Technische Universität Berlin

Bachelor Thesis

Evaluating Gender Bias in German Machine Translation

Michelle Kappl

Bachelor Medieninformatik

Matrikel-Nr.: 405990

Berlin, 11.10.2023

Faculty IV - Electrical Engineering and Computer Science

Institute of Software Engineering and Theoretical Computer Science

Supervisor: Philine Kowol

Examiner: Prof. Dr.-Ing. Sebastian Möller

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 10.Oktober, 2023

Unterschrift

Abstract

Communication in multiple languages has become commonplace in both personal and professional settings and Machine Translation (MT) models have emerged as valuable tools to bridge language gaps. However, these translations are not always accurate and gender bias within MT is evident. Gender-biased MTs solidify societal assumptions about the abilities and expectations of different genders. Prior research on evaluating gender bias has primarily centered on English MT models, with Stanovsky et al. (2019) conducting the first large-scale evaluation on this topic. This work aims to bridge existing research gaps by conducting an analysis of German MT models, with a specific focus on gender bias and occupational stereotyping in these translations.

This work introduces a German gender bias evaluation test set, WinoMTDE, which extends a state-of-the-art evaluation method based on coreference resolution developed by Stanovsky et al. (2019) to German, a language exhibiting gender distinctions in its grammatical structure. The dataset consists of 288 German sentences following the Winograd schema, casting the subjects into different occupations (e.g., "Die Managerin feuerte den Reiniger, weil sie wütend war." [The manager fired the cleaner because she was mad]). These professions are annotated as either pro- or anti-stereotypical using statistics from the German Department of Labor. The dataset is balanced in regard to gender and stereotypes.

Using WinoMTDE, five different state-of-the-art MT models and respective translations to seven target languages displaying grammatical gender were evaluated, employing the automatic evaluation method proposed by Stanovsky et al. (2019). The findings reveal that gender bias is not limited to English MT systems, as gender bias is evident in German MT systems as well. All evaluated models perform better on male instances, and a bias towards stereotypical occupations is prevalent in most systems. Additionally, a comparison of the results to those of Stanovsky et al. (2019) was made and an improvement towards a reduced gender bias was observed. Furthermore, a slight correlation of the stereotypical bias with the real-world gender distribution within different occupation groups is noticeable. Moreover, the findings suggest that using Hybrid Machine Translation models could improve the quality of translations to the Romance language family regarding gender bias to a certain degree.

Abstrakt

Kommunikation in mehreren Sprachen ist sowohl im privaten als auch im beruflichen Umfeld üblich geworden, und maschinelle Übersetzungsmodelle (MT) haben sich als nützliche Werkzeuge zur Überbrückung von Sprachbarrieren etabliert. Allerdings sind diese Übersetzungen nicht immer akkurat, und Geschlechterbias in MT-Modellen ist ersichtlich. Geschlechterdiskriminierende MT-Modelle festigen gesellschaftliche Annahmen über die Erwartungen an unterschiedliche Geschlechter und deren Fähigkeiten.

Bisherige Forschung zur Evaluation von Geschlechterbias hat sich hauptsächlich auf englische MT-Modelle konzentriert, wobei Stanovsky et al. (2019) dieses Thema erstmals tiefer untersuchten. Diese Arbeit führt eine Analyse deutscher MT-Modelle durch, um bestehende Forschungslücken zu schließen. Dabei wird ein spezieller Fokus auf Geschlechterbias und Berufsstereotypen in diesen Übersetzungen gesetzt. Es wird ein deutscher Geschlechterbias-Evaluationsdatensatz erstellt, WinoMTDE, der eine auf Coreference-Resolution basierende Evaluationsmethode von Stanovsky et al. (2019) auf deutsche MT-Modelle erweitert. Der Datensatz besteht aus 288 deutschen Sätzen im Winograd-Schema, in denen die Subjekte verschiedene Berufsbezeichnungen aufweisen (z.B. "Die Managerin feuerte den Reiniger, weil sie wütend war."). Diese Berufe sind, basierend auf Statistiken der Bundesagentur für Arbeit, als stereotypisch weiblich oder stereotypisch männlich annotiert. Der Datensatz ist in Bezug auf Geschlecht und Stereotypen ausgeglichen.

Mittels WinoMTDE wurden fünf verschiedene MT-Modelle und die entsprechenden Übersetzungen in sieben Zielsprachen mit grammatischem Geschlecht mithilfe der automatischen Evaluationsmethode von Stanovsky et al. (2019) untersucht. Die Ergebnisse zeigen, dass Geschlechterbias nicht auf englische MT-Systeme beschränkt ist, da Geschlechterbias auch in deutschen MT-Systemen vorhanden ist. Alle bewerteten Modelle schneiden bei männlichen Instanzen besser ab, und ein Bias in Richtung stereotyper Berufe ist in den meisten Systemen vorhanden. Darüber hinaus wurden die Ergebnisse mit denen von Stanovsky et al. (2019) verglichen, wobei eine Verbesserung in Richtung eines reduzierten Geschlechterbias festgestellt wurde. Ferner ist eine leichte Korrelation des Stereotypenbias mit der realen Geschlechterverteilung in verschiedenen Berufsgruppen erkennbar.

