24 NOVEMBER 2023

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

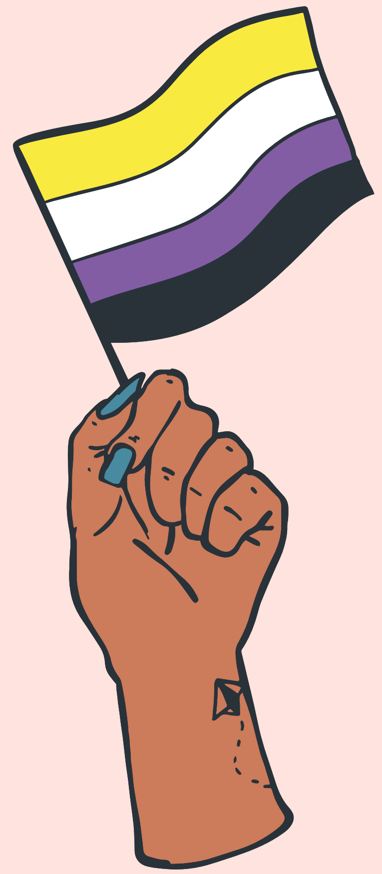
Thesis MSc Artificial Intelligence
Supervised by Dr. Dong Nguyen and Yupei Du
Goya van Boven (she/her)

PROBLEM OUTLINE

- Gender-neutral language is becoming more popular across Western languages. In Dutch the **gender-neutral pronouns** *hen* and *die* are increasingly being used by non-binary individuals

- **Non-binary** indiviuals identify with a gender identity that is outside the female-male binary

- **Transgender** individuals do not identify with the gender they were assigned at birth

