they* and neopronouns. Both datasets reveal poor model performances on gender-neutral pronouns. Finally, Brandl et al. [6] evaluate a Danish coreference resolution model on gender-neutral pronouns. They create a gender-neutral version of a regular coreference resolution corpus by replacing gendered pronouns with gender-neutral ones. They report a small CoNLL score performance drop on the gender-neutral data. However, from their results it does not become entirely clear *how many* more gender-neutral pronouns are incorrectly resolved, as the CoNLL-score does not directly reflect this.

Debiasing. To our best knowledge, only two studies consider debiasing NLP systems for non-binary gender bias. First, Hossain et al. [24] aim to improve language model performance in incorporating the declared preferred pronouns of an individual. They explore few-shot in-context learning using explicit examples and note an improvement in performance. However, the improvement plateaus rapidly, falling short of achieving comparable accuracy levels to those observed for gendered pronouns. Second, Björklund and Devinney [4] aim to improve the POS-tagging performance on the Swedish gender-neutral pronoun *hen*. They first augment the training data with semi-synthetic data by replacing gendered pronouns with gender-neutral ones. Then, they 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. We consider their study the most similar to the current work as (a) both works involve debiasing downstream tasks and (b) their debiasing method is similar to our application of CDA. However, besides considering a different task and language, we make the additional contributions of (1) evaluating a "continual fine-tuning" debiasing configuration, besides fine-tuning from scratch; (2) additionally evaluating the delexicalisation method; and (3) investigating the effect of the debiasing methods on previously unseen pronouns.

4 DATA

In this section, we introduce and analyse the data (Section 4.1); describe the data preprocessing steps (Section 4.2); and finally describe the transformation steps for inserting gender-neutral pronouns into the corpus (Section 4.3).

4.1 Data analysis

We use the 1M-token SoNaR-1 corpus [37, 41], because this is the largest Dutch corpus annotated for coreference resolution to date. SoNaR-1 consists of 861 documents from various domains, e.g. magazines, Wikipedia articles, brochures, websites, legal texts, autocues and press releases. The coreference relations were manually annotated based on the COREA guidelines [22]. The corpus also contains manually checked annotations for syntactic dependency trees, spatio-temporal relations, semantic roles and named entities.

Table 2: Frequency of gender-neutral pronouns and neopronouns in the SoNaR-1 corpus. *Die* appears 1,995 times as a demonstrative pronoun, 5,268 times as a relative pronoun, and 19 times with an alternative label. Its possessive form, *diens*, appears 35 times as a demonstrative pronoun. The 290 occurrences of *hen* function as a third-person plural object personal pronoun. Its possessive counterpart, *hun*, appears 1,865 times as a third-person plural possessive pronoun. *Vij* appears once and *zeer* (also connoting *very* or *sore*) occurs 295 times as an adverb. The remaining neopronouns do not feature in the dataset. None of the considered gender-neutral pronouns and neopronouns appear as third-person singular pronouns in the corpus.

Pronoun	Frequency	Pronoun	Frequency
die	7,282	*zeer*	295
diens	35	*vij*	1
hen	290	*dee, dem, dijr, dij, dem,*	
hun	1,865	*dijr, nij, ner, nijr, vijn*	0
		vijns, zhij, zhaar, zem	

Personal and possessive pronouns constitute 2.9% of all tokens in the corpus. Of these pronouns, third-person singular pronouns are the most prevalent, constituting 59.2%. A striking gender imbalance can be observed: **79.1% of all third-person pronouns are masculine**. Table 2 displays the overall frequency of gender-neutral pronouns and neopronouns. While some of the types frequently appear in the corpus, further analysis of the POS-tags reveals that none of these types are actually used as third-person singular pronouns within the corpus, indicating that **the corpus does not include gender-neutral pronouns or neopronouns**.

4.2 Data preprocessing

We use the genre-balanced 70/15/15 division by Poot and van Cranenburgh [39] to divide the documents over train/dev/test splits. Moreover, we remove all singleton annotations from SoNaR-1. The reason for this is two-fold: (1) While the SoNaR-1 corpus includes singleton annotations, coreference resolution systems are frequently trained and evaluated on corpora without singleton annotations [5, 27, 33, 34]; (2) Pronouns typically refer to proper nouns and thus seldom exist as singletons. Recognising singletons thus appears to be of limited relevance for the current study.

4.3 Construction of pronoun-specific data

To compare the performance on different pronouns, we create four *pronoun-specific* versions of the test split, in which all third-person pronouns are replaced by a specific pronoun set: 1. *hij/hem/zijn* (masculine), 2. *zij/haar/haar* (feminine), 3. *hen/hen/hun* (gender-neutral) and 4. *die/hen/diens* (gender-neutral). Examples can be found in Table 3. We use a rule-based rewriting algorithm based on Zhao et al. [52]. The algorithm consists of three steps. The principle step is swapping pronouns. Moreover, names are anonymised and gendered nouns are replaced by gender-neutral nouns. These operations ensure that the performance on the different *pronoun-specific* test sets can be fairly compared, as they avoid confounding gender

Table 3: Example sentence in the SoNaR-1 dataset, before and after transforming it into different pronoun settings. Words that are changed between the versions are marked in bold. Here, the noun replacement *vrouw* (*wife* in this context, but also means *woman*)→*persoon (person)* changes the meaning of the noun in this context. We consider this replacement incorrect. We manually evaluated a subset of the replacements, and only 1.53% of the replacements was incorrect.

