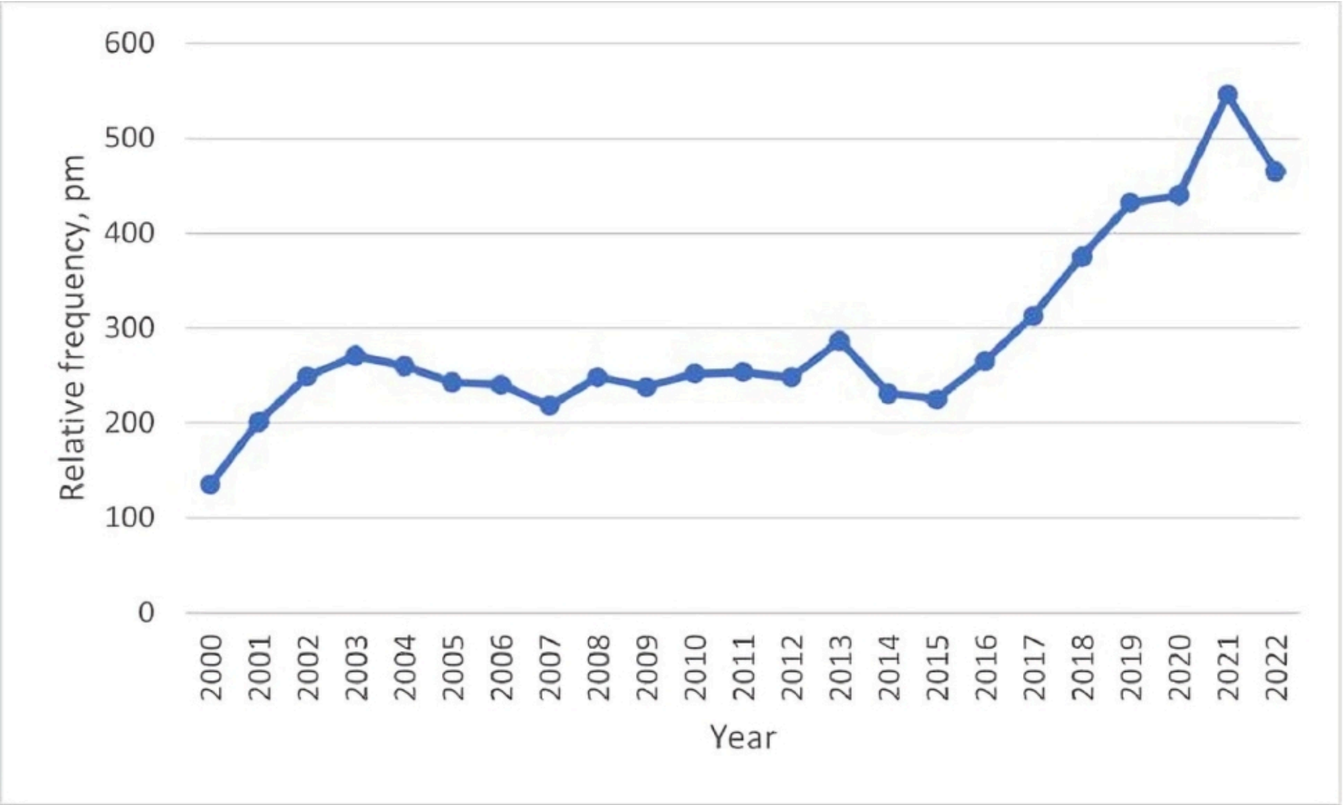
Gender Swapping as a Data Augmentation Technique: Developing Gender-Balanced Datasets for Ukrainian Language Processing | UNLP 2025

Olha Nahurna, Mariana Romanyshyn

INTRODUCTION & MOTIVATION

Gender	Correct	Incorrect
Feminine	Моя подруга є досвідченою **керівницею**.	Моя подруга є досвідченим **керівником**.
Masculine	Іван Петрович був моїм першим **керівником**.	Іван Петрович був моїм першим **керівницею**.

Fig 1: Dynamics of 744 feminine occupational titles in newspaper texts in GRAC.



Source:
Feminine Personal Nouns in Ukrainian: Dynamics in a Corpus
Authors: Vasyl Starko, Olena Synchak

Fig 2: The use of feminine derivatives ending in -знавиця [-znavytsia] (female -logist) in GRAC.