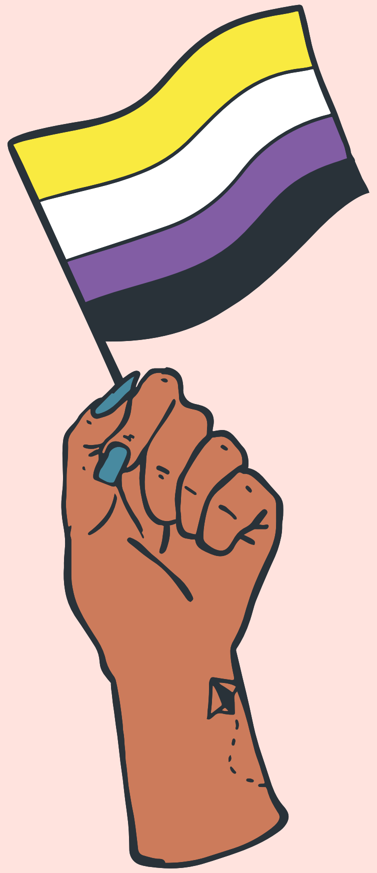


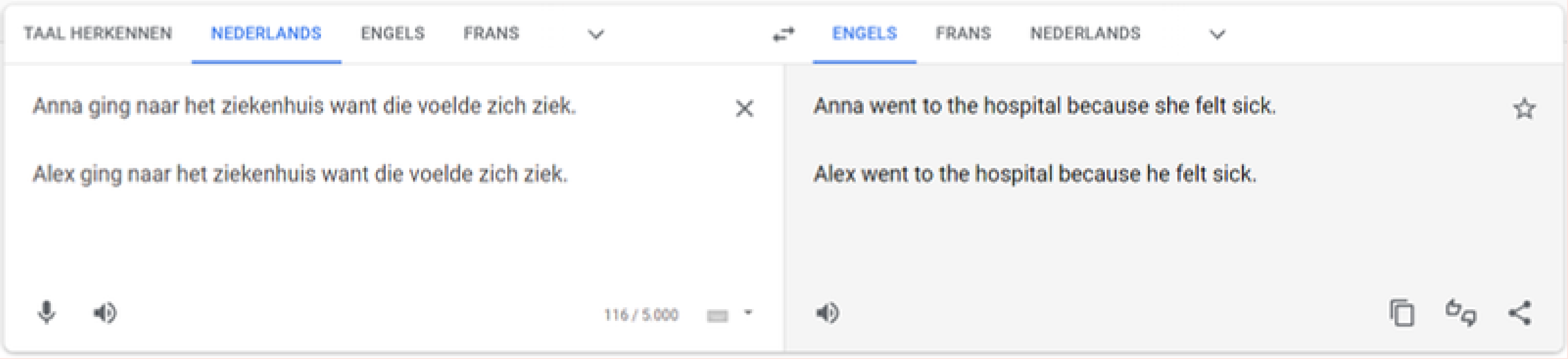
PROBLEM OUTLINE

- In recent years, **gender bias** has become a hot topic in NLP (e.g. Bolukbasi et al., 2016; Rudinger et al., 2018; Zhao et al., 2018; Caliskan et al., 2017)

- However, gender bias research in NLP is typically **trans-exclusive** (Cao and Daumé III, 2021; Dev et al., 2021)

- Trans-exclusive NLP can lead to **erasure** of non-binary gender identities and the **misgendering** of individuals





PROBLEM OUTLINE

Recent **evaluations** of the **representation of non-binary people** in **English pre-trained language models** (Dev et al., 2021; Brandl et al., 2022) and **co-reference resolution systems** (Baumler and Rudinger, 2022; Saunders et al., 2020; Cao and Daumé III, 2021) show **poor performances**.

Research gap:

- No such evaluation is yet done for any Dutch system or language model
- No study (as far as I know) yet evaluates debiasing methods to make coreference resolution systems more trans-inclusive

Coreference resolution

The task of deciding whether two mentions refer to the same entity.

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

Forms the basis of downstream tasks like

- information extraction
- summarization
- question answering

→ Failing to identify gender neutral pronouns could lead to the **erasure** of non-binary individuals

Research Question

Can the **debiasing techniques** *counterfactual data augmentation* and *delexicalisation* improve the ability of **Dutch coreference resolution** systems to **process gender-neutral pronouns**?