Dataset	Gender	Sentence
Original	Masculine	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 hij	Masculine	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.*
Pronoun-specific zij	Feminine	Na **haar** herstel vindt **zij haar persoon** en **haar ouder** terug in Folkestone.
Pronoun-specific hen	Gender-neutral	Na **hun** herstel vindt **hen hun persoon** en **hun ouder** terug in Folkestone.
Pronoun-specific die	Gender-neutral	Na **diens** herstel vindt **die diens persoon** en **diens ouder** terug in Folkestone.

effects in case the gender associations of a name or noun and a pronoun no longer match after pronoun swapping, as for example in the phrase *Vader en haar kind* (*Father and her child*).

1. Swapping pronouns We recognise third-person singular pronouns by their POS-tag[1], and replace them according to the rules stipulated for the targeted dataset version (e.g., *hij* → *hen* for the *pronoun-specific hen* test set).

2. Name anonymisation Following Zhao et al., we recognise names using named entity annotations,[2] considering all tokens with a PER (person) tag. 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:[3]
Jan Jansen is op vrijdag vrij omdat Jan dan voetbalt.
(*Jan Jansen is free on Friday because Jan plays football.*) →
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.*)

3. Replacing gendered nouns We create a Dutch list (Appendix A) of gendered nouns and their gender-neutral counterparts (e.g. *moeder* (*mother*) → *ouder* (*parent*)), using the English list by Zhao et al. [52] as a basis. To ensure the quality of replacements, a panel of six individuals participated in reviewing the list and contributed their suggestions.[4]

Limitations. For some nouns it proved challenging to identify gender-neutral replacements, as numerous Dutch words exclusively have gendered forms (e.g. *nicht* (*niece*) lacks an alternative akin to *cousin*).[5] Other gendered terms have multiple meanings (e.g. *vrouw* means both *woman* and *wife*), forcing the selection of a one interpretation for the replacement, as illustrated in Table 3. Moreover, sometimes changing a noun in Dutch necessitates changing the determiner (e.g. *de dochter* (*the daughter*) → *het kind* (*the child*)), but we did not extend the transformation algorithm to feature this ability. We manually review a subset comprising 12,584 tokens, balanced over the data versions and splits. Out of 1,111 replacements, only 17 are incorrect, giving an error rate of 1.53%. We therefore consider the transformed data to be of good quality.

[1]This is necessary because some Dutch pronouns have identical lexical forms but distinct grammatical functions, such as possessive or personal object *haar* and third-person singular or plural *zij*.
[2]We use the manually checked named entity annotations included in the SoNaR-1 corpus for this.
[3]This step may introduce some processing difficulties to coreference models, because these models are usually 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.
[4]Among these individuals, three use *she/her* pronouns, two use *he/him* pronouns, and one uses *he/they* pronouns. The recruitment of participants was facilitated through our personal network.
[5]In such instances, we settled for a gender-neutral hypernym of the term of interest, such as *familielid* (*family member*), sacrificing a substantial portion of the meaning of the original term. Such instances are marked in Appendix A.

5 MODEL

In alignment with prior debiasing studies, which only consider debiasing neural coreference resolution systems [e.g. 52], we focus on debiasing a neural Dutch model.[6] We select the wl-coref [16] model, which is originally trained on English data. We chose this model because (a) it achieves a competitive performance, (b) it has a low complexity and (c) its base models are available in Dutch.

5.1 Architecture

The wl-coref model [16] architecture consecutively performs the following two steps: (1) Predicting the antecedent for each word individually, or predicting that the word does not have an antecedent. 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, the head of the mention *their roommate* in Figure 1 is *roommate*, as the head of this word, *asked*, is outside of the mention span. (2) Predicting the full mention span boundaries from the mention heads, ultimately culminating in the final coreference predictions for complete mentions.

5.2 Training

We fine-tune the wl-coref architecture on the SoNaR-1 corpus, without making any changes to the core modules. Following Dobrovolskii [16], we train the models for 20 epochs and report the performance of the epoch that performs best on the development data. We make three minor adjustments to the setup:

(1) We change the evaluation metric from the CoNLL score to the LEA score, as the CoNLL metric has been demonstrated to be flawed [36]. For the wl-coref model, the performance scores can be computed both at the word-level (before span

[6]We therefore exclude the rule-based Dutch dutchcoref [46] model and the hybrid Dutch model by van Cranenburgh et al. [47].

Table 4: LEA performances of the wl-coref model on the development set of the SoNaR-1 corpus using three different base models.

Base model	Dev F1
robBERT [12]	45.5
mBERT-base [15]	47.0
XLM-RoBERTa-base [11]	**52.4**

Table 5: Average LEA performance scores of the wl-coref model with XLM-RoBERTa-base on the SoNaR-1 development and test set, using five random seeds.

Data	Precision	Recall	F1
Dev	53.0 ($\sigma = 2.3$)	57.9 ($\sigma = 3.1$)	55.3 ($\sigma = 0.2$)
Test	55.5 ($\sigma = 2.3$)	55.8 ($\sigma = 2.7$)	55.6 ($\sigma = 0.5$)

boundary prediction) and at the span-level (after span boundary prediction). We report the span-level performance, because this represents the performance on the full coreference resolution task.

