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

XX

Sperrvermerk

XX

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 linguistics that uses software to translate text between languages (Lin and Chien 2009). It is part of Natural Language Processing (NLP), which belongs to the broader field of Artificial Intelligence (AI) (Smacchia et al. 2024). MT helps millions of people communicate across languages, in daily life and in areas like healthcare, law, and business (Kappl 2025). Services like Google Translate handle over 200 million users every day (Prates et al. 2019; Shrestha and Das 2022).

The MT market is growing fast. A report by SkyQuest (2025) valued it at 980 million USD in 2023, with projections reaching 2.78 billion USD. New and more advanced translation models keep appearing, and many of them are free to use. As a result, MT tools are now used to translate large volumes of content across domains.

With this widespread use, the output of MT systems increasingly shapes how people receive and interpret information. But automatic translations are not neutral. There is growing concern about the social effects of biased translations. One key issue is gender bias. MT systems are often trained on large datasets that reflect social norms and stereotypes. If the data contains gender bias, the system will likely reproduce it (Cho et al. 2019; Soundararajan and Delany 2024; Smacchia et al. 2024).

A common case is the use of gendered terms in translations of gender-neutral input. For example, the English sentence “The nurse is hard-working” does not say anything about gender. But a translation system may render it in German as “Die Krankenschwester ist fleißig,” which uses the explicitly feminine term *Krankenschwester*. Similarly, “The surgeon is hard-working” may become “Der Chirurg ist fleißig,” using the masculine form *Chirurg*. These choices add gendered assumptions that were not present in the original. Such patterns are not just technical side effects. They can reinforce stereotypes, especially when they appear in job ads, reports, or other public texts.

1.1 Motivation

1.1.1 Social and Ethical Importance of Addressing Gender Bias

Academia has come to the consensus that MT systems do default to male pronouns when gender in the source sentence is ambiguous (Prates et al. 2019; Cho et al. 2019; Rescigno and Monti 2023). In addition, translations often reflect traditional roles, like associating "nurse" with women and "surgeon" with men. This can affect people's perceptions of jobs and reinforce gender roles.

When used in formal contexts like job descriptions or reference letters, biased translations can shape how a candidate is perceived. If a system always assigns male pronouns to leadership roles and female terms to caregiving roles, it may disadvantage those who do not match those stereotypes (Bolukbasi et al. 2016). This is not just a personal issue. It can reduce diversity and go against 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).

Language also shapes thought. Research shows that readers often interpret masculine forms as male-specific, even if they are supposed to be generic (Sczesny et al. 2016). Inclusive forms are more common in official documents, less so in everyday language. However, exposure matters. Frequent use of fair language makes it feel more normal. Detecting and addressing bias in MT can support this shift.

1.1.2 Why Detection Systems Are Needed

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). Common methods include counting gendered forms in outputs and comparing them to demographic baselines or human expectations (Rescigno and Monti 2023; Prates et al. 2019; Savoldi et al. 2024). These are useful, but they do not help users identify specific biased translations in real-time. Evaluations are not enough for accountability.

Other domains, like facial recognition, have already seen progress in active bias detection. For example, Schwemmer et al. (2020) showed that systems tend to label women more accurately if they match stereotypical appearances (e.g., long hair). Some models even linked female images to words like "kitchen" or "cake" based on bias patterns in training data. For MT, a detection layer is still missing. Without such tools, biased translations are likely to spread unnoticed. A detection system could flag potential bias in real time, improving transparency and encouraging more careful use.

1.2 Problem Statement and Research Questions

DRAFT NEED TO REWRITE AFTER IMPLEMENTATION 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.

DRAFT NEED TO REWRITE AFTER IMPLEMENTATION 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

WRITE AFTER IMPLEMENTATION PART This thesis focuses only on English-to-German (EN-DE) MT. This language pair is widely used in MT reserach, offering high-quality models and datasets. Extending the work to other language pairs would require native-level understanding to detect subtle gender patterns and translation errors, which is outside the current scope.

1.4 Limitations

WRITE AFTER IMPLEMENTATION PART 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.

1.5 Overview of Chapters

DRAFT NEED TO REWRITE AFTER IMPLEMENTATION

2 Related Work

This section outlines key findings of related work on gender bias in MT, with a focus on the English-German (EN-DE) language pair to build the theoretical knowledge base. The research aims are 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.

For the literature review I combined incremental and conceptual literature review methods, where each source led to the identification of the next. Based on this progression, I identified key concepts and used them to organize and interpret the literature, aligning with a conceptual approach. The structure followed the qualitative Information Systems framework by Schryen (2015) and was further informed by Shrestha and Das (2022) and Savoldi et al. (2025), who both conducted systematic reviews on gender bias in ML and MT respectively.

2.1 Literature Search Process

2.1.1 Search Sources and Tools

Sources were primarily searched on Google Scholar and Perplexity, which served as an additional search engine. Prompts and outputs from Perplexity have been saved and are included in the appendix. To organize and manage the collected sources, Zotero was used throughout the process.