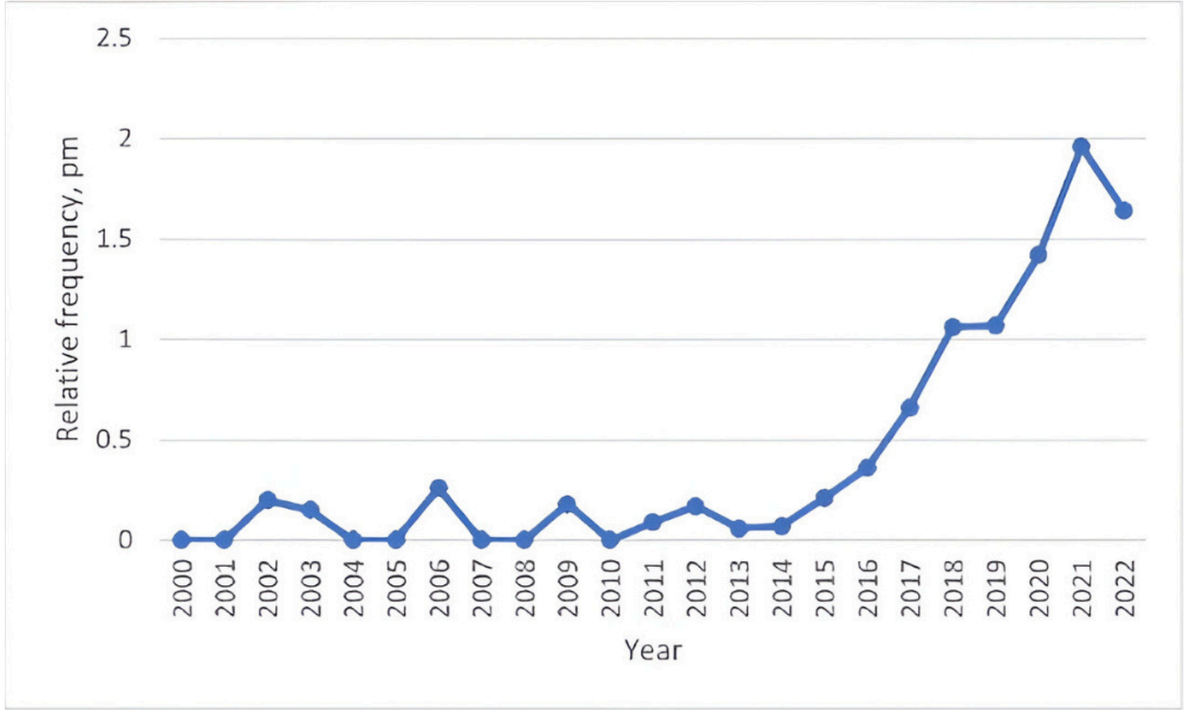
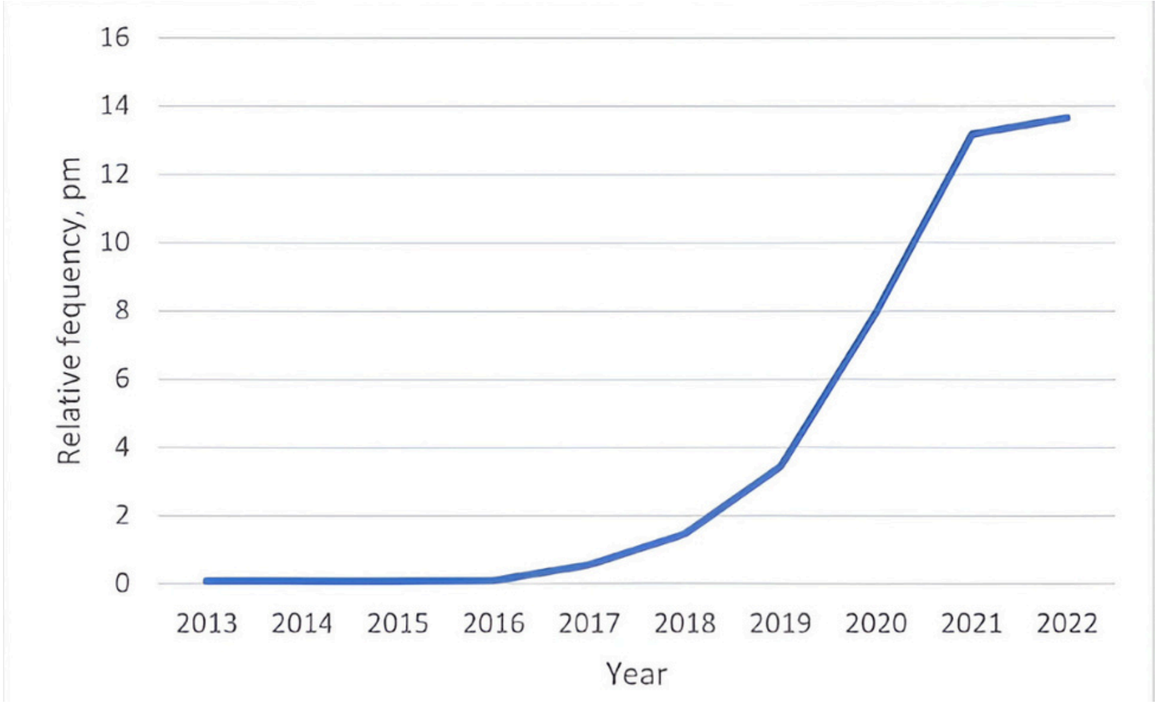