(2) While wl-coref uses speaker and genre information, the SoNaR-1 corpus does not contain speaker information. Therefore, we use the same speaker value (zero) for all instances. We also experiment with taking out the speaker component entirely, but this does not improve the performance.

(3) Although genre information is available for SoNaR-1, we follow Poot and van Cranenburgh [39] in always using the same genre value. We leave exploring the effect of including genre information to future work.

We use the Hugging Face Transformers [50] implementations to compare three base models: the Dutch robBERT model [12], and the multilingual mBERT [15] and XLM-RoBERTa [11] models, all in their *base* versions. Table 4 shows the results on the SoNaR-1 dev set, using the same hyperparameters as Dobrovolskii [16]. As XLM-RoBERTa obtains the best performance (F1-score = 52.4), we use this model in our main experiments. Furthermore, we execute a hyperparameter search (full results in Appendix B) for two hyperparameters: the learning rate (best value = $5e^{-4}$) and the BERT learning rate (best value = $3e^{-5}$). We also experiment with lowering k and using a different learning rate schedule than the standard linear one, but these adaptations do not improve the performance.

We report the final performance scores in Table 5, as the average of five random seeds. The model obtains a test F1-score of 55.6. For a comparison between wl-coref and other Dutch coreference resolution models, refer to Appendix C. Throughout the rest of this study, we report the main results as the average of five random seeds.

6 PRONOUN SCORE

We use link-based entity aware LEA [36] as our evaluation metric. Because the *pronoun-specific* test sets only differ in third-person pronouns, any performance difference between these sets can be attributed to third-person pronouns. However, if the LEA F1-score is e.g. one point lower for *hen* than for *hij* pronouns, it is not directly clear how many more *hen* pronouns are incorrectly resolved. Per illustration, Figure 1a shows a sentence with gold annotations, and 1b shows an example prediction. The prediction is correct, except the pronoun `[they]` is not recognised as a mention, and is thus not considered as part of the *Raven* cluster. This prediction results in a LEA F1-score of $\frac{6}{7}$.[7] From this score alone, it is hard to interpret that $\frac{1}{2}$ of the third-person pronouns is correctly resolved. To get a more direct insight into the model's ability to process pronouns, we introduce the **pronoun score**, complementary to the LEA score, that represents *the percentage of pronouns for which at least one correct antecedent is identified.* As we focus on third-person singular pronouns in this work, we only consider pronouns in this category, and define the metric as:

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

We illustrate this score with the example in Figure 1. We can identify the following gold and predicted antecedents for the third-person pronouns *they* and *their*:

$$gold\_ants(they) = \{Raven\}$$
$$predicted\_ants(they) = \{\}$$
$$gold\_ants(their) = \{they, Raven\}$$
$$predicted\_ants(their) = \{Raven\}$$

Then the pronoun score is computed as follows:

$$gold\_ants(they) \cap predicted\_ants(they) >= 1$$
$$= \{Raven\} \cap \{\} >= 1 = 0 >= 1 = 0$$
$$gold\_ants(their) \cap predicted\_ants(their) >= 1$$
$$= \{they, Raven\} \cap \{Raven\} >= 1 = 1 >= 1 = 1$$
$$pronoun\_score = \frac{0+1}{2} \cdot 100\% = \frac{1}{2} \cdot 100\% = 50\%$$

The pronoun score thus directly reflects the amount of correctly resolved third-person pronouns.

Design considerations. Rather than only considering the one antecedent that is *directly* predicted by the model, we consider *all* antecedents in the predicted cluster in the evaluation, and consider at least one correct antecedent in the prediction cluster to be sufficient. We prefer this more relaxed configuration because prior evaluations show poor performances on gender-neutral pronouns [2, 8], indicating the difficulty of the task. However, the metric is highly adaptable, and can be made stricter for more straightforward tasks. Moreover, we prefer considering one correct antecedent as sufficient, over an alternative such as requiring all pronoun antecedents to be correct, because otherwise the model might be punished twice for a single mistake.[8] The LEA score already provides a holistic view of the model's performance, so we prefer the

[7] For an explanation of how the LEA score is computed, refer to Moosavi and Strube [36]

[8] This is illustrated by the example in Figure 1. In the predictions, *they* is missed as a mention, and therefore *their* also misses one of its correct antecedents. Requiring all the correct antecedent to be found results in a pronoun score of zero for this example, punishing the mistake for *they* twice. We consider this outcome undesirable.

[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 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.

pronoun score to complement the LEA score by zooming in on the pronoun alone.[9]

A potential objection against considering antecedents is that the first mention of a cluster is always excluded from the evaluation. However, pronouns are typically used to replace names or proper nouns, and thus rarely appear as the first mention of a cluster. This objection thus does not appear to be relevant for pronouns.

7 EXPERIMENTS

In Section 7.1 we compare between the performance on gendered pronouns and gender-neutral pronouns. In Section 7.2 we evaluate the effectiveness of two debiasing techniques: Counterfactual Data Augmentation (CDA) and delexicalisation. Next, in Section 7.3 we explore CDA debiasing in low resource conditions. Finally we evaluate the performance of the original and the debiased models on previously unseen neopronouns in Section 7.4.

