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) 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). It is a fast-growing market. 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 This thesis focuses on gender bias in English-to-German (EN-DE) MT. This language pair is widely used in research, with many open datasets and high-quality models available. It also involves a grammatical shift: English has limited gender marking, while German assigns gender to many nouns and pronouns. This structural difference makes gender bias more visible and easier to study in the translation outputs.

The core problem boils down to the significant bias towards the masculine form in EN-DE 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.

The translation system used is Opus-MT, an open-source neural MT model. It is widely used in research, supports EN-DE translation, and is trained on real-world corpora, making it suitable for studying translation bias (Tiedemann and Thottingal 2020). Translations are then 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). The final tool highlights biased parts in a simple web demo. The goal is not a perfect classifier but a working prototype that shows how such detection could be integrated into translation workflows.

The main research question is therefore: **"How can a NLP-based binary classification model detect gender bias in English-German translations?"**.

1.3 Scope

WRITE AFTER IMPLEMENTATION PART This thesis focuses only on EN-DE MT. Other language pairs are out of 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

WRITE AFTER IMPLEMENTATION PART

2 Theoretical Background and 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.

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*, which were combined with *AND/OR*. 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. Beyond these systematic methods, I also included supplementary sources when needed while writing. These consist of contextual references, statistics, or secondary citations that support specific points but were not part of the core conceptual or methodological framework. Supplementary sources were defined as materials identified outside the systematic search, such as papers found through backward citations or targeted queries for statistics and news, which provided support for subordinate arguments without being central to the study's theoretical or analytical structure.

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 Understanding Gender Bias in English-to-German Machine Translation

This section explains the key terms and concepts needed to understand gender bias in English-to-German MT. It defines important ideas like natural language processing (NLP), MT, and gender bias. These concepts provide the background necessary to follow the thesis.

2.2.1 Differentiating between Natural Language Processing and Machine Translation

NLP refers to the development of machine systems that can process and generate human language. The goal is to mimic and understand it as fluently as possible (Smacchia et al. 2024; Ullmann 2022). Common applications are chatbots, translation tools, speech recognition, and image captioning.

MT is a direct application of NLP. It is used to automatically translate text from one language to another (Lin and Chien 2009). MT systems have gone through several stages of development; earlier approaches like rule-based and statistical MT used manually defined grammar rules or pattern matching from large translation corpora (Chakravarthi et al. 2021). For example:

> "The girl reads a book" → "Das Mädchen liest ein Buch"
>
> Rules: "girl" → "Mädchen", "reads" → "lesen", "book" → "Buch"

These systems often struggled with full sentences and complex expressions because they fail to capture context and phrase-level meaning. "She gave him a hand" might be translated literally, missing its idiomatic meaning.

Most modern systems, including Google Translate and DeepL, use **neural machine translation (NMT)** (Wu et al. 2016; DeepL 2021). These systems are trained on large sets of translated texts. They learn to represent the meaning of whole sentences as mathematical structures and generate more fluent and accurate translations. Unlike earlier systems, they aim to consider the full context of a sentence, which helps reduce errors and improves the handling of ambiguous or idiomatic language.

2.2.2 What is Gender Bias in Machine Translation?

A clear definition of gender bias has not yet been established (Stanczak and Augenstein 2021). Determining which features in text indicate bias is difficult, and the characteristics of non-biased text are often unclear. This makes it challenging to hold users accountable for gender bias, detect all harmful signals, and develop standard evaluation benchmarks (Barclay and Sami 2024; Shrestha and Das 2022; Stanczak and Augenstein 2021).

Since there is no clear definition, this work defines gender bias based on specific manifestations described in the following subsection. Any text that exhibits one or more of these forms will be considered gender biased.

2.2.3 Manifestiations of Gender Bias

This section draws from the main studies analyzing gender bias in EN-DE MT (Ullmann 2022; Rescigno and Monti 2023; Lardelli et al. 2024; Kappl 2025). Since existing research does not clearly define the different manifestations, the findings are grouped here into three main categories.

Defaulting to Masculine Forms

In both singular and plural contexts, the *generic masculine* refers to the default use of the masculine grammatical gender. For example, the sentence "Die Studenten sind im Hörsaal" (translation: "The students are in the lecture hall") uses the masculine plural form to refer to a group of students regardless of their 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). In MT, the generic masculine can lead to inaccurate or unfair representations of gender in translated text. 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. This shows that the bias is not minor but occurs quite heavily in those systems.

Reinforcement of Stereotypes

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). These associations can then be reinforced by popular media, such as TV and advertisements (Godsil et al. 2016), just as much as it can be influenced by 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).

A common manifestation of this are **stereotypical job associations**. This can be seen in cases where models assign he/him pronouns to roles like doctors and pilots, and she/her pronouns to roles like nurses and flight attendants (Shrestha and Das 2022), with an even stronger tendency in male-domniated fields such as STEM (Prates et al. 2019). In addition, NLP models have also been shown to **link certain adjectives and traits to genders**. Traits like "masterful," "assertive," and "competitive" are often associated with men, while "friendly," "unselfish," and "emotionally expressive" are more commonly linked to women (Godsil et al. 2016).