2.1.2 Literature Review Framing

To answer the four research aims, I have defined the key concepts in Table 2.1. Key search terms consisted of *gender bias*, *machine translation*, *AI*, *machine learning*, *German*, *stereotypes*, and *detection*. The focus was on literature published between 2019 and 2025 to maintain relevance and currency, while foundational and definitional works from earlier periods were selectively included. The initial search for the term *gender bias in machine translation* returned over 18,000 results. Through my iterative selection process, this was narrowed down to 34 core sources.

Key Concept	**Description**
Foundations of Gender Bias in Natural Language Processing	Traces early research that identified gender bias in language. Focuses on foundational studies that showed why the issue matters and how later work builds on these findings.
Sources and Manifestations of Bias	Explains how stereotypes shape language and persist over time. Describes how societal bias enters training data, model design, and system feedback. Shows how bias appears in machine translation and everyday language.
Linguistic Challenges in English-German Translations	Explores key grammatical differences between English and German that affect translation. Focuses on how the lack of gender in English and its presence in German can lead to biased outputs.
Mitigation Strategies and Current Limitations	Reviews how current research tries to reduce gender bias in NLP. Highlights what these methods can and cannot do. Helps identify where a classification-based approach could fill gaps and improve bias detection in translations.

Table 2.1: Key concepts relevant to this thesis

2.1.3 Citation Tracking

Backward citation searching involved reviewing references cited by selected papers, prioritizing frequently cited and foundational works relevant to gender bias in MT. Forward citation searching used Google Scholar's "cited by" function to identify newer research citing those key papers. Filtering with specific terms (e.g., *German* and *machine translation*) was applied during forward search to maintain focus. In addition to the main review process, supplementary sources were included as needed throughout the writing phase. These consist of contextual references, statistics, or secondary citations that support specific points but were not part of the core conceptual or methodological framework.

2.1.4 Selection Criteria and Screening Process

Titles and abstracts were manually screened to select relevant studies. **Inclusion criteria** required sources to specifically address gender bias in MT, provide examples or discussions of gender-related errors, or explain the significance of gender bias in this context. Sources also had to be available in full text without access restrictions. **Exclusion criteria** filtered out studies focusing on general NLP bias without a direct link to MT, non-gender biases, and highly technical papers lacking contribution to the general understanding of gender bias or that did not provide additional knowledge beyond what was already found in previously published papers. Full texts were reviewed after initial screening to confirm relevance and extract insights. Redundant sources not providing new perspectives aligned with the thesis goals were excluded.

2.2 Foundations of Gender Bias in Natural Language Processing

This section outlines why gender bias is a subject of research in the first place and where it connects to broader social and ethical questions. It first looks at early studies that brought attention to gender patterns in language technologies and raised awareness of their social impact. Understanding these origins helps explain why it continues to be relevant today.

2.2.1 Foundational studies

The existence of gender bias in MT is well-documented. First mentions of this issue date back to over a decade ago, having been recognized by a paper by Schiebinger in 2014. Since then, there has been a general increase in research papers focusing on this topic, especially between 2019 and 2023 (Savoldi et al. 2025).

Prates et al. (2019) conducted a large-scale quantitative study using Google Translate, translating sentences such as "He/She is an engineer" from twelve gender-neutral languages into English. The study revealed a significant overrepresentation of male pronouns, particularly in STEM-related occupations. This was not attributable to actual gender distributions in the labor market, suggesting that the bias stemmed from imbalances in the system's training data. The paper received widespread media coverage, which was then followed by a policy change by Google to present both feminine and masculine official translations for ambiguous queries (Google 2018), acknowledging that their model inadvertently replicated gender biases (see Figure 3).

Following that, **Stanovsky et al. (2019)** introduced WinoMT, a challenge set designed to evaluate gender bias in translations of English sentences into eight target languages with grammatical gender. The study showed that both commercial and academic MT systems failed to preserve correct gender in non-stereotypical roles, while performing better on stereotypical ones. In line with the findings of Prates et al., the study demonstrated a systematic preference for traditional gender roles in translations. This pattern is further supported by **Cho et al. (2019)**, who showed that occupational terms exhibit higher levels of gender bias across systems compared to other semantic categories.

These foundational studies not only confirm the existence of systematic gender bias in MT outputs, but also lay the groundwork for subsequent research that builds upon their findings and approaches to develop more robust evaluation methods and mitigation strategies.

2.2.2 Human-Centered Studies

Not until half a decade later, studies have begun to assess the real-world implications of gender bias by measuring its impact on human effort. Before that, the human component has mostly been neglected. Savoldi et al. (2024) conducted a human-centered evaluation in which approximately 90 participants were tasked with post-editing MT outputs to ensure gender-accurate translations.