7.1 Gender-neutral pronoun evaluation experiment

7.1.1 Setup. We compare the wl-coref model's performance on gendered and gender-neutral pronouns, by comparing its performance on the *pronoun-specific* versions of the test set (Section 4.3). Besides changing pronouns, the transformation performed to create the *pronoun-specific* test sets involves obscuring gender clues through (a) replacing gendered nouns by gender-neutral nouns, and (b) anonymising names. Prior to performing the experiment, we compute how these two transformations affect the model's performance, to isolate the impact of changing the pronouns in the *pronoun-specific* data sets (Table 6, *baseline*). We observe a drop of 4.16 points in F1-score compared to the original test data.

7.1.2 Results. Table 6 reports the performances on the *pronoun-specific* test sets. We now discuss the main observations.

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 4.1).

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 [28, 42, 49].[10]

The performance is worse on gender-neutral pronouns. The pronoun score drops strongly for *hen* pronouns (-12.51 percentage points) and decreases even further for *die* (-30.87 percentage points).

***Hen* pronouns are better resolved than *die* pronouns.** The high standard deviations for *die* additionally indicate an unstable resolution. A potential reason for this 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* (see Table 2).

7.2 Debiasing experiment

7.2.1 Setup. In this experiment we compare two debiasing methods to improve the performance on gender-neutral pronouns. The first debiasing method is **Counterfactual Data Augmentation** (CDA), which involves replacing existing pronouns in the training data with the pronouns of interest [51, 52]. We create a *gender-neutral* version of the data, following the algorithm described in Section 4.3, in which all third-person singular pronouns are substituted by gender-neutral pronouns: *hen* in 50% of the documents, and *die* in the remaining 50%. 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. Given the favourable outcomes demonstrated by CDA in mitigating binary-gender bias within coreference resolution systems [51, 52], we expect this method to be effective.

The second debiasing method is **delexicalisation** [30]. This method is applied by training the model on a *delexicalised* version of the data, wherein all third-person pronouns are replaced by their corresponding POS-tag. We adapt Lauscher et al. [30]'s approach slightly by replacing the POS-tags, which only distinguish between personal and possessive pronouns, by syntactical tags for subjects (<SUBJ>), objects (<OBJ>) and possessives (<POSS>). This adaptation is made to distinguish between various grammatical functions, given that the lexical forms adopted by subjects and objects differ in Dutch. The rationale 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. This method has not previously been tested in

[9] For the same reason, we decide not to incorporate other error types (e.g., false positives) in the pronoun score: this would make the metric less interpretable, and the LEA score already has a recall and precision component.

[10] A difference with English is that in Dutch, the feminine third-person singular pronoun *zij* (250 occurrences in the corpus) is also used as a third-person plural pronoun (481 occurrences). This might increase the model's familiarity with the type, and boost it's recognition as a pronoun.

Table 6: Gender-neutral pronoun evaluation experiment: LEA and pronoun scores on the *pronoun-specific* test sets, as the average of five random seeds. The *baseline* performance refers to the performance on the version of the test set in which gender clues are removed, but the pronouns remain unchanged. The model's performance is lower on gender-neutral than on gendered pronouns.

Data	Precision	Recall	F1	Δ F1 baseline	Pronoun score	Δ with *hij*
Baseline	53.58 (σ=2.24)	49.59 (σ=2.62)	51.41 (σ=0.44)	-		
Hij (masculine)	52.23 (σ=2.30)	49.66 (σ=2.76)	51.29 (σ=0.42)	-0.12	88.36% (σ =0.89)	-
Zij (feminine)	53.18 (σ=2.33)	48.73 (σ=2.56)	50.77 (σ=0.37)	-0.64	86.65% (σ =1.23)	-1.71
Hen (gender-neutral)	53.29 (σ=2.56)	45.82 (σ=3.16)	49.14 (σ=0.68)	-2.27	75.85% (σ =2.93)	-12.51
Die (gender-neutral)	52.55 (σ=1.46)	44.94 (σ=2.24)	48.36 (σ=0.44)	-3.05	57.49% (σ =6.55)	-30.87

a similar setup. Lauscher et al. [30], who introduce this methodology, conduct 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. They additionally try both training and testing on delexicalised data, which results in a satisfactory performance. However, this experimental design does not faithfully simulate a scenario in which a model is debiased through delexicalised data, but deployed on naturally occurring data, which includes pronouns in their lexical forms.

We evaluate both debiasing methods in two conditions: (i) **Fine-tuning the model from scratch** on the respective debiasing dataset, and (ii) **continual fine-tuning** the original wl-coref model, initially trained on the regular SoNaR-1 data, with the respective debiasing dataset. Given that continual fine-tuning is computationally less demanding compared to fine-tuning from scratch,[11] it would be a preferable debiasing approach, provided it achieves a satisfactory performance. For consistency, the same hyperparameters are employed as for the regular model. Debiased models fine-tuned from scratch are trained for 20 epochs, while the continual fine-tuned models are trained for 10 epochs. The models are evaluated on the *pronoun-specific* test sets. The debiasing performance is measured as the difference between 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.

7.2.2 Results. Table 7 displays the performance scores after (a) fine-tuning from scratch and (b) continual fine-tuning the original wl-coref model, using the two debiasing techniques. Here we discuss the main observations.

