**Exposé Bachelor's Thesis**

# Detecting Gender Bias in English-German Translations Using NLP

Khanh Linh Pham, 77211916753

Supervisor: Prof. Dr. Diana Hristova

Hochschule für Wirtschaft und Recht Berlin

May 30, 2025

## 1 Motivation

With the growing availability of free Machine Translation (MT) tools capable of handling complex sentences, ensuring the accuracy and fairness of these translations becomes increasingly significant. However, when translating between languages with and without grammatical gender, there is a risk of incorrect gender assignment. For example, the English sentence "The surgeon is hard-working" is translated into German as "Der Chirurg ist fleißig" using Google Translate and DeepL. This assigns a male gender instead of remaining neutral or offering both masculine and feminine options. Unbeknown to the tool user, such biased translations can amplify discrimination and inequality (Smacchia et al. 2024), given the widespread use of MT in both public and business contexts. Thus, this thesis aims to explore gender bias in English-German MT by developing a binary classification model based on Natural Language Processing (NLP). The goal is to help users identify potential bias in translations, raising awareness and promoting the creation of fairer AI systems.

Recent years have seen extensive research on bias in MT. A systematic review of the literature by Shrestha and Das (2022) offers a thorough overview of this issue, detailing its scope, impact, and key findings relevant to my thesis. Several case studies comparing English to both grammatical-gender languages (e.g., Spanish, Italian) and non-gender-marking languages (e.g., Hungarian) consistently identify biases, including the assignment of stereotypical gender roles and a default tendency to use male pronouns (Stanovsky et al. 2019; Prates et al. 2019; Smacchia et al. 2024). The methodologies used in these studies will be valuable for my own implementation. Lastly, the study by Lardelli et al. (2024) on the challenges of translating English into German lays the foundation for my analysis. It provides data for testing the translation of gender-neutral terms in context, as well as a gender-fair German dictionary, both of which I can build upon.

## 2 Research Question

To address this issue, the following research question will be investigated:

*How can a NLP-based binary classification model detect gender bias in English-German translations?*

This question can be broken down into four sub questions, each depicting a milestone of the research: **(1)** Literature review, **(2)** Data collection and preprocessing, **(3)** Implementation and **(4)** Benchmarking and discussion of results.

## 3 Methodology

**(1) Literature Review:** This thesis places a focus on the practical implementation. The primary aim is to develop a system for detecting gender bias in English-German translations. As such, the literature review does not aim to provide a comprehensive theoretical analysis. Its role is instead limited to establishing the conceptual and technical context for the implementation. An overview of how the literature is used and how sources were selected is provided in table 1.

Since Machine Learning (ML) is a rapidly evolving field and older studies may be outdated, only publications from 2015 onward are considered. Additionally, only publicly available English sources are included. Literature was primarily found using keyword-based searches on Google Scholar. Initial search terms such as "gender bias" and "machine translation" yielded over 18,000 results. Abstract screening and citation analysis helped narrow the scope to a few foundational works, including Prates et al. (2019), Stanovsky et al. (2019), and Cho et al. (2019). These studies clearly demonstrate the presence of gender bias in MT systems and its relevance in the ML context. They were chosen because they do not focus heavily on technical implementation but instead highlight the differences between gendered and non-gendered languages. Case study results are presented in a clear, accessible way, making them a solid foundation to build upon in this thesis.

From these core papers, both backward and forward citation searches were used to expand the literature base. The backward search often led to older, more conceptual texts, which help define the relevance of the topic overall. In contrast, forward citation search surfaced more applied work, including case studies on specific language pairs. For the forward search, the results were refined using more specific keywords such as “gender bias in English-German machine translation” and “occupational bias in translation”. To cover the technical side, terms like “AI” and “ML” were added. One key resource found through this process is Shrestha and Das (2022), which provides a structured overview of current research on gender bias in MT. It helps define the current state of the field and its limitations without requiring deeper engagement in literature theory. Overall, the targeted search approach revealed a clear research gap: while the existence of bias in MT is well-documented, open-source tools addressing this problem, particularly for English-German translation, are limited. Studies focusing on this language pair generally remain rare.

After establishing that bias exists, the search shifted to sources discussing its real-world consequences. These studies quantify the impact of biased translations, especially in gendered languages like German. For example, Savoldi, Papi, et al. (2024) and Bolukbasi et al. (2016) discuss how bias in MT can reinforce stereotypes or increase correction costs in post-processing.