Fig 3: The use of feminine derivatives ending in -логиня [-lohynia] (female -logist) in GRAC.



Source:
Feminine Personal Nouns in Ukrainian: Dynamics in a Corpus
Authors: Vasyl Starko, Olena Synchak

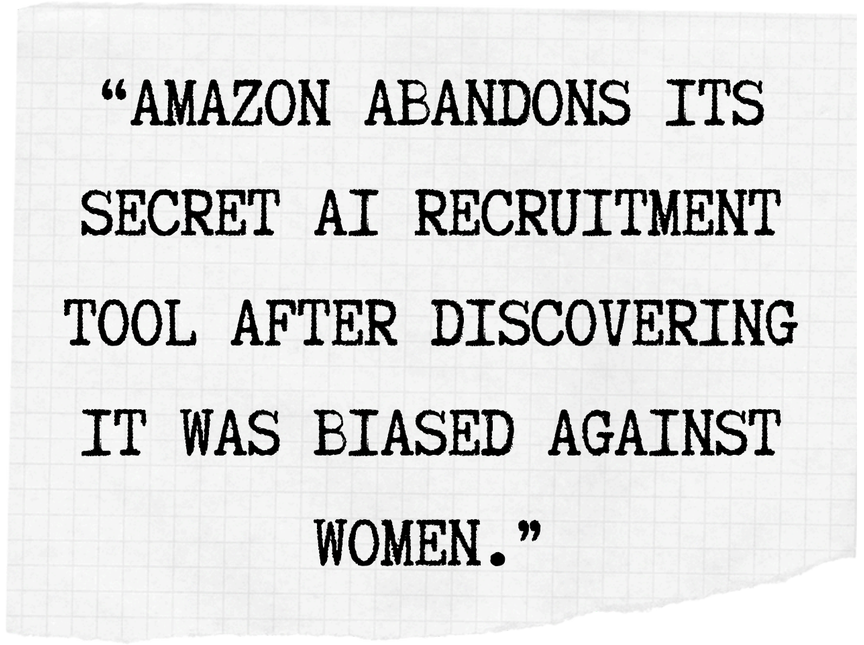
THEORETICAL BACKGROUND

Morphologically Rich Language

Language	Feminine	Masculine
Ukrainian	Талановит**а** співа**чка** виступа**ла** в театрі	Талановит**ий** співа**к** виступа**в** в театрі
Spanish	**La** talentos**a** cantante actuó en el teatro	**El** talentos**o** cantante actuó en el teatro.
Portuguese	**A** cantor**a** talentos**a** se apresentou no teatro.	**O** cantor talentos**o** se apresentou no teatro.
Arabic	المغنية الموهوبة قدمت عرضًا في المسرح.	المغني الموهوب قدم عرضًا في المسرح.