Delexicalisation does not successfully debias the model. The pronoun scores remain low in both conditions, and even deteriorate in the continual fine-tuning condition. This suggests that the removal of lexical information alone is 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. 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*. Furthermore, the fine-tuned from scratch model sustains 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.[12]

Continual fine-tuning with CDA results in the best debiasing outcomes. The F1-scores across pronoun-specific test sets surpass 54.0, an improvement for all pronouns. 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 continual fine-tuning already was the preferred method, due to its computational efficiency.

Lastly, we evaluate the impact of debiasing on the performance on the original test set in Appendix D, to investigate if any knowledge is lost through the debiasing process. Table 14 shows that all debiased models exhibit a small performance drop in comparison to the original model, while this decline is most pronounced for the delexicalised models. The decrease is smaller for the CDA models, with a drop of only -0.58 in the continual fine-tuning setting.

7.3 Low-resource debiasing experiment

We now investigate whether the best debiasing method can also be effectively applied in a scenario with limited data, because large corpora for debiasing may not always be available. Moreover, debiasing with a smaller corpus reduces the computational costs. We use *CDA through continual fine-tuning*, because this method obtains the best debiasing results (Section 7.2), reducing the performance gap between gendered and gender-neutral pronouns to less than 1%.

We implement CDA with the same experimental setup as before, but we only use fractions of the *gender-neutral* training set, specifically 10%, 5%, 2.5%, and 1.25%, corresponding to 62, 30, 15, and 7 documents respectively.[13] Five partitions are used for each training size fraction, of which the average scores are reported. In the interest of computational efficiency, we only use one seed.

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

[12] 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.

[13] 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.

Table 7: Debiasing experiment: LEA F1 and pronoun scores (PS) on the *pronoun-specific* test sets after debiasing, as the average across five random seeds. While delexicalisation does not improve the results, CDA reduces the performance gap between gendered and gender-neutral pronouns.

Model	Metric	Hij (masculine)	Zij (feminine)	Hen (gender-neutral)	Die (gender-neutral)
Original model	LEA	51.29 (σ=0.42)	50.77 (σ=0.37)	49.14 (σ=0.68)	48.36 (σ=0.44)
	PS (%)	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	LEA	53.04 (σ=0.70)	53.31 (σ=0.53)	50.67 (σ=0.79)	50.69 (σ=0.64)
	PS (%)	76.50 (σ=4.56)	82.79 (σ=2.42)	71.55 (σ=4.94)	61.89 (σ=5.53)
CDA	LEA	54.44 (σ=0.41)	54.47 (σ=0.49)	54.40 (σ=0.33)	54.33 (σ=0.41)
	PS (%)	86.88 (σ=1.64)	89.08 (σ=0.93)	88.02 (σ=0.74)	89.37 (σ=0.57)
Continual fine-tuning the wl-coref model					
Delexicalisation	LEA	53.74 (σ=0.78)	53.53 (σ=0.78)	50.51 (σ=1.05)	50.07 (σ=0.90)
	PS (%)	89.29 (σ=1.17)	88.76 (σ=0.98)	72.91 (σ=2.80)	57.17 (σ=1.95)
CDA	LEA	54.57 (σ=0.59)	54.48 (σ=0.63)	54.50 (σ=0.58)	54.36 (σ=0.59)
	PS (%)	90.52 (σ=0.44)	90.60 (σ=0.33)	90.16 (σ=0.51)	89.60 (σ=0.50)

Table 8: Low-resource debiasing experiment: Pronoun scores after applying *CDA continual fine-tuning* with various fractions of the full *gender-neutral* training set. The reported scores represent average across five data partitions. The performances already improve after debiasing with a few documents.

Percentage	# Train docs	Hij (masculine)	Zij (feminine)	Hen (gender-neutral)	Die (gender-neutral)
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%

7.3.1 Results. Table 8 presents the outcomes. **The pronoun scores improve after debiasing with just a few debiasing documents.** With a tiny dataset of 7 documents (3–4 documents per pronoun), substantial improvements (+12.67 percentage points for *die*; +6.33 percentage points for *hen*) are observed. Furthermore, by using 5% of the documents (15 documents per pronoun), the pronoun scores already surpass 80% for the gender-neutral pronouns. This is in line with Björklund and Devinney [4], who observe that including gender-neutral pronouns in 2% of the training instances leads to a satisfying POS-tagging performance on these pronouns.[14] The gap between debiasing with 5% of the documents and all documents is only 2.62 percentage points for *hen* and 6.02 percentage points for *die*. These results show that effective debiasing can be achieved with reduced access to resources.

7.4 Unseen pronouns experiment

7.4.1 Setup. This experiment evaluates the ability of all models to process pronouns that have not previously been encountered by the model. The reason for this evaluation is that novel (neo-)pronouns may be popularised in the future [30]. 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.

We evaluate the original and the debiased models on the *unseen test set*: a version of the dataset wherein all third-person pronouns are substituted by a neopronoun p, 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* }.[15] Prior studies on debiasing coreference resolution systems [30, 51, 52] have not included evaluations that focus on previously unseen pronouns.