The study employed behavioral metrics such as time to edit and the number of edits, measured through human-targeted error rate, to quantify the effort required. The results showed that post-editing feminine translations required nearly twice as much time and four times the number of editing operations compared to masculine counterparts. Consequently this effort gap also translates into **higher economic costs**, suggesting a measurable **quality-of-service disadvantage that disproportionately affects women**. Savoldi et al. concluded that current automatic bias metrics do not sufficiently capture these human-centered disparities, emphasizing the need for evaluation methods that reflect real user experience.

A comparison analysis between AI and human translations was conducted by Smacchia et al. (2024). The study's aim was to understand if gender bias is still present in how people think in society. Their results demonstrated a consistency between the outcomes generated by the AI tools and the human survey responses, suggesting that **AI tools reflect human behaviour** regarding job occupations and gender distributions in society. They also identified a "converging bias", which is a tendency to maintain consistency in the output based on an initial translation. For example, if the *doctor* in *"The doctor arrived"* is translated with a male form, the subsequent input *"The doctor then started working"* is likely to be translated as male too.

2.3 Sources and Manifestations of Bias

To address a problem, one needs to understand its origins. This section outlines how societal bias transfers into data and NLP systems. It looks at different types of bias, with a focus on those that shape model behaviour and outputs.

2.3.1 Types of Technical Bias

Technical bias happens when limitations in the system's design affect how models learn and make predictions (Stanczak and Augenstein 2021). This includes the way data is processed or how the model handles information.

Shah et al. (2020), as described by Ullmann (2022), differentiates between four technical origins of biases:

- **Selection Bias:** Happens when the training data does not reflect the context in which the model is used (e.g., using Wikipedia data for detecting harmful language on Twitter).
- **Label Bias:** Occurs when annotations in the dataset are incorrect or skewed. This can be influenced by the annotators' own biases or lack of awareness of diverse linguistic expressions.
- **Model Overamplification:** During training, models can exaggerate patterns found in the data. This is the cause of the aforementioned "only women cook" example (see subsection 2.3.4).
- **Semantic Bias:** Stems from associative relationships within the data, where certain words or phrases are frequently co-occurring with specific genders (e.g., "he" with "doctor").

2.3.2 Human Bias Transfer

Stereotypes and gender roles stem from historical and cultural perceptions of men's and women's societal roles, many of which are obsolete but still influential. For example, when men and women often take on different roles at work and at home, it shapes how people think about their personalities and qualities. **Correspondence bias** can emerge, where people infer attributes from observable behaviours (Godsil et al. 2016). It is a result of the human brain's automatic categorization of stimuli when faced with incomplete information, often a quick and unconscious process. Common groupings are gender, race and age. These associations can be reinforced by popular media, such as TV and advertisements (Godsil et al. 2016), just as much as it can be influenced by modern technology like MT tools.

Similarly to how humans are shaped by their environemnt, ML models learn from data they are trained on. **Biases are thus reflected and reinforced in the final models** (Stanczak and Augenstein 2021; Smacchia et al. 2024). That bias can, again, reflect back to humans and create a regressive feedback loop (Barclay and Sami 2024; Shrestha and Das 2022). LLMs like GPT-3 were trained on hundreds of billions of words, making it practically impossible to review all of the data, therefore allowing misinformation or offensive content to be reproduced by the system.

Ullmann (2022) notes that the scale of training data (e.g., 175 billion parametres for GPT-3) makes it practically impossible to review all of it, allowing misinformation or offensive content to be reproduced by the system. The author also points out that platforms like Wikipedia and Reddit are male-dominated and often contain harmful or false content, which contribute to gender bias.

2.3.3 Common Gender Bias Phenomena

While the previous sections discussed the sources of gender bias, this section focuses on how such biases appear in MT outputs. Drawing on selected studies, I identify recurring patterns of bias and group them into thematic categories. An overview of the result is prestented in Table 2.2. It is important to note that, due to the complex nature of gender bias, clear boundaries between these categories are not always possible. Definitions in the literature often overlap, and some patterns may extend or intersect with others. The classification presented here reflects my synthesis of the existing research, aiming to provide a structured framework for analysis despite these inherent ambiguities.

Stereotype-driven bias

The majority of gender bias mentioned fall into stereotype-driven categories. Most notably, the persistent use of the **generic masculine** as default (see subsection 2.4.2) and systematic **occupational stereotyping**. When translating gender-neutral terms, most systems overwhelmingly favor masculine forms, e.g., translating "doctor" to masculine forms like "el médico" (Spanish) or "der Arzt" (German) even in neutral contexts (Smacchia et al. 2024; Cho et al. 2019; Prates et al. 2019). This bias becomes particularly evident in professional contexts, where translation outputs closely mirror societal stereotypes rather than linguistic requirements. Smacchia et al. (2024) found that across major commercial systems including Microsoft Azure, DeepL, and Google Translate, male-dominated occupations were translated using masculine forms in **94%** of cases.