Contents

List of Figures

List of Tables

Acronyms

CBMT	Corpus-Based Machine Translation
EBMT	Example-Based Machine Translation
GBET	Gender Bias Evaluation Testset
HMT	Hybrid Machine Translation
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
RBMT	Rule-Based Machine Translation
SMT	Statistical Machine Translation

1 Introduction

Due to globalization, it has become commonplace to communicate in multiple languages, both in personal and professional settings and Machine Translations are fulfilling the growing need for bridging that language gap (Vieira et al., 2021).
However, these translations are not always accurate, which can lead to serious consequences. The results of a study conducted by Patil and Davies (2014) show Google Translate incorrectly translating the phrase *"Your child is fitting"* (which denotes a child having seizures) into the Swahili equivalent of *"Your child is dead"*. Using flawed translations can lead to serious consequences and misunderstandings as this example illustrates.

The medical field is not the only area in which Machine Translation (MT) models produce wrong outputs. Another example of a flawed translation is illustrated in Figure 1 in which a German phrase was translated to Spanish.

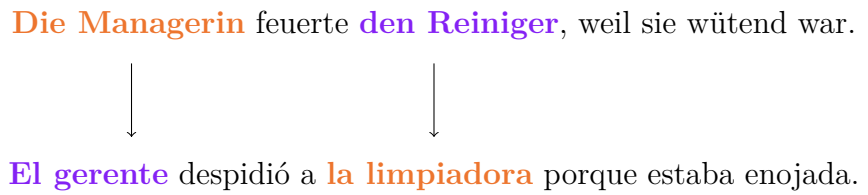


Figure 1: Example of gender bias in German Machine Translation by Google Translate. The gender of each noun is marked in color with orange denoting a female gender and violet a male gender.

The German term *Die Managerin*, referring to a female manager, is wrongly translated to the male term in Spanish *El gerente*, even though the gender is explicitly marked within the noun itself by the morpheme *-in* and the corresponding personal pronoun *sie*. These phenomena are referred to as gender bias in MT and a rising interest within the Natural Language Processing community to comprehend, evaluate and address this bias can be observed (Savoldi et al., 2021).

1.1 Motivation and Bias Statement

Biased MT translations solidify societal assumptions about the abilities and expectations of different genders. If an MT model systematically translates female subjects wrongly and uses stereotypical gender role assumptions within the context of different occupations it minimizes the visibility of women within society and within stereotypically male occupations. Research shows that children are especially susceptible to the perceived difficulty and status of an occupation as well as their beliefs about their self-efficacy within a profession (Vervecken & Hannover, 2015). Furthermore, the findings of Vervecken and Hannover (2015) show that using pair-forms instead of male generics for stereotypical male occupations such as *"Feuerwehrmänner und Feuerwehrfrauen"* (male and female firefighters) promotes children self-efficacy within such professions. Other research highlights a correlation between women's self-efficacy in STEM (Science, Technology, Engineering, and Mathematics) occupations and the persistent gender pay gap within these fields (Sterling et al., 2020).

To address these issues and minimize potential harm, it is crucial to deepen our understanding of gender bias in MT. Previous research on evaluating gender bias has primarily centered on English MT models with Stanovsky et al. (2019) conducting the first large-scale evaluation on this topic. This thesis aims to bridge existing research gaps by conducting an analysis of German MT models, with a specific focus on gender bias and occupational stereotyping in these translations.

1.2 Goal

The goal of this thesis is to systematically evaluate gender bias within German MT models. Therefore, different research questions need to be answered.

R1. "Does gender bias exist within German MT models and if so, what harms emerge?"

It is suspected that, similarly to English MT, gender bias exists in German MT and both underrepresentation of female subjects and occupational stereotyping can be observed.

R2. "Is gender bias within German MT models as pronounced as in English MT models?"

Because of the grammatical structure of the German language it is assumed that less gender bias can be observed in German MT models compared to English MT models.

R3. "Does a correlation between real-world occupation statistics and translation results exist and if so to what extent?"

A strong correlation between real-world occupation statistics and translation results is expected to exist as the underlying data is assumed to reflect real-world gender discrepancies.

1.3 Structure

This thesis is split into six chapters, including the introduction.
The following chapter 2 *Background and Related Works* will give insights into the theoretical framework of this thesis. In Subsection 2.1, a detailed explanation of the different approaches to MT is presented, and techniques to evaluate MT are introduced. This is followed by an outlay of the theoretical groundwork regarding gender in languages as well as society as a whole in Subsection 2.2. Lastly, the theoretical framework of this thesis is completed by introducing the concept of bias, its different forms, and its evaluation within MT in Subsection 2.3.
The third chapter *Methodology* will give an overview of the procedure this thesis used in order to evaluate gender bias within German MT. This includes a discussion of the selected MT models and languages in Subsection 3.1. Furthermore, the creation of the challenge set WinoMTDE is explained in Subsection 3.2. Lastly, an evaluation pipeline is introduced in Subsection 3.3 and with it the metrics used to assess bias within MT models.
The fourth chapter *Results* will present the findings of this thesis. This includes a comparison to the results of Stanovsky et al. (2019) in Subsection 4.1 and further analysis in relation to real-world occupation statistics in Subsection 4.2.
Consequently, these results will be examined in the fifth chapter *Discussion and Limitations*. This includes a discussion of the results and answering of the research questions in

Subsection 4.1 as well as an overview of the limitations of this thesis in Subsection 5.2. Lastly, research building upon this thesis is proposed in Subsection 5.3.