Experimental setup

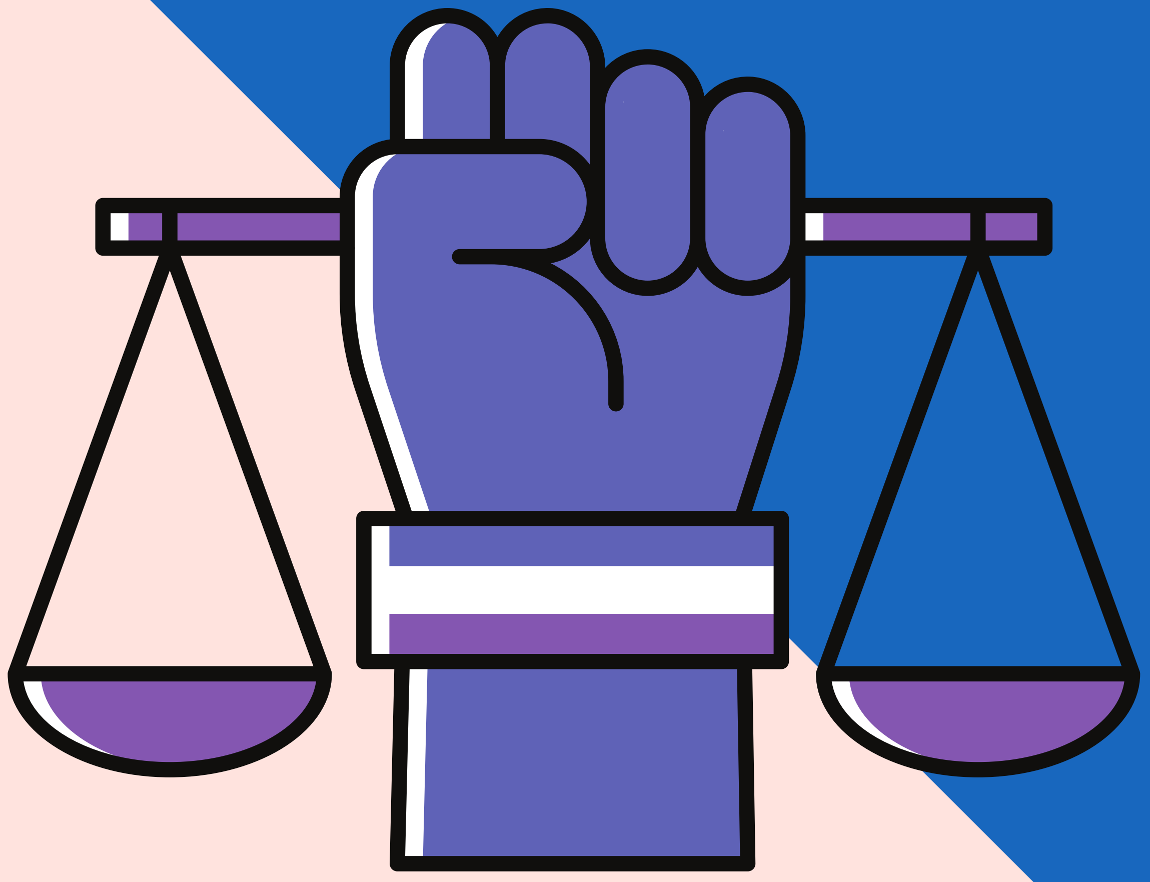
Dataset transformation

- Replace gendered pronouns by gender-neutral pronouns

Evaluate existing **Dutch coreference system** on **gender-neutral pronouns**

Apply two **debiasing techniques** (CDA and delexicalisation) and **evaluate** their effectiveness

Data transformation



SoNaR-1 corpus (Schuurman et al., 2010)
Originally **79% of the third person pronouns are male**;
the corpus does not include gender-neutral pronouns

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.*

Rewritten:
Hij stierf toen **ANON_1** 27 jaar was en op het toppunt van **zijn** creatieve periode.
Zij stierf toen **ANON_1** 27 jaar was en op het toppunt van **haar** creatieve periode.
Hen stierf toen **ANON_1** 27 jaar was en op het toppunt van **hun** creatieve periode.
Die stierf toen **ANON_1** 27 jaar was en op het toppunt van **diens** creatieve periode.

Model evaluation on pronouns

Evaluate the wl-coref model (Dobrovolskii, 2021), with XLM-RoBERTa base (Conneau et al., 2020) as its base model

Evaluation metric = a **pronoun score** for third person pronouns:

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

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

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

Model evaluation on pronouns

Evaluate the wl-coref model (Dobrovolskii, 2021), with XLM-RoBERTa base (Conneau et al., 2020) as its base model

Evaluation metric = a **pronoun score** for third person pronouns:

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

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

$$gold\_antecedents(they) = \{Raven\}$$

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

$$predicted\_antecedents(they) = \{\}$$

Model evaluation on pronouns

Evaluate the wl-coref model (Dobrovolskii, 2021), with XLM-RoBERTa base (Conneau et al., 2020) as its base model

Evaluation metric = a **pronoun score** for third person pronouns:

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

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

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

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

$$predicted\_antecedents(their) = \{Raven\}$$

Model evaluation on pronouns

Evaluate the wl-coref model (Dobrovolskii, 2021), with XLM-RoBERTa base (Conneau et al., 2020) as its base model

Evaluation metric = a **pronoun score** for third person pronouns:

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

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

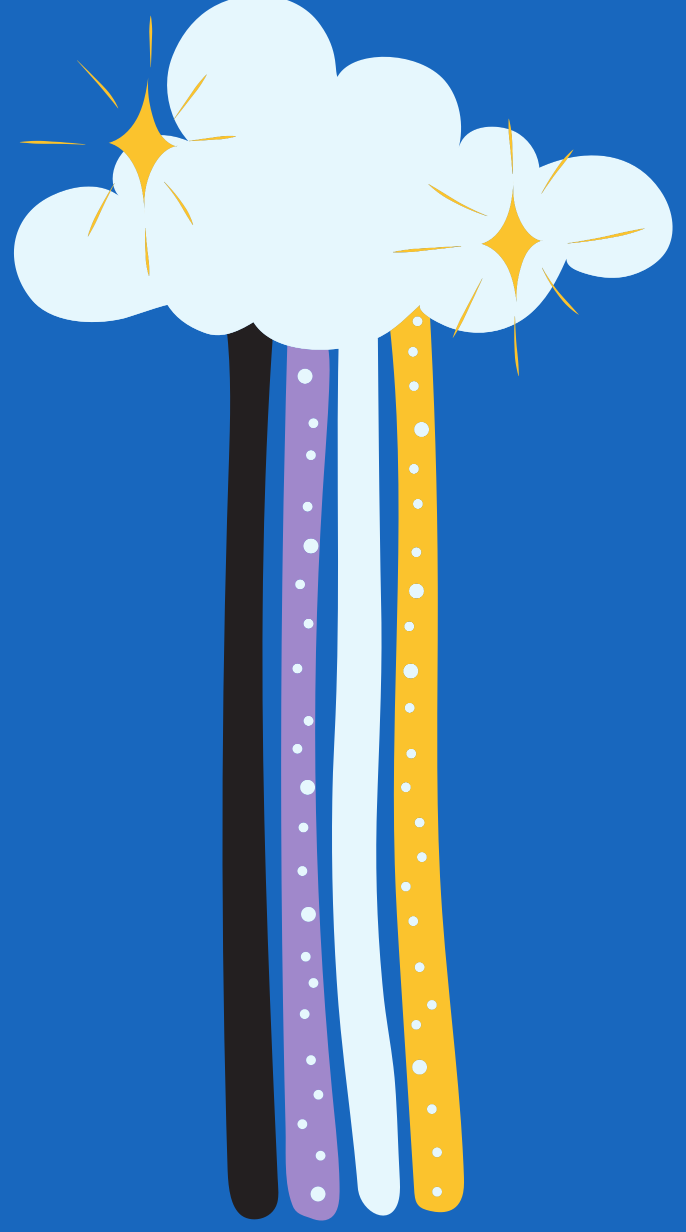
Predicted: [Raven] entered the kitchen. "Did [you] sleep well?", they asked [[their] roommate] "No [Raven]", said [Thorn] annoyed, "[Tobi] called me way too early"

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

Model evaluation on pronouns

Use the wl-coref model (Dobrovolskii, 2021)

Evaluate on pronoun-specific versions of the test set, that only contain one type of third person pronouns



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 9.3: Percentage of pronouns for which at least one correct antecedent is identified. Scores are computed as the average of five random seeds, over a version of the test set that only contains the pronoun of interest.

Debiasing

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.*

Counterfactual Data Augmentation :

Insert gender-neutral pronouns into the training data

Train model on a *gender-neutral* version of the training data:

Hen stierf toen **ANON_1** 27 jaar was en op het toppunt van **hun** creatieve periode.
Die stierf toen **ANON_1** 27 jaar was en op het toppunt van **diens** creatieve periode.

The usage of *hen/die* is alternated (50/50) between documents

Delexicalisation (Lauscher et al., 2022):

Remove all lexical forms of pronouns and replace them with their POS-tag

The core idea : this way the model will learn to identify *any* type as a pronoun

<SUBJ> stierf toen **ANON_1** 27 jaar was en op het toppunt van **<POSS>** creatieve periode.

Debiasing results

Full retraining:

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)
Delex	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)

Pronouns scores after debiasing. Models are trained for 20 epochs. Results are the average of 5 random seeds.

Debiasing results

Full retraining:

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)
Delex	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)

Pronouns scores after debiasing. Models are trained for 20 epochs. Results are the average of 5 random seeds.

Fine-tuning:

Model	Hij	Zij	Hen	Die
Original model	88.51% (σ=0.73)	86.95% (σ=0.80)	77.61% (σ=2.40)	68.38% (σ=6.55)
Delex	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)

Pronouns scores after debiasing. Models are fine-tuned for 10 epochs. Results are the average of 5 random seeds.

Debiasing results

Full retraining:

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)
Delex	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)

Pronouns scores after debiasing. Models are trained for 20 epochs. Results are the average of 5 random seeds.