Stereotype-driven bias extends beyond occupational terms. As demonstrated by Prates et al. (2019), adjective associations exhibit similar gendered patterns. "Shy" or "desirable" disproportionately trigger feminine pronouns, while "guilty" or "cruel" default to masculine forms. These associations occur independently of professional contexts, revealing deeper linguistic biases embedded in translation systems.

Contextual analysis

Current MT systems exhibit fundamental limitations in handling gender context, particularly when compared to human translation capabilities. Where professional translators successfully interpret both linguistic markers (pronouns, grammatical agreements) and extra-linguistic knowledge (cultural norms, real-world contexts) to determine gender (Rescigno and Monti 2023), automated systems struggle with consistent implementation.

Coreference resolution is the technical process that should leverage contextual cues for gender. In translation, this helps systems use the right gender based on context. But as shown by Cho et al. (2019), many systems still struggle with this, especially when the gender is mentioned earlier in the text or in another sentence. Datasets like WinoMT (Stanovsky et al. 2019) show that translation models often miss these links and ignore gender cues, leading to biased results.

Sometimes models even actively disregard *explicit* gender markers in the immediate sentence. For example:

- "The doctor asked the nurse to help **her**" → incorrectly translated with masculine "el doctor" (Spanish) despite the feminine pronoun.
- "The doctor finished **her** shift" → rendered as masculine "der Arzt" (German) by Google Translate (Stanovsky et al. 2019)

2.3.4 Implications of Gender Bias in Natural Language Processing

As outlined in subsection 1.1.1, this section builds on the social and ethical foundations by examining how gender bias in NLP can lead to the amplification of existing social biases (Rescigno and Monti 2023). Ullmann (2022) illustrates this with an example: if a dataset predominantly associates cooking with women, the system may amplify this pattern, reinforcing the assumption that cooking is an activity exclusive to women. This not only reproduces but also strengthens a social stereotype, potentially resulting in

Bias Type	**Definition**	**Example**
Stereotype-Driven Biases		
Occupational Stereotyping	Gender assigned to occupations based on societal stereotypes rather than context	"Engineer" → masculine; "nurse" → feminine, regardless of source
Beyond Occupations	Gender bias appears in adjective associations and other non-professional contexts	Shy/desirable" → female pronouns; "guilty/cruel" → male pronouns
Generic Masculine	Masculine form used by default in gender-ambiguous cases	"The students are late" → "Die Studenten (m) sind zu spät". translated as male, even when gender is unclear
Contextual Failures		
Coreference Resolution Failures	MT fails to track gender associations across sentences/phrases	"Anna is a teacher. **She** teaches math." → "teacher" translated as male
Ignoring Contextual Cues	MT actively disregards explicit gender markers in the immediate context	"She is a baker" → translated as male baker

Table 2.2: Summary of common gender bias manifestations in MT systems.

representational harm, namely, the continued spread of reductive or biased portrayals of a particular gender (Stanczak and Augenstein 2021).

This also contributes to the invisibility of women in professions traditionally dominated by men (Kappl 2025). Studies show that gender bias in machine-generated text, such as children's stories or job advertisements, can **influence how young people view themselves** (Soundararajan and Delany 2024; Kappl 2025). It may shape their interests, hobbies, and decisions about education and careers. This effect is especially noticeable in Science, Technology, Engineering, and Mathematics (STEM) fields (Prates et al. 2019), where stereotypes are more deeply rooted. When job descriptions or mock interviews use gender-exclusive pronouns, women report feeling less sense of belonging, lower motivation, and weaker identification with the role (Godsil et al. 2016). As a result, they may self-select out of the application process, reducing the pool of female talent available to employers and

reinforcing existing gender gaps in the workforce.

On the other hand, research shows that using gender-inclusive language, e.g. "she and he" or "one", can lead to more positive reactions from women when considering job opportunities. It helps reduce stereotype threat and improves how women perceive and engage with different environments (Godsil et al. 2016). Hence, complying with gender-inclusive langauge may offer companies both social and competitive advantages.

2.4 Linguistic Challenges in English-German Translation

Because this thesis focuses on EN-DE translation, we need to take a closer look at how gender works in German. This section will explore the main linguistic features that affect gender expression in German. Then, it will review key studies that examine how these features influence bias in translation.

2.4.1 Grammatical Gender

Although both English and German originate from the Indo-European language family (Baldi 2008), they have different characteristcs. English does not assign grammatical gender to nouns. The article "the" is used universally, independent of what it refers to. On the contrary, German assigns one of three grammatical gendered articles to nouns: "der" (m), "die" (f) and "das" (n). The form or ending of a noun may also change depending on its grammatical gender. While English has a few gendered word pairs, such as "actor" (m) and "actress" (f), gender distinctions in German apply broadly across the entire noun system. "Der Student" refers to a male student, whereas "die Studentin" refers to a female student. Note that grammatical gender has no connection to societal or biological gender. It is a rule of the language rather than a reflection of identity. For example, the German word Mädchen (girl) is grammatically neuter and takes the article "das". This is not because the referent lacks gender, but because the suffix "-chen" automatically assigns neuter gender. This illustrates that grammatical gender in German follows structural rules, even when they contradict real-world gender associations.

