B. Sc. Information Systems
Berlin School of Economics and Law
Department 1: Business and Economics

Bachelor's Thesis

Detecting Gender Bias in English-German Translations using Natural Language Processing

Khanh Linh Pham

Supervisors: Prof. Dr. Diana Hristova, Prof. Dr. Markus Schaal

Semester: Summer Semester 2025
Matrikel-Nr.: 77211916753
Email: klpham04@gmail.com

Date: xx.xx.2025

Abstract

Recent years have seen extensive research on bias in MT. A systematic review of the literature by Shrestha and Das 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. 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.

Sperrvermerk

Die vorgelegte Masterarbeit basiert auf internen, vertraulichen Daten und Informationen des Unternehmens ..... In diese Arbeit dürfen Dritte, mit Ausnahme der Gutachter und befugten Mitgliedern des Prüfungsausschusses ohne ausdrückliche Zustimmung des Unternehmens und des Verfassers keine Einsicht nehmen. Von diesem Verbot ausgenommen sind außerdem jene Personen, die auch ansonsten zur Einsichtnahme in die genannten Daten und Informationen befugt sind. Eine Vervielfältigung und Veröffentlichung der Masterarbeit ohne ausdrückliche Genehmigung – auch auszugsweise – ist nicht erlaubt.

Berlin, den 01. Januar 2099

.....

(Unterschrift des Verfassers)

Contents

List of Figures

List of Tables

1 Introduction

Machine Translation (MT) is a sub-field of computational linguistic that uses computer software to translate texts between languages (Lin and Chien 2009). It is a major area within Natural Language Processing (NLP), a branch of Artificial Intelligence (AI) (Smacchia et al. 2024). This technology helps millions of people communicate across languages, whether in every situations or high-stakes domains like healthcare, law and business (Kappl 2025). Users rely on it to translate everything from casual conversations to medical prescriptions and legal documents. Tools like Google Translate serve over 200 million users daily (Prates et al. 2019; Shrestha and Das 2022), with new advanced translation models appearing on the market frequently. According to a market analysis by SkyQuest (2025), the MT market size was valued at 980 million USD in 2023 and is projected to reach 2.78 billion USD by 2023.

With this growing availability and accessibility of free MT tools capable of handling complex sentences, their use in translating large volumes of online content is increasing (Thompson et al. 2024). This not only expands their influence on global access to information, but also shapes how readers perceive and interpret that content. Automated and unsupervised translations raise new concerns: not just about quality, but also about bias. One aspect is gender bias. Several studies (Smacchia et al. 2024; Cho et al. 2019; Stanczak and Augenstein 2021; Soundararajan and Delany 2024) confirm that MT systems trained on large-scale datasets that incorporate societal biases, can learn and perpetuate gender biases present in the training data. In short, if the training data reflects gender stereotypes, the translation system is likely to repeat them. This includes, but is not limited to, the insertion of gendered pronouns in place of originally gender-neutral terms, and the use of stereotypically gendered occupational titles.

There is a risk of incorrect gender assignment when translating between languages with and without grammatical gender. For example, the gender-neutral English sentences "The surgeon is hard-working" and "The nurse is hard-working" are translated into German as "Der Chirurg ist fleißig" and "Die Krankenschwester ist fleißig" respectively, as seen in Figure 1 and Figure 2. These translations introduce occupational gender stereotypes. “Der Chirurg” is the masculine form of “surgeon” in German, adding a male gender where the

original English sentence did not specify one. Similarly, “Die Krankenschwester” is the feminine form of “nurse,” again assigning a gender that was not present in the source. These patterns are not just technical flaws. They can reinforce harmful stereotypes in real-world contexts. The following section outlines the broader motivation behind this thesis.