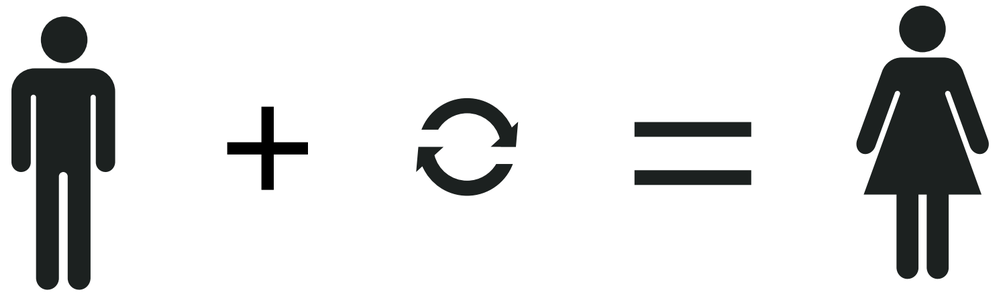
*translation by DeepL

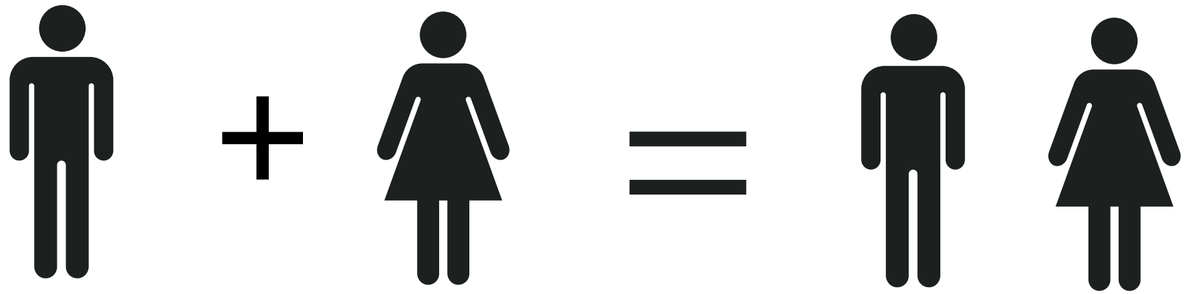
Gender bias


“AMAZON ABANDONS ITS SECRET AI RECRUITMENT TOOL AFTER DISCOVERING IT WAS BIASED AGAINST WOMEN.”

Data Augmentation

(Counterfactual Data Augmentation)

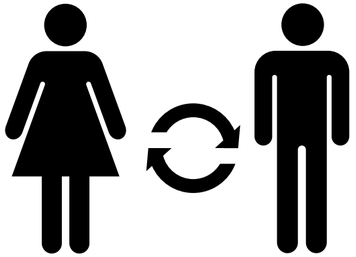




METHODOLOGY

Dataset Selection

Choosing a suitable dataset with gendered entities for analysis



Gender-Swapping with LLM

Using language models to perform gender inversion in sentences

Human in the loop

Involving annotators to correct and refine swapped sentences

LLM Fine-tuning

Training the model using the revised dataset for better performance

Dataset

NER-UK

- 560 text files (train: 391, test: 169)
- 21,993 NER entities
- 13 entity types

file.txt

У Чернівцях завершилися змагання Кубка України з боксу серед жінок і чемпіонат країни серед молоді .

П’ятиденний марафон жіночого боксу зібрав у ФОК « Олімпія » двісті кращих спортсменок України у трьох вікових категоріях : юніорки , молодь та дорослі .

file.ann *(Brat Standoff Format)*

T1	LOC	2	11	Чернівцях
T2	MISC	33	46	Кубка України
T3	ORG	146	161	ФОК « Олімпія »
T4	JOB	176	187	спортсменок
T5	LOC	188	195	України

Gender-swapping

Тим часом **Христоф** та **двоє найманців** , відшукавши високий та зручний пагорб , спостерігали за всім згори .

Original

Тим часом **Христина** та **дві найманкині** , відшукавши високий та зручний пагорб , спостерігали за всім згори .

gpt-4o-mini

Human in the loop

- 1,513 sentences containing JOB-related titles.
- 17 volunteers for manual annotation or validation
- Clear instructions, including requirements, annotation rules, and reference sources to ensure consistency and quality.

Тим часом **Христоф** та **двоє найманців** , відшукавши високий та зручний пагорб , спостерігали за всім згори .

Original

↓

Тим часом **Христина** та **дві найманкині** , відшукавши високий та зручний пагорб , спостерігали за всім згори .

gpt-4o-mini

↓

Тим часом **Христоф** та **дві найманки**, відшукавши високий та зручний пагорб , спостерігали за всім згори .

Human

Results

1. *to Accept*: **58.5%** of the generated sentences did not need any correction.
2. *to Correct*: **37.6%** of the generated sentences were updated by annotators.
3. *to Dismiss*: **3.9%** of examples were dismissed as complex or ambiguous.

83% of created gender-swapped pairs could be found in the dictionary.