To inform the technical design, the search was extended using Perplexity.ai. This tool supports precise prompting and is especially useful for identifying datasets, translation models, and technical papers. Prompts such as “Find datasets for training and testing a gender bias detection system for English-German translations” or “What are lightweight, open-source English-German translation models for Python?” were used. Once specific tools or models were identified, their documentation and supporting literature were manually reviewed.

So far, 17 papers have been collected to support the thesis. They cover key concepts, technical methods, and tools related to gender bias in English-German translation. This selection will be adjusted as the research develops and specific requirements become clearer. The full list of references can be found in section 7.

**(2) Data Collection and Preprocessing:** As mentioned in (1), I have collected relevant datasets from existing research. The preliminary list is shown in the table below, alongside short descriptions and content summaries.

Data preprocessing aims to ensure the quality and fairness of the data. The bias labels and translations of the chosen datasets will be validated first. This requires checking that each example has the bias

| Purpose | Method of Finding | Preliminary Sources |
|---|---|---|
| Define core concept of gender bias in MT | Keyword-based search | Prates et al. (2019), Stanovsky et al. (2019), Cho et al. (2019), Shrestha and Das (2022) |
| Establish relevance of the topic | Backward citation search | Bolukbasi et al. (2016), Savoldi, Papi, et al. (2024), Godsil et al. (2016), Smacchia et al. (2024) |
| Justify technical design decisions | Focused search via Perplexity.ai + manual search | Devlin et al. (2019), Lardelli et al. (2024), Tiedemann and Thottingal (2020), Stella et al. (2021), Savoldi, Cupin, et al. (2025), Papineni et al. (2001) |
| Identify research gap | Forward citation search | Lardelli et al. (2024), Gete and Etchegoyhen (2024), Kappl (2025), UNIOR NLP Research Group, University of Naples " et al. (2023) |

Table 1: Overview of literature use and discovery methods

| Dataset | Description | Content |
|---|---|---|
| mGeNTE en-de (Savoldi, Cupin, et al. 2025) | Multilingual Gender-neutral Translation Evaluation dataset used to assess gender bias in MT systems. | ~1,500 English sentences with gender-neutral and gendered . |
| Building Bridges Dictionary (Lardelli et al. 2024) | Curated bilingual dictionary for promoting gender-fair language in English-German MT. | ~1,000 German gender-neutral and gender-inclusive sentences and their English translations. |
| Translated Wikipedia Biographies (Stella et al. 2021) | Dataset of English Wikipedia biographies automatically translated into multiple languages, used to test gender bias in MT outputs. | ~1,500 translated biography sentences. |

Table 2: Overview of Selected Datasets

labels applied correctly and that the translations are accurate. Validation will be done through a mix of manual checks and automated checks using language models and translation systems. The dataset will then be checked for balance to make sure that gender-related terms and categories are not significantly over- or under-represented to avoid adding bias to the model. If necessary, under-represented categories will be added as needed, and over-represented categories will be decreased.

Ultimately, the goal is to combine the validated data from all datasets. The merged dataset should have at least 6,000 entries to support a proper binary classification task (Pecher et al. 2024). This dataset will be used to fine-tune the classification model explained in (3). Each data point will contain:

- English sentence "eng_source_sentence"
- German translation "ger_translation"

- Source Gender "source_gender" (male/female/neutral)
- Translation Gender "translation_gender" (male/female/neutral)
- Bias label "bias_label" (1 for biased, 0 for unbiased)

**(3) Implementation:** The focus of this thesis will be on **Fine-tuning a Pre-trained BERT Model to Detect Gender Bias in English-German Translations**. The demo will allow the user to input an English sentence, which is translated into German. If a bias is detected, the translation will be flagged.