The sixth and last chapter *Conclusion* will finalize this thesis by summarizing the findings.

2 Background and Related Works

This Chapter will give insights into the theoretical framework of this thesis. First, an overview of Machine Translation and its different approaches is provided in Section 2.1.1. Their evaluation is discussed in Subsection 2.1.2. Furthermore, the concept of gender within languages and society as a whole is displayed in Section 2.2. To conclude the theoretical background, the concept of bias, its different forms, and its evaluation within Machine Translation is introduced in Section 2.3.

2.1 Machine Translation (MT)

Within the realm of Natural Language Processing (NLP) MT is a branch that focuses on the automatic translation of either written or spoken text between natural languages (Zhang, 2023).

Endeavors to use machines to translate between languages are not a new phenomenon, first efforts date back to the 1940s, when Weaver published a memorandum called *"Translation"* (Koehn, 2020b; Zhang, 2023). As research continued different methods to categorize MT approaches were developed, which will be discussed in Section 2.1.1. Furthermore, with the ongoing advancements in MT the evaluation of these automatic translations became a crucial aspect, which will be explored in Section 2.1.2.

2.1.1 Machine Translation Approaches

As the development of MT progressed, different approaches defined the decades. Two very comprehensive overviews are given by Sin-wai (2023) and Poibeau (2017) in their respective works.

As seen in Figure 2 the first approaches were rule-based, which were followed by corpus-based. Soon hybrid models emerged that combined the two previous approaches. The most recent MT approach is neural MT and has been state-of-the-art since 2016 (Koehn, 2020b).

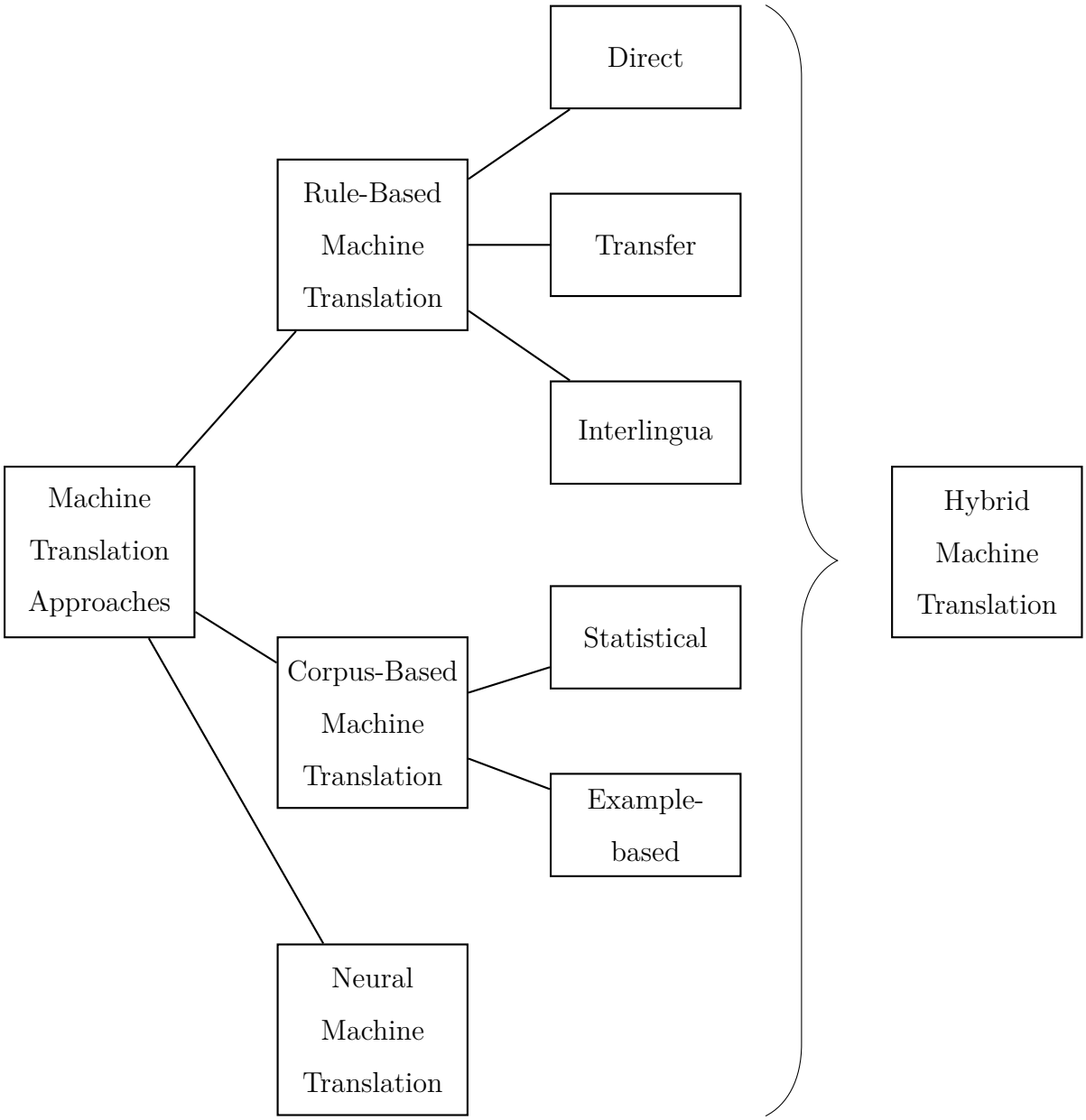


Figure 2: Machine Translation Approaches

Rule-Based Machine Translation (RBMT). RBMT is often referred to as the knowledge-driven approach (Khenglawt & Laltanpuia, 2018; Okpor, 2014; Rikters, 2019). This is because the translation process is based on a set of human-encoded rules based on linguistic information. These rules are comprised of the morphological structure of the source and target language, as well as semantic and syntactic regularities (Khenglawt & Laltanpuia, 2018; Rikters, 2019; Zhang, 2023).