We do not expect our previous debiased models to improve the performance on neopronouns: although the delexicalisation method was specifically designed to enable the model to process pronouns of diverse lexical forms, its debiasing performance showed unsatisfactory for gender-neutral pronouns (see Section 7.2). Consequently, we expect this method will similarly fall short on debiasing unseen pronouns. Moreover, CDA relies on instructing the model to process a particular pronoun by exposing it directly to the pronoun's lexical form. Consequently, it is also expected that CDA will not enhance performance on unseen pronouns, as the model lacks exposure to these specific pronouns.

[14] 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%).

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

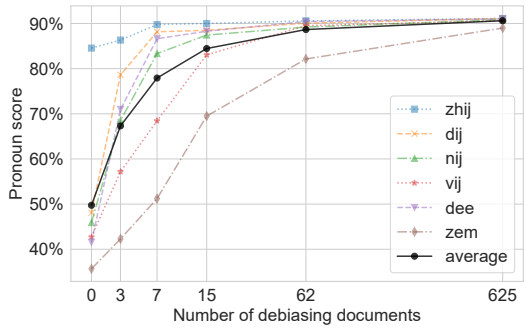

Figure 2: Neopronouns debiasing experiment: Pronoun scores across six neopronouns as a function of the number of debiasing documents included in *CDA continual fine-tuning*. The black dotted line indicates the average pronoun scores across the different neopronouns. Reported scores are the average of five data partitions.

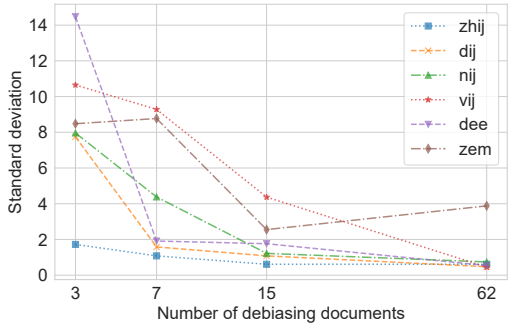

Figure 3: Neopronouns debiasing experiment: Standard deviations of the pronoun scores across five different data partitions, for six neopronouns, as a function of the number of debiasing documents included in *CDA continual fine-tuning*. The zero and 625 training documents settings are excluded from this figure, because these two settings only use one data partition (an empty set or the full set).

7.4.2 Results. Table 9 reports the results for this experiment. The original model has an unsatisfactory performance on unseen pronouns. Moreover, consistent with our expectations, **neither of the debiasing methods improves the performance on unseen pronouns:** the highest scores are for the *continual fine-tuned CDA* model, which achieves a pronoun score of 36.3 percent points lower than on *die* pronouns (Table 7).

7.5 Neopronouns debiasing experiment

In light of the poor performance on processing previously unseen pronouns, we now assess whether continual CDA fine-tuning, the most successful method in our debiasing experiment, is able to improve the processing of neopronouns. We adopt a low-resource setting for two reasons: (1) it is computationally desirable as it requires minimal resources and computational costs; (2) the experiments in Section 7.3 showed satisfactory debiasing results in such a setup. 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: 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.[16]

7.5.1 Setup. The setup is similar to that of the low-resource debiasing experiment (Section 7.3). However, here, we debias each pronoun individually, to allow a comparison between neopronouns. For each pronoun, we create a debiasing and a test set with this particular pronoun inserted. Similarly, we vary the amount of debiasing documents between 3 (0.625%), 7 (1.25%), 15 (2.5%) and 62 (10%) documents, using five data partitions. For comparison, we include the model performance (a) before debiasing and (b) after debiasing with the full training set (625 documents) as baselines.

7.5.2 Results. The results of the debiasing process, in terms of pronoun scores, are presented in Figure 2.

The performance on the different pronouns before debiasing varies a lot. 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 that least resembles known pronouns, exhibits the lowest initial score of 35.66%.

A small number of debiasing documents can improve the performance. We observe a consistent debiasing trend across the pronouns. For example, using only three training documents results in a substantial average performance improvement from 49.7% to 67.3% (+17.6 percentage points), despite high standard deviations (see Figure 3). A high standard deviation is sensible in this setting, considering the significant variation in length and pronoun frequency across training documents. Moreover, as the number of debiasing documents increases, the performances keep improving and standard deviations generally decrease. With the inclusion of 15 debiasing documents, the average performance on neopronouns already improves to 84.5%. It is worth noting that for certain pronouns (*zhij, dij, nij*), performance begins to plateau with additional debiasing documents, indicating a saturation of debiasing data; whereas others (particularly *zem*) continues to benefit from further debiasing. Additionally, we discuss the effect of the tokeniser in Appendix E.

Taken together, the outcomes of this experiment are promising. Satisfactory results are achieved across various sets of neopronouns using CDA continual fine-tuning with a limited dataset. These findings demonstrate the feasibility of future-proof gender-inclusive debiasing with minimal resource requirements and low computational costs.

[16]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.

Table 9: Unseen pronouns experiment: 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 continual fine-tuning, using delexicalisation or CDA. The reported scores are the average of five random seeds. All models perform poorly on previously unseen pronouns.

	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				
Delexicalisation	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)
Continual fine-tuning the wl-coref model				
Delexicalisation	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)

8 DISCUSSION AND CONCLUSION

The results of this study have two main implications. First, the debiasing results show that applying CDA manifests a considerable improvement, even when applied through continual fine-tuning with just a handful of documents. This outcome aligns with the finding of Björklund and Devinney [4] 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.