2.4.2 Gender-Fair Language

The Generic Masculine

In both singular and plural contexts, the *generic masculine* refers to the default use of the masculine grammatical gender. It is commonly used in spoken German (Lardelli et al.

2024; Schmitz 2022), although research has consistently shown that the generic masculine creates a male bias in mental representations, leading readers or listeners to think more of male than female examples (Sczesny et al. 2016). Similarly, Rescigno and Monti (2023) observed a predominance of masculine forms in translation outputs (approximately 90% in Google Translate and 85–88% in DeepL for EN-IT and EN-DE), even when the original sentences contained relatively few masculine references. These linguistic biases in human language naturally carry over into ML systems. Since most models for NLP are trained on large datasets of human-generated text, they inadvertently learn and reproduce the same sociolinguistic biases present in the data (Cho et al. 2019).

All students are male

The English sentence "The students are studying" does not indicate the genders of the individuals involved. There are various ways to translate that sentence into German. The plural forms of the gendered term "student" would be "die Studenten" (multiple male students) and "die Studentinnen" (multiple female students). The problem arises when there is a mix of male and female students or when the genders are unknown. Using the common generic masculine, the sentence translates to *die Studenten lernen*, with the male term referring to a (potentially mixed-gender) group. As Schmitz (2022) pointed out, if the female form is not explicitly mentioned, the phrase is understood as all students are male.

The rise of the gender-fair language (GFL) debate was a response to this structural asymmetry. It refers to the use of language that treats all genders equally and aims to reduce stereotyping and discrimination (Sczesny et al. 2016). There are four main approaches to GFL in German identified by Lardelli et al. (2024). I will not discuss two of them further because they are less common and face greater hurdles for broader public and professional acceptance.

- **Gender-neutral rewording:** This uses neutral terms instead of gendered nouns, e.g., *die Studierenden lernen*. A challenge for this version is that neutral alternatives do not exist for every noun and cannot be consistently applied.
- **Gender-inclusive characters:** This combines masculine and feminine forms using a character like ***, *:*, or _, e.g., *die Student*innen lernen*. This method is consistent but may interrupt reading flow and lacks standardization.

Another approach not mentioned by Lardelli et al. is to simply name both forms (pair form), e.g., *die Studenten und Studentinnen lernen*. It is currently the most used GFL form

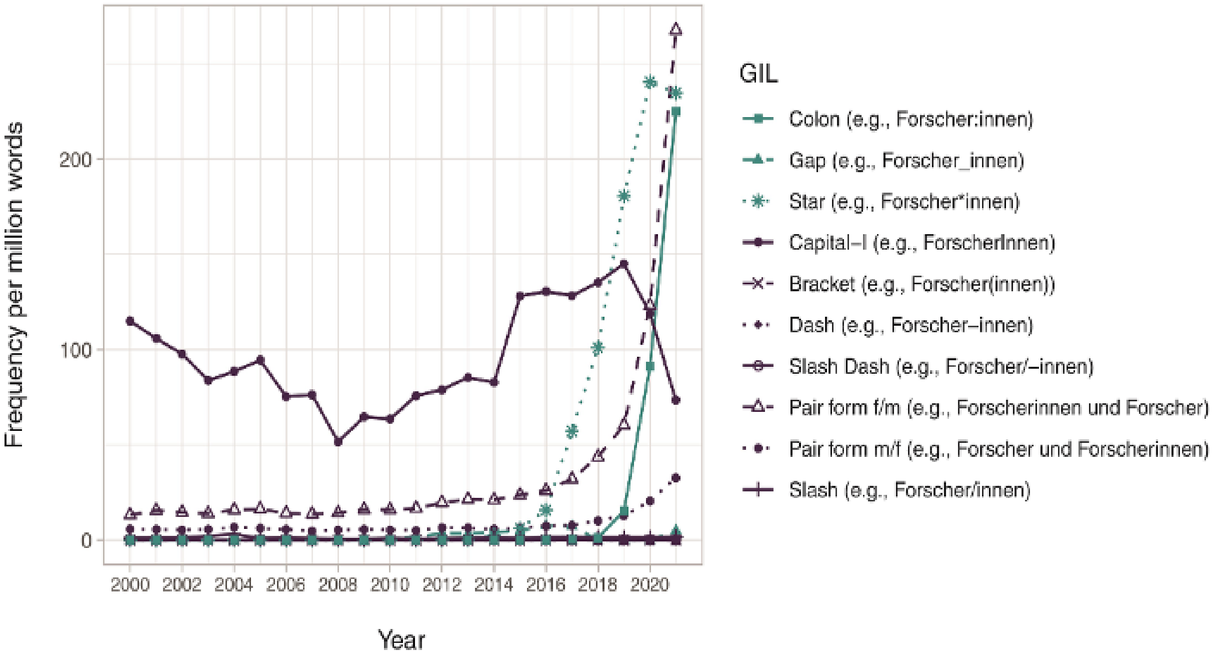


Figure 2.1: Frequency of different types of gender-inclusive language. Source: Waldendorf (2024).

in German (Waldendorf 2024), briefly surpassing the star and colon characters as seen in Figure 2.1.