The model will analyze translated text for the following labeling rules:

| Source Gender | Translation Gender | Label | Example |
|---|---|---|---|
| male or female | matches source gender | 0 | "She is a doctor" → "Sie ist Ärztin" |
| male or female | neutral | 1 | "She is a doctor" → "Arzt" |
| male or female | opposite gender | 1 | "She is a doctor" → "Er ist Arzt" |
| neutral | neutral | 0 | "The teacher" → "Lehrkraft" |
| neutral | gendered (male or female) | 1 | "The teacher" → "Lehrer" |

Table 3: Bias labeling rules based on source and translation gender

Fine-tuning a pre-trained multilingual BERT model was chosen over training a model from scratch due to constraints on computational resources and time. Fine-tuning enables the adaptation of a pre-trained model to a specific task, in this case, binary classification of gender bias. Since BERT has been trained on a large, multilingual corpus, it is well-suited to capture complex language patterns. This approach not only enhances performance but also accelerates training, especially when data is limited (Devlin et al. 2019). This approach improves performance and speeds up training compared to starting from scratch, especially when data availability is limited.

The implementation will be carried out in the steps outlined below:

1. **Dataset Preprocessing**
   - **Purpose:** Prepare the merged dataset for input into the BERT model.
   - **Details:** Combine the English source sentence and German translation, separating them with the [SEP] token.
   - **Example:** "She is a doctor. [SEP] Sie ist Ärztin."
2. **Tokenization**
   - **Purpose:** Tokenize the sentences, converting them into token IDs compatible with BERT.
   - **Details:** Use BERT's multilingual tokenizer to split the sentences into subword tokens.
   - **Tool:** BertTokenizer.from_pretrained("bert-base-multilingual-cased")
3. **Model Building**
   - **Purpose:** Use BERT for sequence classification with a binary classification head.
   - **Details:** Use "BertForSequenceClassification" to build the model. This adds a layer for classification on top of the pre-trained BERT model.
   - **Tool:** BertForSequenceClassification.from_pretrained("bert-base-multilingual-cased", num_labels=2)
4. **Model Training**
   - **Purpose:** Fine-tune BERT on the dataset to classify translations.
   - **Details:** Train the model using backpropagation and a cross-entropy loss function to minimize classification errors.

- **Tool:** Hugging Face's Trainer API.
- **Optional Hyperparameter Tuning:** Experiment with learning rate, batch size, and number of epochs to optimize validation performance.

5. **Model Evaluation**
   - **Purpose:** Evaluate the trained model to assess its classification performance.
   - **Details:** Use standard metrics such as accuracy, precision, recall, and F1-score.
   - **Tool:** Hugging Face's Trainer API's Trainer.evaluate().
6. **Model Saving and Inference**
   - **Purpose:** Persist the fine-tuned model and set up an inference pipeline for new inputs.
   - **Details:** Saving the model ensures reproducibility and avoids expensive re-training. An inference pipeline allows the demo to load the model, tokenize new sentence pairs and predict bias.

The technology stack includes **Python** for the backend and **Streamlit** for the web demo. I will build the application locally. Figure 1, which was created using Streamlit, illustrates what the User Interface could look like, though it does not include the bias flagging feature yet.

For the translation model, I will use **Opus-MT**. I considered two criteria in my choice: 1) Open-source availability and 2) Lightweight design for optimizing performance. While Opus-MT will be the primary translation model, I may consider adding more models or even Large Language Models (LLMs) that fit my two criteria if it is feasible within the scope of the project. Additionally, as an optional add-on, I might include a justification for the gender flag. This would better explain the reason for the flag to the user and increase the usability of the tool.