1.1 Motivation

Academia has come to the consensus that MT systems do default to male pronouns when gender in the source sentence is ambiguous. In addition, as shown in the earlier example where "the surgeon" and "the nurse" were translated with stereotypical genders, the reinforcement of occupational stereotypes is an increasing concern. When MT is used for job descriptions, recommendation letters, or resumes and it inserts or reinforces unfairly gendered language (Bolukbasi et al. 2016), it may discourage individuals whose gender is misrepresented or stereotyped. In turn, this would reduce their chances of success in recruitment processes. Failing to address this issue can bring broader consequences for business, leading to the exclusion of qualified candidates, reduce diversity and contradict international standards. Organizations like the United Nations, UNESCO, and the European Union stress the importance of gender equality and inclusive language, making gender equality one of the 17 Sustainable Development Goals for 2030 (Sczesny et al. 2016; United Nations 2023). Moreover, language influences how people think. There have been consistent findings that speakers do not understand masculine forms as referring to both genders equally, but interpret them in a male based way (Sczesny et al. 2016). Gender-fair language is more commonly used in official texts, muss less in private messages. Still, exposure matters. The more often people see inclusive forms, the more normal they become. Promoting awareness of these patterns is an important step toward reducing bias in society.

Current research on this topic tends to focus more on the quantitative measurement of gender bias (Rescigno and Monti 2023; Barclay and Sami 2024; Smacchia et al. 2024), e.g. counting the occurences of gendered pronouns or grammatical forms in outputs when prompting models with a neutral input. It is then often compared against a standard or desired outcome like real-world demographic distributions (Smacchia et al. 2024; Prates et al. 2019) or human evaluation (Lardelli et al. 2024; Savoldi et al. 2024). However, current evaluations are not enough for accountability. Few approaches address an active gender bias detection layer. While this gap remains in translation systems, similar issues have been addressed in other domains. For example, as summarized by Shrestha and Das (2022), Schwemmer et al. (2020) propose a detection framework to uncover gender bias in

facial recognition technologies. Their findings show that these systems are more accurate in identifying individuals as women when the images conform to stereotypical feminine features like long hair or makeup. In some cases, systems even associated such images with stereotypically gendered labels like "kitchen" or "cake," despite these elements not being present. A detection system specifically for MT would increase linguistic transparency, because without the development of bias-aware tools, problematic translations are likely to scale without oversight. Therefore, addressing gender bias in MT becomes both a social and ethical necessity.

1.2 Problem Statement and Research Questions

The core problem boils down to the significant bias towards the masculine form in English-German MTs, sometimes consituting 93-96% of translations for isolated words (Lardelli et al. 2024). These outputs often reflect social stereotypes rather than objective translations, yet current systems offer no mechanism to detect or signal when such bias occurs (Rescigno and Monti 2023). To address this, this thesis deploys a blackbox approach to explore how fine-tuning a pre-trained multilingual BERT model can help detect gender bias in MT outputs. The model takes an input sentence and its corresponding German translation and predicts whether the translation introduces gender bias. It focuses on identifying two common cases: added gendered pronouns and wrongly gendered nouns.

The translation system used is Opus-MT, an open-source neural MT model. Translations are passed through BERT, trained on a dataset I have constructed by combining and adapting several existing datasets from other researchers. The classifier is lightweight and efficient, aiming for transparent behavior and easy integration into other tools (Devlin et al. 2019). Its predictions are used to highlight biased parts in a web-based demo. The goal is not to build a perfect detector, but a working proof of concept that shows how bias can be flagged automatically. This supports more critical use of MT systems and encourages further development of bias-aware translation tools.

The main research question is: **"How can a NLP-based binary classification model detect gender bias in English-German translations?"**. This involves building a suitable training dataset, selecting features that capture bias patterns, and evaluating how well the model generalizes across different domains.