Second, the results observed in this study suggest that there exists 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. The implementation of NLP technologies in this context holds the potential to facilitate the wider adoption of these innovative structures within societies, by showing people an example of how to correctly use these structures.

A limitation of this study is that we only consider a single model in our evaluation. Future research could extend the scope by exploring potential trends across various models. Moreover, future research could assess the applicability of the current setup to other languages. For example, for languages like Italian or French, in which gender is more intricately woven into grammatical structures, the debiasing task may be more complex.

9 ETHICAL CONSIDERATIONS

We 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[17] and style. Despite the fact that we 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 [7].

Moreover, this study has not actively involved non-binary and transgender individuals in the designing, debiasing and evaluation process, and instead was for the main part conducted by cisgender individuals in a binary gendered environment. We recognise that this may have led to overlooking important barriers, risks or opportunities relating to the emancipation of non-binary individuals. We hope that 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 [9], is essential [14].

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

ACKNOWLEDGMENTS

Dong Nguyen is funded by the Veni research programme with project number VI.Veni.192.130, which is (partly) financed by the Dutch Research Council (NWO). We want to thank Antal van den Bosch, Guusje Juijn, Tim Koornstra, Michael Pieke, Daan van der Weijden, Shiyi Butter and Silin Chen for their constructive feedback. We thank the anonymous reviewers for their efforts in reviewing this paper and their interesting questions.

REFERENCES

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

[2] 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*. Association for Computational Linguistics, Seattle, United States, 3426–3432. https://doi.org/10.18653/v1/2022.naacl-main.250

[3] Sander Becker. 2020. Is het Nederlands klaar voor het genderneutrale 'Hen loopt'? *Trouw* (8 Oct. 2020). https://www.trouw.nl/cultuur-media/is-het-nederlands-klaar-voor-het-genderneutrale-hen-loopt~b5fb2e1b/

[4] 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*, Bharathi R. Chakravarthi, B. Bharathi, Joephine Griffith, Kalika Bali, and Paul Buitelaar (Eds.). INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 54–61. https://aclanthology.org/2023.ltedi-1.8

[5] Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics* 11 (2023), 212–226. https://doi.org/10.1162/tacl_a_00543

[6] 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*. Association for Computational Linguistics, Seattle, United States, 3624–3630. https://doi.org/10.18653/v1/2022.naacl-main.265

[7] 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*. Association for Computational Linguistics, Online, 4568–4595. https://doi.org/10.18653/v1/2020.acl-main.418

[8] 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 (Nov. 2021), 615–661. https://doi.org/10.1162/coli_a_00413

[9] 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*. Association for Computational Linguistics, Online, 27–35. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.4

[10] 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*. Association for Computational Linguistics, Florence, Italy, 173–181. https://doi.org/10.18653/v1/W19-3824

[11] 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*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[12] 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*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

[13] 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*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1968–1994. https://doi.org/10.18653/v1/2021.emnlp-main.150

[14] 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* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2083–2102. https://doi.org/10.1145/3531146.3534627

[15] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[16] Vladimir Dobrovolskii. 2021. Word-Level Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7670–7675. https://doi.org/10.18653/v1/2021.emnlp-main.605

[17] EditieNL. 2021. Lij of vij: is er een nieuw non-binair persoonlijk voornaamwoord nodig? *RTLnieuws* (25 May 2021). https://www.rtlnieuws.nl/editienl/artikel/5232738/nieuw-persoonlijk-voornaamwoord-non-binaire-personen-hen-hun-die

[18] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[19] Marinel Gerritsen. 2002. Language and gender in Netherlands Dutch: Towards a more gender-fair usage. *Gender Across Languages* 2 (2002).

[20] 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* (, Montréal, QC, Canada,) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 901–912. https://doi.org/10.1145/3600211.3604672

[21] 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 (2015). https://doi.org/10.3389/fpsyg.2015.00893

[22] Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A Coreference Corpus and Resolution System for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/49_paper.pdf

[23] Het Neutrale Taal collectief. [n. d.]. Lijst van populaire voornaamwoorden. https://nl.pronouns.page/voornaamwoorden. Accessed: 2022-10-14.

[24] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5352–5367. https://doi.org/10.18653/v1/2023.acl-long.293

[25] Robin Mattias Hurkens. 2021. Genderneutraal: geen 'hij', 'zij', 'hen', 'dij', maar 'ij'. *De Volkskrant* (26 May 2021). https://www.volkskrant.nl/columns-opinie/genderneutraal-geen-hij-zij-hen-dij-maar-ij~b3c890b9/

[26] 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* (2016). https://transequality.org/sites/default/files/docs/usts/USTS-Full-Report-Dec17.pdf

[27] 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 (2020), 64–77. https://doi.org/10.1162/tacl_a_00300

[28] 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*. Association for Computational Linguistics, Florence, Italy, 166–172. https://doi.org/10.18653/v1/W19-3823

[29] 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*, Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 116–127. https://doi.org/10.18653/v1/2023.latechclfl-1.13

[30] 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*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1221–1232. https://aclanthology.org/2022.coling-1.105

[31] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 377–392. https://doi.org/10.18653/v1/2023.acl-long.23

[32] 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*. Association for Computational Linguistics, Copenhagen, Denmark, 188–197. https://doi.org/10.18653/v1/D17-1018

[33] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 687–692. https://doi.org/10.18653/v1/N18-2108