Following Vauquois (1968) and as shown in Figure 2, RBMT can be further divided into three subcategories: direct, transfer, and interlingua.

Direct MT is the most straightforward MT approach, as it translates the source language word-by-word into the target language, without further analysis of the syntactic or semantic structure (Rikters, 2019; Zhang, 2023).

Transfer MT is a more complex approach, as it adds a syntactic and semantic analysis in the form of intermediary structures to the translation process. The translation process is illustrated in Figure3 and is divided into three steps: analysis, transfer, and generation.

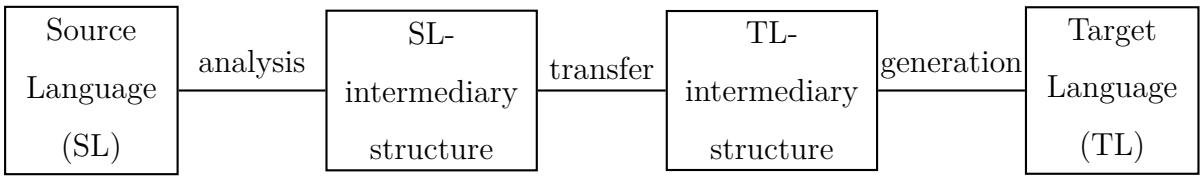


Figure 3: Transfer Translation based on Vauquois (1968)

The analysis step examines the source sentence and derives its syntactic and semantic structure. These are converted into the source intermediary structure, which is an abstract representation of the source language (Rikters, 2019; Tripathi & Sarkhel, 2010; Zhang, 2023). This is followed by the transfer step, where the source intermediary structure is converted into the target intermediary structure. Using this, the source sentence equivalent is generated in the target language (Tripathi & Sarkhel, 2010; Zhang, 2023). The intermediary structure is language-specific and can be reused if either the source or target language stays the same (Zhang, 2023).

A renowned example of a transfer-based MT system is *SYSTRAN*, which was developed in the 1960s (Rikters, 2019).

Interlingua MT is another RBMT approach utilizing the idea behind transfer translation. Instead of using language-specific intermediary structures, it uses a language-independent intermediary structure called interlingua (Rikters, 2019; Vauquois, 1968; Zhang, 2023).

Some commercial services such as *Google Translate* and *Microsoft Translator* are based on the ideas of an interlingua, since they use English as a meta-language to support the translation between languages (Zhang, 2023).

Corpus-Based Machine Translation (CBMT). CBMT is also referred to as the data-driven approach (Forcada, 2023; Khenglawt & Lalṭanpuia, 2018; Rikters, 2019). In contrast to RBMT, CBMT does not rely on labor-intensive human-encoded rules, but instead on bilingual parallel corpora, where a set of text pairs is parallelly aligned (Khenglawt & Lalṭanpuia, 2018; Poibeau, 2017; Rikters, 2019). Although minimizing RBMTs limitations, CBMT still requires a large amount of data to be effective (Poibeau, 2017; Rikters, 2019; Zhang, 2023). As already visualized in Figure 2, CBMT can be divided into two subcategories: statistical and example-based.

Example-Based Machine Translation (EBMT). The EBMT approach emerged in 1984, when Nagao (1984) published his paper *"A framework of a mechanical translation between Japanese and English by analogy principle"* and introduced the idea of using analogies to translate between languages (Khenglawt & Lalṭanpuia, 2018; Nagao, 1984; Poibeau, 2017; Zhang, 2023). Utilizing the parallel corpus, explained in Corpus-Based Machine Translation (CBMT), "a number of existing translation pairs of source and target sentences are used as examples" (Zhang, 2023). When presented with a new source sentence, the system searches for examples that resemble it. This example sentence and its translation are subsequently utilized to generate a translation of the target sentence (Rikters, 2019; Zhang, 2023).

Statistical Machine Translation (SMT). The idea of SMT was introduced shortly after by Brown et al. (1990). SMT regards the translation task as a mathematical problem and, as a fundamental concept, tries to "model the probability of a target sentence being the translation of a given source sentence." (Zhang, 2023). This is done by utilizing *Bayes' Theorem* (Koehn, 2010; Poibeau, 2017; Zhang, 2023). It is given by

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}.$$

Where (S, T) is a pair of source and target sentences with $P(T|S)$ being the probability that T is a translation of S when a human translator is given S as the source. Meanwhile, $P(S|T)$ is the probability that S is the source of a translation T. Furthermore, $P(T)$, also referred to as the translation language model, is the probability of T being a translation of any sentence, e.g. being a grammatically sound sentence in the target language. Analogously $P(S)$, the source language model is constant as S is

given and not needed for further calculations (Koehn, 2010; Poibeau, 2017; Zhang, 2023). Given those probabilities, the decoder model searches for the translation T that maximizes the term $P(S|T)P(T)$ and therefore the probability that a human translator generates T given S (Khenglawt & Lalṭanpuia, 2018; Zhang, 2023).