1.3 Scope and Limitations

This thesis focuses only on English-to-German machine translation due to my fluency in both languages, allowing me to evaluate the outputs and datasets directly. Extending the work to other language pairs would require native-level understanding to reliably identify subtle gender patterns and translation errors, which is beyond the current scope. More generally, gender bias in language is a complex issue altogether that goes far beyond simple word associations. It becomes especially difficult to detect when sentences contain multiple subjects, indirect references, or ambiguous pronouns. For example, as Barclay and Sami (2024) explain, the sentence "He went to see her mother" clearly implies three people, while "He went to see his mother" could refer to either two or three. These types of structures introduce ambiguity that makes annotation and evaluation much harder. Creating a dataset that captures such linguistic complexity would require significant effort and careful control of variables. One broader limitation in building datasets for complex scenarios with multiple subjects is the difficulty of isolating the influence of each gendered entity (Lardelli et al. 2024). When working with natural language sources, it becomes hard to tell what caused the bias in the translation. Because of this, the focus of this thesis is on simpler sentence structures with a single subject. This makes it easier to identify and explain bias patterns. It also fits the intended use case: translating business texts like job advertisements or reports, which rarely involve multiple nested clauses or ambiguous pronouns.

1.4 Overview of Chapters

2 Related Work

This section outlines key findings of related work on gender bias in MT, with a focus on the English-German language pair. The aim is to (1) define the core concept of gender bias in MT, (2) establish the relevance of the topic, (3) identify the research gap, and (4) justify technical design choices. To support this, I examine datasets, model types, and tools used in previous studies.

This literature review was structured according to the framework for qualitative Information Systems literature reviews developed by Smacchia et al. (2024) and further guided by the findings of Shrestha and Das (2022), who conducted a systematic literature review on gender bias in ML and AI. It served as a reference point for identifying key areas of the topic and provided a basis for defining relevant keywords and thematic categories. This process helped identify empirical findings and technical approaches relevant to this thesis and supported the positioning and justification of its methodological choices.

2.1 Literature Sourcing Methodology

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.

- refer to 03 of methodology

- Define the review's purpose: synthesis, interpretation, and guidance for future research. - Describe databases used (ACL, Google Scholar), search terms, filtering process. - Establish inclusion and exclusion criteria for selecting literature. - Utilize qualitative data analysis tools (e.g., NVivo) for coding and thematic analysis. - Ensure transparency and replicability in the review process.

2.2 Gender Bias in Machine Translation

Bibliography

Barclay, P. J. and Sami, A. (2024). Investigating Markers and Drivers of Gender Bias in Machine Translations.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Cho, W. I., Kim, J. W., Kim, S. M., and Kim, N. S. (2019). On Measuring Gender Bias in Translation of Gender-neutral Pronouns.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Kappl, M. (2025). Are All Spanish Doctors Male? Evaluating Gender Bias in German Machine Translation.

Lardelli, M., Attanasio, G., and Lauscher, A. (2024). Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7542–7550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lin, G. H.-c. and Chien, P. S. C. (2009). Machine Translation for Academic Purposes. *Proceedings of the International Conference on TESOL and Translation 2009*, pages pp.133–148.

Prates, M. O. R., Avelar, P. H. C., and Lamb, L. (2019). Assessing Gender Bias in Machine Translation – A Case Study with Google Translate.

Rescigno, A. A. and Monti, J. (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*, pages 1–11, UNIOR NLP Research Group, University of Naples "L'Orientale", Naples, Italy. INCOMA Ltd., Shoumen, Bulgaria.

Savoldi, B., Papi, S., Negri, M., Guerberof-Arenas, A., and Bentivogli, L. (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*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. (2020). Diagnosing Gender Bias in Image Recognition Systems. *Socius*, 6:2378023120967171.

Sczesny, S., Formanowicz, M., and Moser, F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7.

Shrestha, S. and Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence*, 5:976838.

SkyQuest (2025). Machine Translation (MT) Market Size, Growth & Trends Report | 2032. https://www.skyquestt.com/report/machine-translation-market.

Smacchia, M., Za, S., and Arenas, A. (2024). Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools. In Braccini, A. M., Ricciardi, F., and Virili, F., editors, *Digital (Eco) Systems and Societal Challenges*, volume 72, pages 355–373. Springer Nature Switzerland, Cham.

Soundararajan, S. and Delany, S. J. (2024). Investigating Gender Bias in Large Language Models Through Text Generation. *Association for Computational Linguistics*, Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024):410–424.