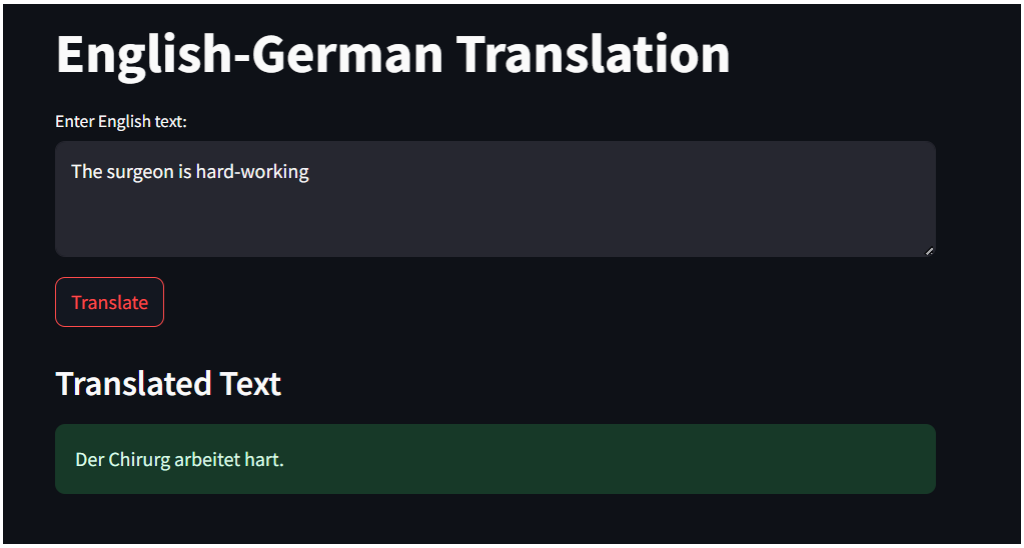



Figure 1: Frontend Example

**(4) Benchmarking and Discussion of Results:** After implementing the model, I plan to manually evaluate it by inputting various sentences and different types of bias to assess its performance. This includes the analysis of any errors to gain a deeper understanding of the model's behavior and limitations. I will highlight notable examples of both errors and successes to provide valuable insights into its strengths and weaknesses. The performance results will then be visualized using graphs.

This process also offers an opportunity to refine the implementation in order to make adjustments where needed. Moreover, I will compare my evaluation outcomes with the claims made in the analyzed literature, determining whether they confirm or challenge existing findings.

Finally, I will address the limitations of my work, outline potential improvements, and suggest what could be added to enhance the model's performance and applicability.

# 4 Expected results

**(1) Literature Review:** Summary of recent research on gender bias in MT (esp. English-German); identification of key bias types (e.g., male defaults, stereotypes); foundational explanations of NLP concepts and translation systems; research gap: lack of lightweight, explainable detection tools.

**(2) Data Collection and Preprocessing:** Cleaned, balanced dataset with bias labels, stored in one CSV for reproducibility; Python code for validation and preprocessing; insights on bias distribution across datasets.

**(3) Implementation:** Working classification model that detects gender bias in translations with a target accuracy of at least 85%; Streamlit demo with user input, bias flagging, and optional explanations.

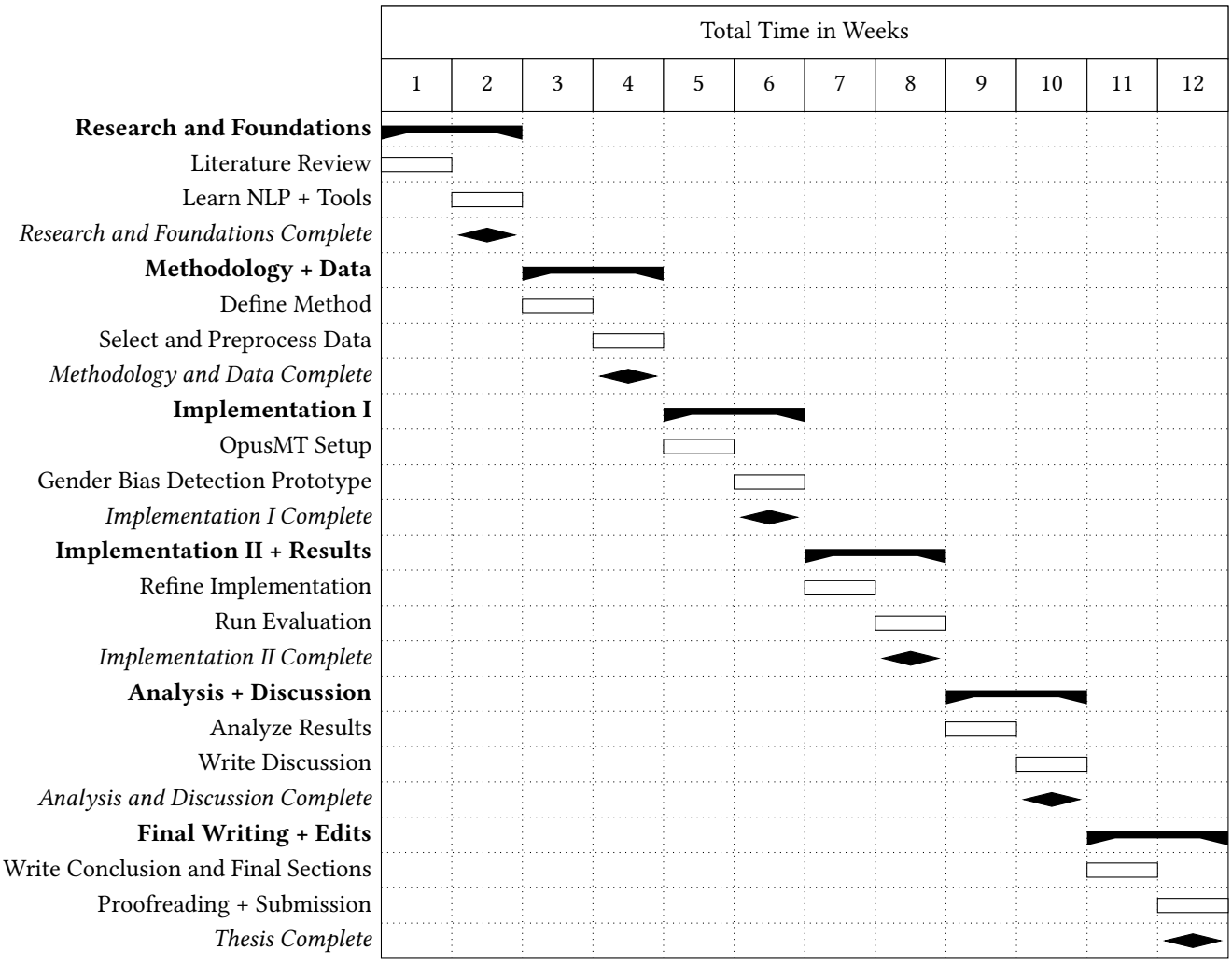
**(4) Benchmarking and Discussion of Results:** Performance metrics (accuracy, precision, recall); example successes and failures; comparison with findings from literature; discussion of limitations and future improvements.