Final dataset contained **1,403** parallel sentence pairs

EXPERIMENTS

Experiment 1.
Gender Swapper UK

Create gender-parallel dataset

1. Collect a gender-swapped sentence pair dataset.
2. Ensure balanced distribution of male → female and female → male sentence pairs.
3. Enrich the dataset by incorporating additional gendered terms from the gender pairs dictionary.
4. Split the dataset into train/test (80/20).

LLM Fine-tuning

LLM: Aya-101 (13B)

Size of Dataset: ~2700 pairs

Hardware: A100 GPU via Google Colab Pro

Prompt_1: *“Перефразуй це речення, змінивши його гендерні сутності на протилежні*
(чоловічий <-> жіночий)”

Prompt_2: *Перефразуй це слово, змінивши його гендер на протилежний (чоловічий <-> жіночий)*

Evaluation

Tab 1: Evaluation of LLMs performing the gender-swapping task on the parallel gender-swapped test set.

Metric	Aya-101 original	Aya-101 fine-tuned	GPT-4o-mini
Exact Match	0.21	**0.52**	0.51
Exact Match w/o PERS	0.34	**0.73**	0.70
JOB Match	0.76	**0.87**	0.62
Token Count Match	0.64	**0.93**	0.91
BLEU	0.79	**0.87**	0.85
ROUGE-L	0.21	0.21	**0.22**
BERTScore (F1)	0.97	**0.99**	**0.99**

Experiment 2.
Enhancing Gender-Marked Entity Recognition

Data Augmentation

(sentence level)

Original Sentence:
Тим часом Христоф та двоє найманців, відшукавши високий та зручний пагорб, спостерігали за всім згори.

+

Gender-Swapped Sentence:
Тим часом Христоф та дві найманки, відшукавши високий та зручний пагорб, спостерігали за всім згори.

↓

Augmented Dataset:
Тим часом Христоф та двоє найманців, відшукавши високий та зручний пагорб, спостерігали за всім згори. (Original Sentence)
Тим часом Христоф та дві найманки, відшукавши високий та зручний пагорб, спостерігали за всім згори. (Gender-Swapped Sentence)

Gender composition

Tab 3: Gender composition of **JOB** entities for the Original (only gender-marked sentences) and Augmented NER-UK 2.0 datasets.

Dataset	Total	Gender Distribution		
		Masculine	Feminine	Common
Original	1,982	83%	3.8%	11.3%
Augmented	3,715	49.2%	37.4%	10.5%

Tab 4: Gender composition of **PERS** entities for the Original (only gender-marked sentences) and Augmented NER-UK 2.0 datasets.

Dataset	Total	Gender Distribution	
		Masculine	Feminine
Original	6,235	34.0%	20.6%
Augmented	7,517	30.2%	26.8%

NER for the Gender-Balanced Data

Baseline - state-of-the-art NER model for Ukrainian
uk_ner_web_trf_13class (NER-UK 2.0 paper)

Tools:

- Ukrainian version of the RoBERTa-large model
- spaCy framework

Tab 5: Performance comparison of Original NER and Gender-Balanced NER for **JOB** and **PERS** entities across different test sets.

Test Set	Original NER				Gender-Balanced NER			
	Entity Type	P	R	F1	Entity Type	P	R	F1
Original	JOB	74.39	65.45	**69.64**	JOB	75.05	59.06	66.10
	PERS	96.20	96.60	**96.40**	PERS	97.01	95.18	96.08
Gender-swapped	JOB	89.08	71.26	79.18	JOB	90.63	80.78	**85.42**
	PERS	98.60	98.60	98.60	PERS	98.60	98.88	**98.74**
Augmented	JOB	80.53	67.75	73.59	JOB	82.51	68.43	**74.81**
	PERS	96.58	97.00	**96.79**	PERS	97.15	95.62	96.38

Tab 6: **Recall** comparison for **JOB** by gender category between Balanced NER and Original NER.

Category	Original NER	Gender-Balanced NER
Feminine recall	0.69	**0.80**
Masculine recall	**0.64**	0.59
Common-gender recall	0.85	**0.87**

CONCLUSIONS

Contributions

- Dataset of parallel sentences
- Gender Swapper via LLM
- Gender-swapped NER Dataset

Source code on GitHub

Models available on Hugging Face

Limitations and Future Work

- Extend gender swapping to document level with better context handling.
- Replace proprietary GPT-4o-mini to improve reproducibility.
- Fix name bias by using valid name variants.
- Expand to other gendered entities and improve model performance.

Q & A

Github:

https://github.com/linndfors/ner_for_fem

Contacts:

onahurna@gmail.com

mariana.romanyshyn@grammarly.com