Stanczak, K. and Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation.

Thompson, B., Dhaliwal, M. P., Frisch, P., Domhan, T., and Federico, M. (2024). A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism.

United Nations (2023). Achieve Gender Equality And Empower All Women and Girls. https://sdgs.un.org/goals/goal5.

Appendix

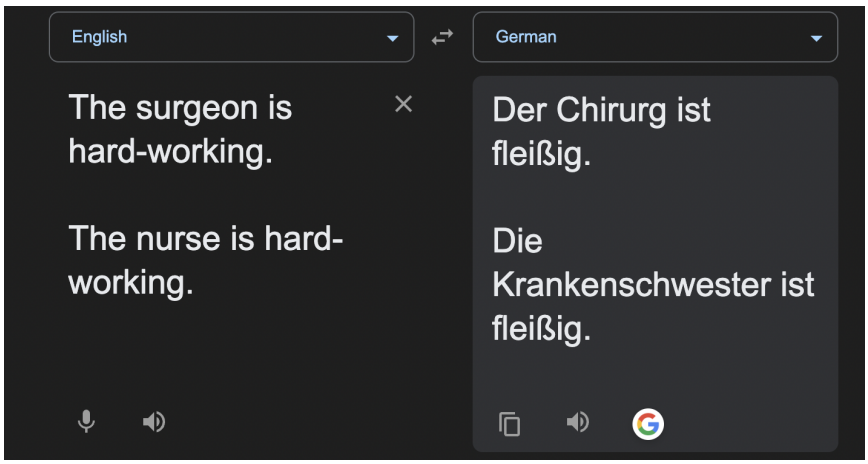

Figure 1: Google Translate assigns stereotypical genders to occupational roles.

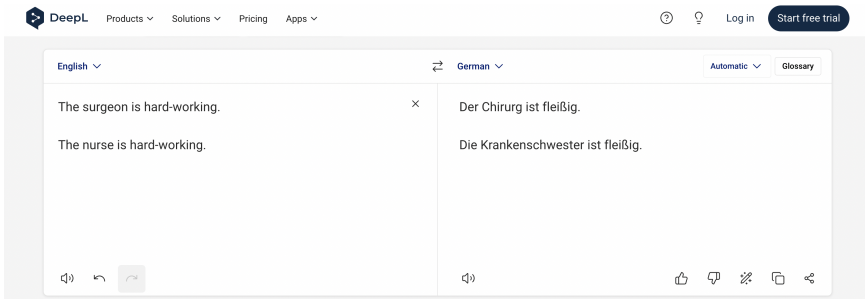

Figure 2: DeepL shows a similar bias in the same sentence, highlighting consistent patterns across MT tools.

1. Hiermit versichere ich,

   - dass ich die von mir vorgelegte Arbeit selbständig abgefasst habe,
   - dass ich keine weiteren Hilfsmittel verwendet habe als diejenigen, die im Vorfeld explizit zugelassen und von mir angegeben wurden,
   - dass ich die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken (dazu zählen auch Internetquellen und KI-basierte Tools) entnommen sind, unter Angabe der Quelle kenntlich gemacht habe und
   - dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe.

2. Mir ist bewusst,

   - dass ich diese Prüfung nicht bestanden habe, wenn ich die mir bekannte Frist für die Einreichung meiner schriftlichen Arbeit versäume,
   - dass ich im Falle eines Täuschungsversuchs diese Prüfung nicht bestanden habe,
   - dass ich im Falle eines schwerwiegenden Täuschungsversuchs ggf. die Gesamtprüfung endgültig nicht bestanden habe und in diesem Studiengang nicht mehr weiter studieren darf und
   - dass ich, sofern ich zur Erstellung dieser Arbeit KI-basierter Tools verwendet habe, die Verantwortung für eventuell durch die KI generierte fehlerhafte oder verzerrte (bias) Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate trage.

Berlin, den May 30, 2025

.....

(Unterschrift des Verfassers)