# 5 Expected Outline

1. **Introduction**
   - Motivation
   - Research Question and Objectives
2. **Theoretical Foundations**
   - Terminology and Definitions
   - Social Implications of Gender Bias in Language
   - State of Research
3. **Methodology**
   - Research Design
   - Implementation Overview
4. **Data Collection and Processing**
   - Dataset Selection
   - Preprocessing Steps
   - Data Limitations and Challenges
5. **System Implementation**
   - Tool Architecture and Workflow
   - Integration of Translation and Bias Detection
   - UI

6. **Results and Evaluation**
7. **Discussion**
    - Interpretation of Results
    - Comparison with Existing Work
8. **Conclusion**
9. **Limitations and Future Work**
    - Methodological Limitations
    - Suggestions for Further Research

## 6 Expected Timeline




## 7 References

This is an initial selection of works relevant to the thesis and will be updated as research develops.

Bolukbasi, Tolga et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: URL: https://doi.org/10.48550/arXiv.1607.06520.

Cho, Won Ik et al. (2019). *On Measuring Gender Bias in Translation of Gender-neutral Pronouns*. DOI: 10.48550/arXiv.1905.11684. arXiv: 1905.11684[cs]. URL: http://arxiv.org/abs/1905.11684 (visited on 04/21/2025).

Currey, Anna et al. (2022). "MT-GenEval: A Counterfactual and Contextual Dataset for Evaluating Gender Accuracy in Machine Translation". In: *Association for Computational Linguistics* (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing), pp. 4287–4299. DOI: 10.18653/v1/2022.emnlp-main.288. URL: https://aclanthology.org/2022.emnlp-main.288/.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 04/09/2025).

Gete, Harritxu and Thierry Etchegoyhen (2024). *Does Context Help Mitigate Gender Bias in Neural Machine Translation?* DOI: 10.48550/arXiv.2406.12364. arXiv: 2406.12364[cs]. URL: http://arxiv.org/abs/2406.12364 (visited on 02/27/2025).

Godsil, Rachel D. et al. (2016). "The Effects of Gender Roles, Implicit Bias, and Stereotype Threat on the Lives of Women and Girls". In: *THE SCIENCE OF EQUALITY* 2 (Perception Institute). URL: https://equity.ucla.edu/wp-content/uploads/2016/11/Science-of-Equality-Volume-2.pdf.