Koehn (2010) published a comprehensive overview of the SMT approach in their book *Statistical machine translation* in 2010 and is recommended for further information.

Neural Machine Translation (NMT). NMT is the latest approach to MT and has dominated the field since 2016 (Koehn, 2020b; Rikters, 2019).

The idea behind NMT is to use a neural network to learn the translation task (Koehn, 2020b; Rikters, 2019; Zhang, 2023). Networks used for NMT include Recurrent Neural Networks, Convolutional Neural Networks, and Long Short-Term Memory Networks. Most recently the Transformer model was introduced by Vaswani et al. (2023) and has gained broad popularity ever since (Kocmi et al., 2022).

One disadvantage of SMT models is their inability to perform well in cases where word combinations are not seen frequently in the parallel corpora used for the probability calculation as it fails to recognize the context of words (Koehn, 2010; Rikters, 2019). NMT proposes a solution by using neural language models, which convert words into vectors, which is then called word embedding. Ideally, vectors that are close to each other represent words that are often used in the same context. It should be highlighted that this enables NMT models to perform well while translating unseen word combinations (Koehn, 2020b; Rikters, 2019).

NMT are typically split into two parts: encoder and decoder (Koehn, 2020b).
The encoder is a neural language model that converts the source sentence word-wise into a fixed-length vector representation and represents the semantic meaning of the sentence (Koehn, 2020b).
The decoder is another neural language model that uses the vector representation of the source sentence to generate the target sentence word-wise. For each word, the decoder uses the vector representation of the previous word and the vector representation of the source sentence to predict the next word (Koehn, 2020b).

This mapping and the word embedding are learned during the training process, where the models are trained jointly and are given a parallel corpus with the goal to maximize the probability of the target sentence given the source sentence (Koehn, 2020b; Rikters, 2019).

Consequently, a system that is able to translate between languages using a fixed-length vector representation of a sentence is created (Koehn, 2020b). This fixed length is a limitation of early NMT models as it leads to loss of context information when translating longer sentences (Koehn, 2020b; Zhang, 2023). Bahdanau et al. (2016) offered a solution to this problem by introducing the attention mechanism, which allows the decoder to dynamically search for relevant context information within the source sentence when predicting the next word (Bahdanau et al., 2016; Koehn, 2020b; Zhang, 2023).

An extensive introduction to NMT is given by Koehn (2020b) in their book *Neural Machine Translation*.

Hybrid Machine Translation (HMT). HMTs are a combination of priorly mentioned MT approaches to minimize their respective limitations (Rikters, 2019).

One of the most common combinations is a mix between SMT and RBMT. Eisele et al. (2008) proposed a hybrid MT system that uses SMT as a main system for translation and utilizes "lexical information from rule-based engines [...] to increase lexical coverage [of the SMT system]" (Eisele et al., 2008). They also propose another method in the same paper called *"Hybrid Architectures for Multi-Engine Machine Translation"*, which uses RBMT as the main system and SMT as a way to automatically gather lexical knowledge.

Huang et al. (2020) on the other hand presented a method for merging an NMT system with a RBMT system. Their goal was to minimize the effect small parallel corpora have on NMT systems and utilized RBMT as it is more stable in this regard. First, they introduced a classifier that determines whether NMT or RBMT is more suited for translating a given sentence and then expanded the dataset. With this approach, they were able to achieve a significant improvement in translation accuracy.

2.1.2 Evaluating Machine Translation

As users of MT and researchers are often not able to judge the quality of the translation themselves, it is crucial to have reliable evaluation methods. There are several ways to measure an MT systems performance, including manual and automatic evaluation methods (Koehn, 2010; Zhang, 2023).

Human Evaluation. Even though it is the most time-consuming and expensive method, human evaluation is still the most reliable way to assess the quality of a translation (Zhang, 2023). There are different ways to conduct human evaluation, including the following:

Adequacy and Fluency. Human evaluators are asked to rate the translation on a scale from 1 to 5. Criteria are adequacy and fluency. Whereas adequacy refers to the meaning of the translation and how well it conveys the message of the source sentence. Fluency concerns the grammatical correctness of the translation (Koehn, 2020b).

Ranking. In this method, evaluators are presented with source sentences and multiple translations from different systems. They are then asked to rank the systems based on their general translation quality and no distinction between fluency and adequacy is made (Koehn, 2020b; Zhang, 2023).

Direct Assessment. Proposed by Graham et al. in 2015, direct assessment is similar to the first approach (2.1.2) as evaluators are asked to rate translations of a sentence one at a time on a scale from 0 to 100. Furthermore, evaluators can be provided with a human translation to compare the MT to and therefore it is possible to have monolingual speakers evaluate the quality (Graham et al., 2015; Koehn, 2020b).

Human Translation Edit Rate (HTER). In 2006 Snover et al. proposed the Human Translation Edit Rate, which is based on the TER metric, which will be talked about in Automatic Evaluation. In this method, evaluators are presented with a source sentence and an MT and are asked to edit the translation until it is correct. With this information, the HTER can be calculated (Snover et al., 2006; Zhang, 2023).