2.4.3 English-German Studies

This language pair in particular is sparsely analysed in academia. I found **four relevant papers about gender bias in EN-DE MT** that fit my inclusion criteria defined in chapter 2.1.4. Some other sources include German among multiple target languages (e.g., Stanovsky et al.'s foundational study), but these do not provide detailed analysis specific to German. Therefore, I do not consider them EN-DE focused sources. The following studies provide a closer look at gender bias specifically in this language pair.

Ullmann (2022) performed a corpus-linguistic analysis of training data, meaning they studied large collections of text to identify patterns and structures related to gender bias. The dataset consisted of 17.2 million sentence pairs sourced from *Common Crawl*. They then tested different techniques to reduce gender bias in a MT system trained on that corpus. Their findings support the broader patterns discussed in this thesis: masculine forms dominate by default, gender stereotypes shape translations, and professions are translated in line with societal roles. Their key contribution lies in testing mitigation strategies. They show that fine-tuning with a small, gender-balanced dataset can reduce bias in MT outputs.

Rescigno and Monti (2023) evaluated gender bias in Google Translate and DeepL for EN-IT and EN-DE using the MT-GenEval dataset. They focused on how often professions were translated with male or female forms, both with and without gender-revealing context. Without context, both systems defaulted strongly to masculine forms (over 85%) for both languages. Contextual information generally improved alignment with reference translations, but in a few cases, context led to incorrect gender disambiguation that had not occurred without it. This suggests that contextual cues can occasionally misguide the system rather than improve performance. The authors also noted that most users are unaware of gender bias, especially if they lack fluency in the source language. Currently there is no system in place to inform them when biased translations occur.

Lardelli et al. (2024) created a Gender-Fair German Dictionary that includes professions and common nouns for people. They tested several MT systems and evaluated translations from Wikipedia and parliamentary texts. Translations were manually annotated as masculine, feminine, gender-neutral, or gender-inclusive. They also used zero-shot detection with GPT models, where GPT tries to identify gender fairness without specific training. Results showed strong masculine bias and poor automatic detection of GFL, requiring human review and therefore proving zero-shot detection to be challenging. Unlike most research focusing on professions, this study covers a broader range of terms.

Kappl (2025) introduced WinoMTDE, a German gender bias evaluation test set based on Stanovsky et al. 2019's WinoMT. It contains 288 balanced German sentences with clearly gendered subjects and tests occupational stereotyping in MT from German to other gendered languages. The study found that gender bias persists due to model architecture and training data, not source language ambiguity. Major limitations of the study include the small dataset size and broad occupation categories. It also misses some bias types and faces alignment issues affecting accuracy estimates. They call for future researchers to expand the dataset, improve annotations and include diverse gender terms. This study's evaluation pipeline is the most advanced among EN–DE studies for automated bias detection.

2.4.4 Cross-Language Perspectives

Gender bias in MT is not limited to English and German. Many other language pairs show similar patterns, revealing how bias is shaped by both language and the systems behind it. This section includes a few examples from other languages to show that the issue is not specific to German in order to keep the broader context in mind and avoid a narrow, language-specific perspective.

Some studies looked at **back-translation from English through gender-neutral languages** like Finnish, Indonesian, and Turkish, then back to English. They found different pronoun patterns depending on the language. This shows why it is important to study many languages to understand gender bias better. Verbs played a big role in how gender was inferred in translations. New metrics, like Adjusted Uncertainty, helped capture these details. Some translation systems showed signs of reducing bias over time (Barclay and Sami 2024).

When translating **gender-neutral Korean into English**, MT systems often leaned toward masculine pronouns. This happened because the training data had more male examples. Some systems made technical changes that sometimes favored feminine forms, which suggests bias mitigation is possible, however ideally, translations should stay neutral or balanced (Cho et al. 2019). **Japanese and Chinese** demonstrated exceptionally low percentages of female pronouns in translations, going as low as 0.196% for Japanese and 1.865% for Chinese (Prates et al. 2019).

Even when translating **between languages that both use grammatical gender**, like German and Spanish, Ukrainian, or Russian, gender bias still shows up (Kappl 2025). This goes against the assumption that clear grammatical cues would reduce ambiguity and help systems make better choices. Instead, the bias often stays or even gets worse, suggesting that the problem is not just about language structure but also how MT systems learn and generalize from data.