[34] 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*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 993–1005. https://doi.org/10.18653/v1/2022.findings-emnlp.70

[35] 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)*, Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, and Roman Yangarber (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 146–154. https://doi.org/10.18653/v1/2023.bsnlp-1.17

[36] 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). Association for Computational Linguistics, Berlin, Germany, 632–642. https://doi.org/10.18653/v1/P16-1060

[37] Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. SoNaR User Documentation. 1, 4 (2013). https://www.ivdnt.org/images/stories/producten/documentatie/sonar_documentatie.pdf

[38] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (, Chicago, IL, USA,) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1246–1266. https://doi.org/10.1145/3593013.3594078

[39] 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*. Association for Computational Linguistics, Barcelona, Spain (online), 79–90. https://aclanthology.org/2020.crac-1.9

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

[41] 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)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/549_Paper.pdf

[42] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 8–14. https://doi.org/10.18653/v1/N18-2002

[43] 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*. 1121–1131.

[44] 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)* (2021). https://www.transgendernetwerk.nl/wp-content/uploads/Inclusion4All-National-report-Netherlands_NL.pdf

[45] Transgender Netwerk Nederland. 2016. ZO MAAK JE NA TOILETTEN OOK TAAL GENDERNEUTRAAL. https://www.transgendernetwerk.nl/non-binair-voornaamwoord-uitslag/. *Transgender Netwerk Nederland Nieuws* (June 2016). https://www.transgendernetwerk.nl/non-binair-voornaamwoord-uitslag/

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

[47] 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*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 47–56. https://doi.org/10.18653/v1/2021.crac-1.5

[48] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6790–6809. https://doi.org/10.18653/v1/2023.acl-long.375

[49] 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 (2018), 605–617. https://doi.org/10.1162/tacl_a_00240

[50] 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*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[51] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064

[52] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003

[53] Lal Zimman. 2018. Transgender Language, Transgender Moment: Toward a Trans Linguistics. In *The Oxford Handbook of Language and Sexuality*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212926.013.45 arXiv:https://academic.oup.com/book/0/chapter/358161844/chapter-ag-pdf/50001811/book_42645_section_358161844.ag.pdf

A GENDERED NOUNS REWRITING RULES

Table 10: Rewriting rules for gendered Dutch nouns to a gender-neutral version of this word (part 1). Not all Dutch words have a gender-neutral alternative however. * marks difficult cases, for which some meaning is lost.

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

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 12: Learning rate tuning of the wl-coref model, using XLM-RoBERTa [11] 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 [16]. The best F1-score is found for a learning rate of $5e^{-4}$.

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

Table 13: BERT learning rate tuning of the wl-coref model, using XLM-RoBERTa [11] 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 [16]. The best F1-score is found for a BERT learning rate of 3^{-5}.

BERT learning rate	Precision	Recall	F1
$1e^{-4}$	50.2	58.9	54.2
$1e^{-5}$ (original value)	51.2	58.6	54.7
$2e^{-5}$	53.2	57.3	55.2
$3e^{-5}$	52.0	**59.5**	**55.5**
$4e^{-5}$	52.0	57.9	54.8
$5e^{-5}$	**54.1**	56.6	55.3
$5e^{-6}$	51.2	56.6	53.8

B HYPERPARAMETER TUNING

The results of the hyperparameter search for the learning rate is present in Table 12 and the corresponding results for the BERT learning rate can be found in Table 13.

C MODEL COMPARISON

Compared to the neural e2e-Dutch model,[18] wl-coref's test F1-score is 6 points lower. However, improvements over e2e-Dutch are (1) a lower complexity [16] and a smaller difference between the development and test performance (F1 $\Delta = -3.7$ for e2e-Dutch and F1 $\Delta = +0.3$ for wl-coref), suggesting minimal overfitting. Moreover, the wl-coref model notably outperforms the rule-based dutchcoref model [46] (+11.6 in test F1-score).

[18] https://github.com/Filter-Bubble/e2e-Dutch

Table 14: Debiasing experiment: Average LEA F1-scores achieved by the debiased models on the regular SoNaR-1 test data, as the average across five random seeds. This evaluation aims to assess potential losses in abilities through the debiasing process. CDA leads to a smaller drop in performance than delexicalisation.

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

D DEBIASING IMPACT ON GENERAL MODEL PERFORMANCE

We evaluate the impact of debiasing on the model's performance on the original test set, in order to inspect whether any knowledge is lost through the debiasing process. We compare the two debiasing techniques (see Section 7.2). Table 14 shows that applying CDA results in a smaller performance drop than delexicalisation.

E TOKENISATION

We directly use the XLM-RoBERTa-Base sentence piece tokeniser, which is trained on multilingual pre-training data. The tokeniser represents both gender-specific (*hij, zij*) and gender-neutral pronouns (*hen, die*) as single tokens, suggesting that their performance disparities are not a result from the tokenisation. However, for neo-pronouns, debiasing efficiency seems to correlate with token quantity. In Figure 2, the neopronouns *zhij/dij/dee/nij/vij/zem* (ordered by their debiasing efficiency) are represented with 3/2/2/2/1/1 tokens. One notable comparison is between *vij* and *dee*: they have similar performance before debiasing, but after debiasing *dee* (2 tokens) outperforms *vij* (1 token) by a substantial margin.