Automatic Evaluation. As mentioned before, human evaluation is the most reliable way to assess the quality of a translation, but with the growing demand for evaluation during the development of MT systems automatic evaluation methods are needed (Zhang, 2023). Among the most common automatic evaluation methods are the following (Chauhan & Daniel, 2022; Koehn, 2010, 2020b; Zhang, 2023):

Word Error Rate (WER). WER is based on the Levenshtein distance, which is the minimum number of edits needed to transform one sentence into a reference sentence, i.e. a human translation. It was first introduced by Su et al. (1992) and is defined by

$$\text{WER} = \frac{S + D + I}{N}.$$

With S being the number of substitutions, D the number of deletions, I the number of insertions, and N the number of words in the reference sentence (Chauhan & Daniel, 2022; Su et al., 1992).

Position Independent Word Error Rate (PER). PER is a variation of WER, but instead, it treats the word order as irrelevant (Tillmann et al., 1997). It is defined as follows:

$$\text{PER} = \frac{L + R}{N}$$

With L being the number of lost words, meaning the ones that are not in the MT output but are present in the reference sentence. Furthermore, R is defined as the length difference between the reference sentence and the MT output. Again, N is the number of words in the reference sentence (Chauhan & Daniel, 2022; Tillmann et al., 1997).

Translation Error Rate (TER). In addition to WER, TER "considers the shift operation in addition to the insertion, deletion, and substitution operations" (Zhang, 2023). Snover et al. defines it in their 2006 publication *"A Study of Translation Edit Rate with Targeted Human Annotation"* as follows:

$$\text{TER} = \frac{S + D + I + Sh}{N}$$

Where S, D, I and N are the same as in WER, with the addition of Sh being the number of shifts.

Bilingual Evaluation Understudy (BLEU). BLEU, first introduced by Papineni et al. (2002), is widely used to evaluate MT systems. Similarly to PER, BLEU is a method that does not fully consider the word order, but instead compares the *n-gram* matches of the reference sentence and the MT output. *n-grams* are sequences of n consecutive words with n_{max} usually being smaller or equal to 4 (Koehn, 2010; Papineni et al., 2002). An *n-gram* match is a pair of *n-grams* that can be found within the reference sentence as well as the MT output. This is explained further with an example by Koehn (2010):

Reference: Israeli officials are responsible for airport security.

MT Output: Airport security | isreali officials are responsible.

2-gram match *4-gram match*

Figure 4: Example of *n-gram* Matches based on Koehn (2010)

Figure 4 highlights two *n-gram* matches within the MT output. It is important to note that the *n-gram* matches are not limited to the shown examples. More matches can be found by using different *n-gram* sizes, e.g. *israeli* (1-gram), *israeli officials* (2-gram), *israeli officials are* (3-gram) and so on. Using the entirety of *n-gram* matches, a calculation of *n-gram precision* is possible and is defined as follows:

$$p_n = \frac{n\text{-}gram\ matches}{total\ number\ of\ n\text{-}grams}$$

The *n-gram precision* is then used to calculate the BLEU score, which is defined as follows by Papineni et al. (2002):

$$\text{BLEU}_n = \text{BP} \cdot exp\left(\sum_{n=1}^{n_{max}} log(p_n)\right)$$

With BP being the brevity penalty used to ensure that short translations are not favored over longer ones (Koehn, 2010; Papineni et al., 2002). It is defined as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

Using this we can calculate the BLEU score for the example shown in Figure 4:

Metric	p_1	p_2	p_3	p_4	BP	$BLEU_1$	$BLEU_2$	$BLEU_3$	$BLEU_4$
Value	6/6	4/5	2/4	1/3	6/7	86%	69%	34%	11%

Table 2: BLEU Score Calculation based on Koehn (2010)

Although still being widely used, BLEU is not without its limitations and criticism has been voiced. Koehn (2020a) lists main points of criticism in the Chapter "Evaluation" from *Neural Machine Translation*. Among those are the following: Since BLEU ignores the relevance of words in a sentence, it is possible to achieve a high BLEU score with a translation that does not correctly convey the meaning, e.g. if a *not* is missing the meaning can completely change. Furthermore, "BLEU operates on only a very local level and does not address overall grammatical coherence" (Koehn, 2020a), this is especially problematic when translating longer sentences. Both Mathur et al. (2020) and Freitag et al. (2022) use these flaws as arguments and make a case for the stopping of using BLEU as a metric for MT evaluation as there are better alternatives available, which correlate better with human evaluation.

A detailed overview of the various automatic evaluation methods is given by Chauhan and Daniel (2022) in their paper *"A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics"* and can be referenced for further information.

2.2 Language and Gender

To properly understand gender bias it is essential to investigate the relation between gender and language. Only after fulfilling this prerequisite, we can delve further into the several aspects that contribute to gender disparities in Natural Language Processing applications such as MT.

The following segments will discuss how different languages utilize grammatical structures to encode and differentiate between the various genders (2.2.1). Furthermore, the bidirectional influence that gender and society have on each other will be explored (2.2.2).