Neglecting Contextual Information

Coreference resolution refers to the process of using contextual information to determine the correct gender in translation (Stanczak and Augenstein 2021). In MT, this means identifying links between words like pronouns and the nouns they refer to. While human translators use both linguistic cues (such as pronouns and grammar) and real-world knowledge to correctly assign gender (Rescigno and Monti 2023), MT systems often fail to do so reliably, especially when gender information appears earlier in the text or across sentence boundaries (Cho et al. 2019; Stanovsky et al. 2019). For example, if a biography introduces a person with a female name at the beginning, but later refers to that person only by name, translation systems may lose the link and default to masculine forms for the remaining text.

Rescigno and Monti (2023) found that including previous sentences improved coreference resolution and reduced masculine defaults, though some systems benefited more than others. However, the use of context also introduced occasional new errors. Additionally, Savoldi et al. (2024) highlighted that correcting biased translations toward feminine forms required significantly more time and edits than masculine ones, revealing a notable cost disparity.

Similarly, Lardelli et al. (2024) showed that even with natural passages from Wikipedia and Europarl, systems still largely defaulted to masculine forms. Feminine and inclusive translations remained rare, while gender-neutral alternatives appeared mainly when the noun itself suggested them.

2.3 Societal Relevance and Impact of Gender Bias in Machine Translation

2.4 Research Gaps in Gender Bias Detection for English-to-German Machine Translation

2.5 Approach and Justification of the Technical Setup

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

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*.

Chakravarthi, B. R., Rani, P., Arcan, M., and McCrae, J. P. (2021). A Survey of Orthographic Information in Machine Translation. *SN Computer Science*, 2(4):330.

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

DeepL (2021). How does DeepL work? https://www.deepl.com/en/blog/how-does-deepl-work.

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).

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.

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.

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.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. *European Association for Machine Translation*, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation:479–480.

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.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Appendix

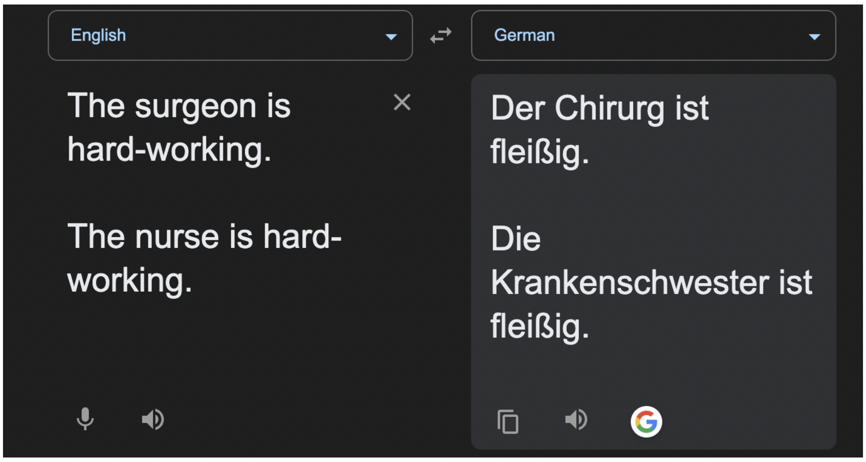

Figure 1: Google Translate assigns stereotypical genders to occupational roles.

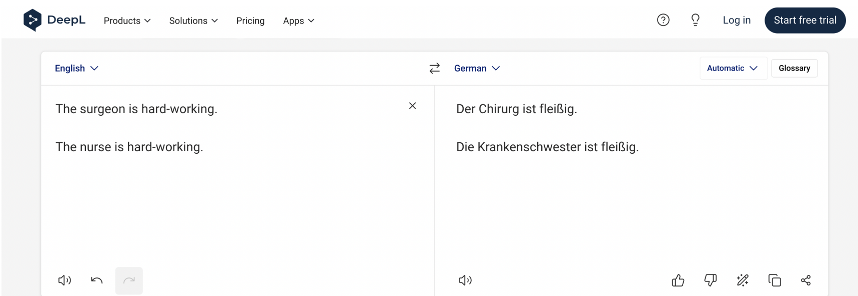

Figure 2: DeepL shows a similar bias in the same sentence, highlighting consistent patterns across MT tools.

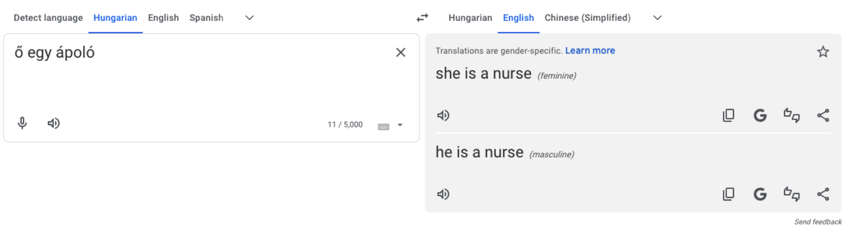

Figure 3: Gender-specific translation by Google Translate for ambiguous pronouns.

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 29, 2025

....

(Unterschrift des Verfassers)