Fine-tuning:

Model	Hij	Zij	Hen	Die
Original model	88.51% (σ=0.73)	86.95% (σ=0.80)	77.61% (σ=2.40)	68.38% (σ=6.55)
Delex	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)

Pronouns scores after debiasing. Models are fine-tuned for 10 epochs. Results are the average of 5 random seeds.

→ **CDA** is **effective** at debiasing in **both** settings, while **delexicalisation** is in **neither**

Debiasing with a smaller dataset

Percentage	# Train documents per pronoun	Hij	Zij	Hen	Die
100%	312	90.76%	90.60%	89.94%	89.67%
10%	31	92.41% (σ=0.19)	91.26% (σ=0.41)	88.64% (σ=0.79)	85.42% (σ=0.94)
5%	15	92.02% (σ=0.48)	90.66% (σ=0.43)	87.32% (σ=0.90)	83.65% (σ=1.08)
2.5%	8	91.40% (σ=0.64)	89.96% (σ=0.54)	85.09% (σ=0.89)	79.48% (σ=0.94)
1.25%	4	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%

Table 9.11: Pronoun scores after fine-tuning the wl-coref model using the debiasing technique CDA and various fractions of the full *gender-neutral* training set.

Debiasing with just a few documents (8 documents per pronoun) already improves the pronoun score by +6.3% (*hen*) and +13.7% (*die*)

Debiasing neopronouns

Lausscher et al. (2022) point out the importance of debiasing in a future-proof manner: what if different neopronouns are popularised in a few years?

→ **Can existing debiasing techniques also improve the performance on previously unseen pronouns?**

CDA:

Hen stierf toen **ANON_1** 27 jaar was en op het toppunt van **hun** creatieve periode.
Die stierf toen **ANON_1** 27 jaar was en op het toppunt van **diens** creatieve periode.

Delexicalisation:

<OBJ> stierf toen **ANON_1** 27 jaar was en op het toppunt van **<POSS>** creatieve periode.

Debiasing neopronouns

Can existing debiasing techniques also improve the performance on previously unseen pronouns?

Evaluate the performance of the models on a set of neopronouns previously unseen by the model:

$p \in \{dee/dem/dijr, dij/dem/dijr,\ nij/ner/nijr,\ vij/vijn/vijns,\ zhij/zhaar/zhaar,\ zem/zeer/zeer\}$

	Pronoun score
Original model	46.68% (σ=2.31)
Delex full	48.03% (σ=2.01)
Delex fine	49.56% (σ=2.07)
CDA full	51.72% (σ=2.90)
CDA fine	53.37% (σ=3.55)

Neither of the debiasing techniques improves the performance on previousy unseen pronouns.
Can we also use CDA to learn the model to use these completely new pronouns?

Debiasing neopronouns

→ **Can CDA learn the model to process these completely new pronouns?**

Apply CDA by fine-tuning with *dee/dem/dijr* pronouns inserted.
Evaluate the pronoun score for this pronoun set

Percentage	# Train documents	Pronoun score
2.5%	15	88.28% (σ=1.76)
1.25%	7	86.62% (σ=1.91)
0.625%	3	70.97% (σ=14.47)
Original model	0	41.55%

Fine-tuned for 10 epochs

CDA **effectively learns** the model to use **neopronouns** with just **7 documents**, showing that **future-proof debiasing** is **possible** with **low resources** and **computational costs**