Kappl, Michelle (2025). *Are All Spanish Doctors Male? Evaluating Gender Bias in German Machine Translation*. DOI: 10.48550/arXiv.2502.19104. arXiv: 2502.19104[cs]. URL: http://arxiv.org/abs/2502.19104 (visited on 04/10/2025).

Lardelli, Manuel et al. (2024). "Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German". In: *Findings of the Association for Computational Linguistics ACL 2024*. Findings of the Association for Computational Linguistics ACL 2024. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 7542–7550. DOI: 10.18653/v1/2024.findings-acl.448. URL: https://aclanthology.org/2024.findings-acl.448 (visited on 04/06/2025).

Papineni, Kishore et al. (2001). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. the 40th Annual Meeting. Philadelphia, Pennsylvania: Association for Computational Linguistics, p. 311. DOI: 10.3115/1073083.1073135. URL: http://portal.acm.org/citation.cfm?doid=1073083.1073135 (visited on 02/27/2025).

Pecher, Branislav et al. (Apr. 26, 2024). *Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance*. DOI: 10.48550/arXiv.2402.12819. arXiv: 2402.12819[cs]. URL: http://arxiv.org/abs/2402.12819 (visited on 04/27/2025).

Prates, Marcelo O. R. et al. (2019). *Assessing Gender Bias in Machine Translation – A Case Study with Google Translate*. DOI: 10.48550/arXiv.1809.02208. arXiv: 1809.02208[cs]. URL: http://arxiv.org/abs/1809.02208 (visited on 04/03/2025).

Savoldi, Beatrice, Eleonora Cupin, et al. (2025). *mGeNTE: A Multilingual Resource for Gender-Neutral Language and Translation*. DOI: 10.48550/arXiv.2501.09409. arXiv: 2501.09409[cs]. URL: http://arxiv.org/abs/2501.09409 (visited on 04/08/2025).

Savoldi, Beatrice, Sara Papi, et al. (2024). "What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics, pp. 18048–18076. DOI: 10.18653/v1/2024.emnlp-main.1002. URL: https://aclanthology.org/2024.emnlp-main.1002 (visited on 04/06/2025).

Shrestha, Sunny and Sanchari Das (2022). “Exploring gender biases in ML and AI academic research through systematic literature review”. In: *Frontiers in Artificial Intelligence* 5, p. 976838. ISSN: 2624-8212. DOI: 10.3389/frai.2022.976838. URL: https://www.frontiersin.org/articles/10.3389/frai.2022.976838/full (visited on 04/06/2025).

Smacchia, Marco et al. (2024). “Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools”. In: *Digital (Eco) Systems and Societal Challenges*. Ed. by Alessio Maria Braccini et al. Vol. 72. Series Title: Lecture Notes in Information Systems and Organisation. Cham: Springer Nature Switzerland, pp. 355–373. ISBN: 978-3-031-75585-9 978-3-031-75586-6. DOI: 10.1007/978-3-031-75586-6_19. URL: https://link.springer.com/10.1007/978-3-031-75586-6_19 (visited on 02/27/2025).

Stanovsky, Gabriel et al. (2019). *Evaluating Gender Bias in Machine Translation*. DOI: 10.48550/arXiv.1906.00591. arXiv: 1906.00591[cs]. URL: http://arxiv.org/abs/1906.00591 (visited on 04/03/2025).

Stella, Romina et al. (2021). *A Dataset for Studying Gender Bias in Translation*. URL: https://research.google/blog/a-dataset-for-studying-gender-bias-in-translation/ (visited on 04/10/2025).

Tiedemann, Jorg and Santhosh Thottingal (2020). “OPUS-MT – Building open translation services for the World”. In: *European Association for Machine Translation* Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61/.

UNIOR NLP Research Group, University of Naples " et al. (2023). “Gender Bias in Machine Translation: a statistical evaluation of Google Translate and DeepL for English, Italian and German”. In: *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*. International Conference on Human-informed Translation and Interpreting Technology 2023. INCOMA Ltd., Shoumen, Bulgaria, pp. 1–11. DOI: 10.26615/issn.2683-0078.2023_001. URL: https://acl-bg.org/proceedings/2023/HiT-IT%202023/pdf/2023.hitit2023-1.1.pdf (visited on 02/27/2025).

*Vorlage Expose* (2025). URL: https://www.overleaf.com/project/67ed0a04d9a03067f037255d (visited on 04/25/2025).