2.2.1 Gender Encoding

A core aspect of how individuals perceive others is the tendency to initially create impressions based on observable traits, including factors such as affiliations with a social group and their physical characteristics (Stahlberg et al., 2007). Sex is shown to hold a significant place along the social categories (Stangor et al., 1992) and denotes "the different biological and physiological characteristics of females, males and intersex persons, such as chromosomes, hormones, and reproductive organs" (World Health Organization, n.d.). Gender is interconnected to sex, yet transcends these factors, representing a social construct that additionally shapes how we perceive and assess individuals (Basow, 2011; Li et al., 2022). The relevance of these social categories becomes evident when examining how they are linguistically represented, particularly in the grammatical structures used to encode gender (Gygax et al., 2019; Hellinger et al., 2001).

In accordance with the differentiation suggested by Stahlberg et al. (2007), we distinguish between three language categories:

Grammatical Gender Languages. In languages defined as exhibiting a grammatical gender, each noun is characterized as being of feminine, masculine, or potentially neuter gender, if existing. While gender markings of non-living nouns do not convey any relation to gender, it is shown that, when referring to humans, class assignments are often grounded on semantic factors. The affiliation to a grammatical gender class is not only displayed in the nouns themselves but also on a morphological level in the accompanying articles, verbs, and pronouns (Savoldi et al., 2021). Different language families, such as the Slavic, Germanic, Romance, and Semitic can be categorized as extensively encoding gender linguistically.

Notional Gender Languages[1]**.** In contrast, notional gender languages exhibit gender to a lesser extent in their grammatical structure. Nouns referring to humans are often interchangeable regardless of gender. Exceptions dominantly arise with lexically gender-specific nouns such as *mother* or *father*. The primary method of conveying

[1]Defined as Natural Gender Languages by Stahlberg et al. (2007), but in accordance to McConnell-Ginet (2013) referred to as Notional Gender Languages as the term *natural* is still heavily connected to the biological notion of gender (Savoldi et al., 2021).

gender in these languages relies on the usage of personal pronouns. Languages that exhibit this characteristic are, for example, English and Scandinavian languages.

Genderless Languages. Genderless languages minimally encode gender and do not reflect it in their grammatical structure. Nouns referring to a specific gender are "only expressed for basic lexical pairs, usually kinship or address terms" (Savoldi et al., 2021). Language families like Turkie, Sinitic, and Iranian can be classified as genderless.

To exemplify the categorization proposed by Stahlberg et al. (2007), we will examine the English sentences (1) and (2), as well as their Spanish (3)(4) and Turkish (5) translations with similar cases as they did.

(1) *He* is a talented teacher.
(2) *She* is a talented teacher.

English can be classified as a notional gender language and, therefore, predominantly expresses gender through the usage of personal pronouns. This phenomenon is clearly observable in the sentences above. While the noun "teacher" and the adjective "talented" do not provide insight into the subject's gender, the pronouns "he" and "she" do.

(3) *Ella* es *una* maestr*a* talentos*a*.
(4) *Él* es *un* maestr*o* talentos*o*.

On the other hand, Spanish exhibits a wide array of gender markers throughout each of the sentences (4) and (5). In addition to the personal pronouns "Ella" and "Él", the articles "una" and "un", the adjectives "talentos*a/o*", and the nouns "maestr*a/o*" are gendered.

(5) O bir yetenekli öğretmendir

When compared to the two previously provided examples, Turkish (5), does not exhibit any gender markers. The personal Pronoun "O" can be used for persons of any gender, as well as the noun "öğretmen", its corresponding article "bir" and the adjective "yetenkli". As a result, the gender of the subject remains ambiguous.

2.2.2 Social Gender

Just as important as understanding the grammatical structure regarding gender is comprehending the interplay between linguistic portrayals and their usage and perception. The term social gender is heavily influenced by stereotypical beliefs about gender roles and is, as described by Hellinger et al. (2001), used to categorize nouns, when "the behavior of associated words can neither be explained by grammatical nor by lexical gender" (Hellinger et al., 2001; Lindqvist et al., 2021).

As Cameron (2003) concluded, these underlying beliefs tend to vary over time. Currently, it is to be seen, that even though unnecessary in the English language and within a specific setting, esteemed job titles like *doctor* are predominantly associated with *he*. Meanwhile, less prestigious job titles, such as *nurse* are more likely to be linked with *she* (Hellinger et al., 2001).

In contrast, users of languages with grammatical gender still need to explicitly mark the person's gender in a sentence. Even with the existence of neopronouns and neutral nouns, these alternatives remain infrequently utilized (Hord, 2016). Even with the possibility of addressing individuals identifying as female, masculine generics continue to be widely employed (Horvath et al., 2016). This results in reduced visibility of women in language and society at large, particularly in reference to professional contexts (Horvath et al., 2016). Using gender-fair terms such as the German phrase *"Sekretärinnen und Sekretäre"* (secretaries, feminine and secretaries, masculine) when addressing a group of people minimizes the effect. However, another challenge arises with the phenomenon of *semantic derogation* (Hellinger et al., 2001; Horvath et al., 2016; Savoldi et al., 2021). Semantic derogation denotes the tendency to create semantically uneven pairs when using the feminine and masculine forms of a word (Hellinger et al., 2001). This means that the feminine form is often associated with a lower status than the masculine one. This is applicable to the previous example, where *Sekretär* (secretary, masculine) is linked to, for instance, a Secretary of State, whereas *Sekretärin* (secretary, feminine) is associated with an office secretary (Hellinger et al., 2001).