Discussion

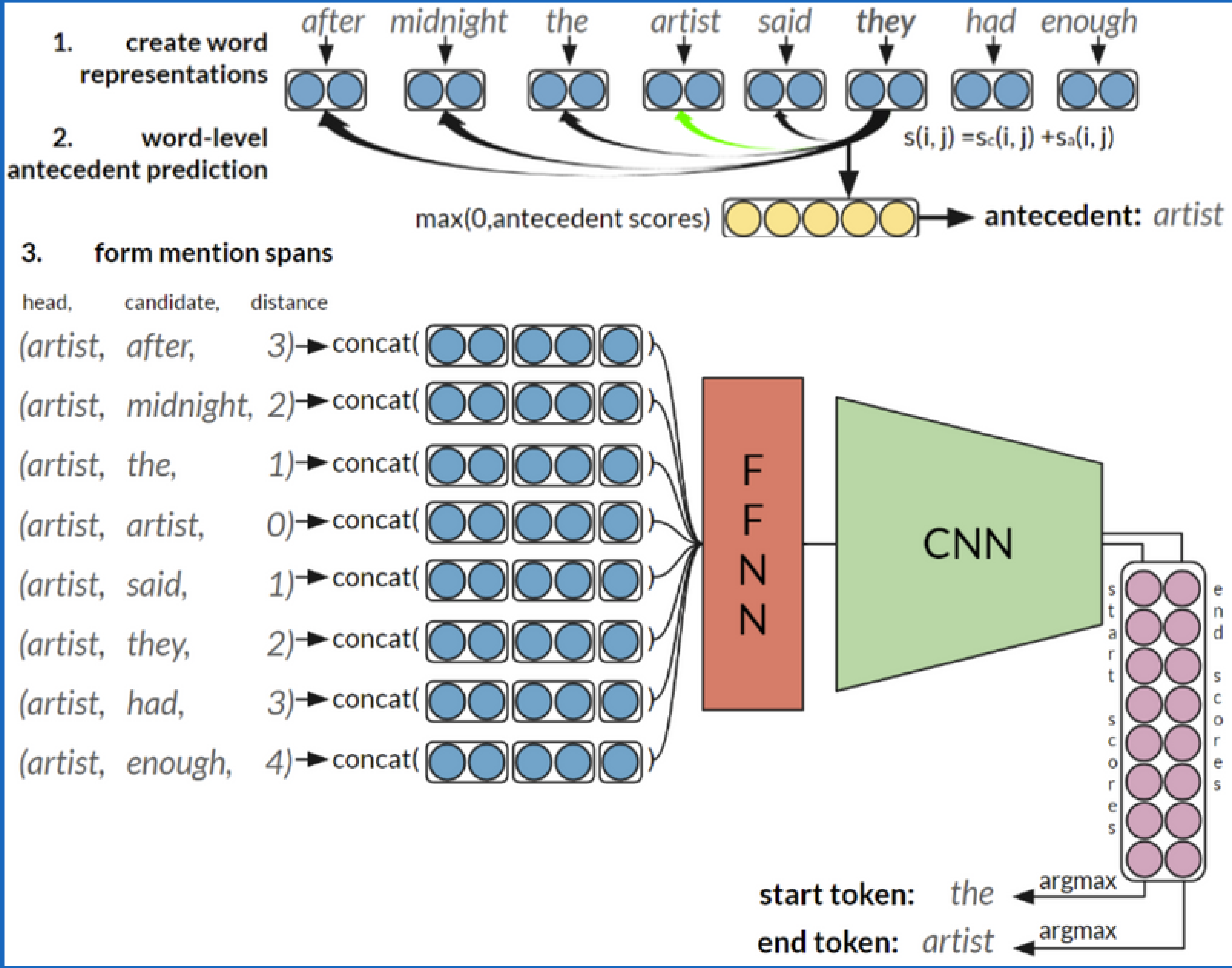
- **Promising findings**, show that NLP technologies have a **possibily** to be at the **forefront** of **emanciplation** movements

- A strong **limitation** of this study is that it zooms in **only** on **pronouns**, and **does not evaluate** the **language usage of non-binary individuals** in a broader sense

- **Future work** might investigate debiasing **other NLP tasks** in a **non-binary context**; and evaluate **other languages** than Dutch

QUESTIONS
OR
FEEDBACK?

wl-coref



1. create word representations
after midnight the artist said they had enough
2. word-level antecedent prediction
s(i, j) =sc(i, j) +sa(i, j)
max(0,antecedent scores)
antecedent: artist
3. form mention spans
head, candidate, distance
(artist, after, 3)→concat(
(artist, midnight, 2)→concat(
(artist, the, 1)→concat(
(artist, artist, 0)→concat(
(artist, said, 1)→concat(
(artist, they, 2)→concat(
(artist, had, 3)→concat(
(artist, enough, 4)→concat(
FFNN
CNN
start scores
end scores
argmax
argmax
start token: the
end token: artist