Most studies focus on English paired with another Western language, with only a few exceptions including West or East Asian languages. This adds an Anglocentric bias to the existing gender bias problem (Savoldi et al. 2025).

2.5 Mitigation Strategies and Current Limitations

Different approaches have been tested to mitigate gender bias in MT. Despite of various proposals, no single solution has emerged as definitively superior (Savoldi et al. 2025). The following section gives an overview of these strategies, takes a closer look at one selected approach, and highlights key limitations in current research.

2.5.1 Technical Mitigation Approaches

Savoldi et al. (2025) recently grouped the mitigation approaches suggested in the past decade of research about gender bias in MT.

A common focus is to create new test sets or ways to measure bias in MT. Instances of these are WinoMT, MT-GenEval and GeNTE. They serve the purpose of determining the extent of gender bias present. Usually these approaches incorporate statistical evaluations or bias metrics, which can then be used for actual mitigation / detection systems. A few papers compare different MT systems or add additional types of input like a document level approach or image guided MT. It tests whether changing the system's structure and/or adding more context can reduce gender bias. However, as previously stated in subsection 2.3.3, many systems still struggle with coreference resolution.

Stepping inside the realm of LLMs, zero-shot detection has been deployed to automatically evaluate outputs regarding gender bias. Zero-shot in this case is the prompting of GPT models to identify gender bias in the translated text without providing specific examples nor fine tuning the models. The findings suggest that the technology is not yet ready to reliably detect biased or neutral instances without human oversight (Lardelli et al. 2024).

Furthermore, by extending the reserach of Tomalin et al. (2021), Ullmann (2022) concerned herself with the pre-processing of data. The approach is to manipulate the training data *before* it is fed into a ML model. This again can be divided into three strategies: (1) Downsampling, which removes data until the ratio of gendered terms is balanced, (2) Upsampling, which duplicates data to balance the ratio of gendered terms and (3), Counterfactual Augmentation by introducing opposite sentences of the under-represented terms. For example, if one corpus contains "He is a doctor", the counterfactual sentence "She is a doctor" would be added (Ullmann 2022). All of the three strategies led to substantially worse translation performances. It has been proven that the implementation of pre-processing is not feasible if the overall translation quality is significantly compromised.

Generally, all solutions operate in a narrow area, not across all languages, types of biases and systems. This again proves the sheer difficulty of finding a fix to such a multifaceted issue spanning multiple disciplines. One approach, however, has shown more promise than others in balancing bias mitigation and translation quality: model adaptation.

2.5.2 Model Adaptation as a practical solution

Model adaptation (or domain adaptation) is the fine-tuning of a MT system *after* it has been trained. It was introduced as a response to the pre-processing approaches yielding subpar results (Tomalin et al. 2021).

This technique, as described by Tomalin et al. (2021), makes use of a small gender-balanced dataset called "Tiny", containing 388 sentence pairs which were either profession-based or adjective-based. The structures of the sentences are simple and follow the following scheme:

"The [PROFESSION] finished [his/her] work" or *"The [ADJECTIVE] [man/woman] finished [his/her] work"*. In order to prevent "catastrophic forgetting", a result in which the model loses its performance on the original data while learning from the new dataset, Elastic Weight Consolidation (EWC) was applied. It helps the model maintain its general translation quality while still working towards the reduction of gender bias.

This approach is particularly effective because manually removing biases from massive corpora is far too computationally intensive and unsustainable to be a reasonable solution. In contrast, model adaptation requires only a small, curated dataset, making it a more feasible and scalable solution worth further investigation.

2.5.3 What counts as fair?

One major limitation is that gender bias is not yet fully discussed in society or in language studies, so there is **no agreed standard for gender-fair language** (GFL) (Lardelli et al. 2024; Savoldi et al. 2025) and "fairness" heavily depends on personal views, culture, and context. Generally said, bias lies on a spectrum and changes with the chosen definition. Some argue that removing all biases is impossible (Ullmann 2022); which, for instance, leads to questions about group fairness and individual fairness. Group fairness seeks to achieve the same statistics for all groups. Individual fairness aims for similar treatment of similar people. For example, in hiring, group fairness might enforce equal hire rates for men and women. Individual fairness might enforce equal chances for two equally qualified applicants, regardless of gender. These aims can conflict, where many settings cannot satisfy both at once.

Due to the unclear definition in academia, I have to define what I consider fair to set a clear direction. In this thesis, I focus on reducing harmful bias rather than chasing a fully unbiased state. I deem fair **a system that does not predict gender incorrectly when the correct gender is clear**. This choice guides my work.

Further challenges are ethical and linguistic considerations, which I will not further elaborate in this section. See more under

3 Conceptual Frameowrk

3.1 What type of gender bias will i refer to and why

3.2 What is Machine Bias?

3.3 Binary Classification in NLP

3.4 Pre-trained Language Model: BERT

Bibliography

Baldi, P. (2008). English as an Indo-European Language. In Momma, H. and Matto, M., editors, *A Companion to the History of the English Language*, pages 127–141. Wiley, 1 edition.

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.

Godsil, R. D., Tropp, L. R., Goff, P. A., Powell, J. A., and MacFarlane, J. (2016). The Effects of Gender Roles, Implicit Bias, and Stereotype Threat on the Lives of Women and Girls. *THE SCIENCE OF EQUALITY*, 2(Perception Institute).

Google (2018). Reducing gender bias in Google Translate. https://blog.google/products/translate/reducing-gender-bias-google-translate/.

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., Bastings, J., Bentivogli, L., and Vanmassenhove, E. (2025). A decade of gender bias in machine translation. *Patterns*, page 101257.

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.

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature*, 507(7490):9–9.

Schmitz, D. (2022). In German, all professors are male.

Schryen, G. (2015). Writing Qualitative IS Literature Reviews—Guidelines for Synthesis, Interpretation, and Guidance of Research. *Communications of the Association for Information Systems*, 37.

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.

Shah, D., Schwartz, H. A., and Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264.

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.

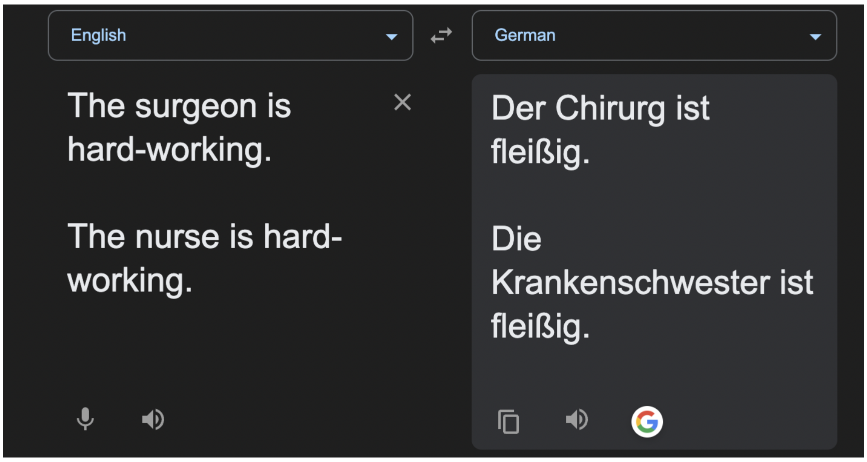
Tomalin, M., Byrne, B., Concannon, S., Saunders, D., and Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, 23(3):419–433.

Ullmann, S. (2022). Gender Bias in Machine Translation Systems. In Hanemaayer, A., editor, *Artificial Intelligence and Its Discontents*, pages 123–144. Springer International Publishing, Cham.

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

Waldendorf, A. (2024). Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, 40(2):357–374.

Appendix



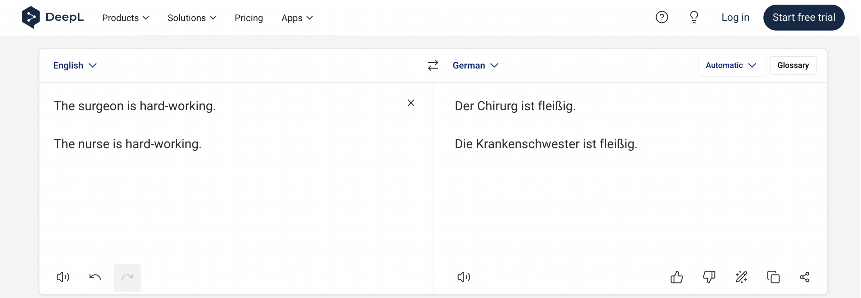
The surgeon is hard-working.

The nurse is hard-working.

Der Chirurg ist fleißig.

Die Krankenschwester ist fleißig.

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



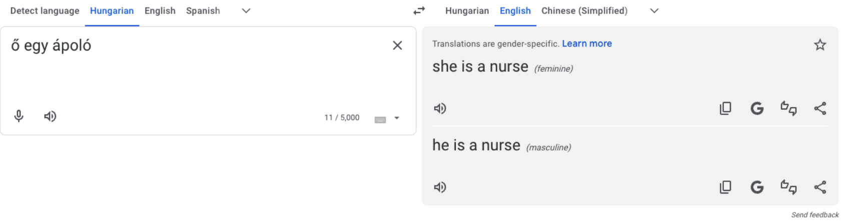
The surgeon is hard-working.

The nurse is hard-working.

Der Chirurg ist fleißig.

Die Krankenschwester ist fleißig.

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



ő egy ápoló

Translations are gender-specific. Learn more

she is a nurse *(feminine)*

he is a nurse *(masculine)*

Figure 3: Gender-specific translation by Google Translate for ambiguous pronouns.

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 June 27, 2025

. . . .

(Unterschrift des Verfassers)