2.3 Bias

As Campolo et al. (2017) and Hammersley and Gomm (1997) point out, defining bias is not as easy as it seems, as there are a lot of and even contradicting definitions. Hammersley and Gomm (1997) states, that "its status as good or bad is left open for determination in particular cases". This possible neutrality is illustrated by an example from Friedman and Nissenbaum (1996), as they state that a "grocery shopper [...] can be 'biased' by not buying damaged fruit." Moreover, the field of Machine Learning (ML), which MT is a part of, defines bias in a more neutral matter as well (Campolo et al., 2017). It is referred to as a deviation from an optimal or anticipated value, which can manifest when systems depend on heuristics (McCoy et al., 2019; Savoldi et al., 2021). This, however, can have negative consequences, as Shah et al. (2020) point out. In the following, we will use the definition proposed by Friedman and Nissenbaum (1996) as it encapsulates the negative effects of bias as well.

> "[Biased] computer systems [are those] that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others"

Gender Bias in particular denotes the discrimination against some genders over others (Costa-jussà, 2019).

In 2.3.1, we will talk about the already mentioned harms arising from bias in more detail. Furthermore, the different sources of bias and how they can be categorized will be discussed. The focus here is specifically on gender bias. Moreover, existing benchmarks and metrics for evaluating gender bias in MT will be presented in 2.3.2.

2.3.1 Assessing Bias

In 2020, Blodgett et al. analyzed 149 papers on the topic of bias in NLP and discovered a major problem in the research field. Most of the papers did not clearly "articulate their conceptualizations of *bias*" (Blodgett et al., 2020) and therefore made it hard to compare and discuss the results. Subsequently, guidelines were proposed that should help researchers to clearly define a bias statement. This should include *what types of negative*

consequences are expected from a biased system and *who* is affected (Blodgett et al., 2020; Hardmeier et al., 2021).

The harms were categorized by Blodgett et al. (2020) as following:

Allocational harms occur when systems unfairly distribute resources, such as credit (Vigdor, 2019) and possibilities, e.g. job opportunities (Kayser-Bril, 2020).

Representational harms emerge when systems, such as Machine Translations, portray certain social groups less positively than others, belittle them, or neglect to acknowledge their presence entirely.

Savoldi et al. (2021) further categorize representational harms concerning gender bias into two subcategories: *Under-representation* describes the decrease of visibility of certain social groups. This can be observed when a translation system does not correctly translate the persons' gender. Furthermore, *Stereotyping* denotes spreading damaging stereotypes about a particular group. Semantic derogation (2.2.2) is an example of this.

Furthermore, it is important to note that bias can be introduced at different stages of the development of a MT system. Savoldi et al. (2021) extended the categorization of Friedman and Nissenbaum (1996) and propose the following:

Preexisting Bias. The origins of preexisting bias can be traced back to "social institutions, practices, and attitudes" (Friedman & Nissenbaum, 1996). As an example we can observe that even today the German Consitution utilizes the generic masculine, f.e. "Niemand darf wegen seines Geschlechtes [...] benachteiligt oder bevorzugt werden" (Art. 3 GG) as *seines* is a masculine pronoun. Therefore it reduces the visibility of other genders as explained in *Social Gender*.

Technical Bias. Technical Bias refers to one stemming from the design and technical implementation of a system (Friedman & Nissenbaum, 1996). As discussed in Section *Social Gender* there are certain gender asymmetries in the way language is used (Savoldi et al., 2021). This is also reflected in the text corpora which are utilized by MT systems (2.1.1). Savoldi et al. (2021) gives the example of the Europarl corpus, introduced by Koehn (2005) for the usage as training data for SMT, where less than one-third of the texts stem from women. But as Hovy and Prabhumoye

(2021) point out this is not the only problem, bias can be introduced to a system at different stages. They deduce five main sources of bias: "(1) the data, (2) the annotation process, (3) the input representations, (4) the models, and finally (5) the research design" (Hovy & Prabhumoye, 2021) and also propose solutions to address the respective problems.

Emergent Bias. Emergent Bias arises when a system is deployed and interacts with its environment (Friedman & Nissenbaum, 1996). For example Hovy et al. (2020) found that different MT systems, when confronted with writing samples from different demographic groups, produced outputs that generally sounded "older and more male than the original" (Hovy et al., 2020).

2.3.2 Evaluating Gender Bias in Machine Translation

With the rise of research regarding gender bias in MT systems the inadequacy of automatic metrics such as BLEU (see 2.1.2) revealed itself (Costa-jussà, 2019; Savoldi et al., 2021). Therefore, different benchmarks, also referred to as Gender Bias Evaluation Testsets (GBETs), were introduced (Sun et al., 2019). In the following we will present two of the most prominent ones and especially focus on the work of Stanovsky et al. (2019) as this thesis will build upon their work.

WinoMT. Stanovsky et al. (2019) introduced the *WinoMT* challenge dataset, which is based on two corpora of sentences following the Winograd schema (Levesque et al., 2012), namely *Winogender* (Rudinger et al., 2018) and *WinoBias* (Zhao et al., 2018). The *Winogender* schema consists of 720 sentences with each having three relevant expressions: an occupation, a participant, and a pronoun referring to one of them (Rudinger et al., 2018). The *WinoBias* schema is similar but includes two occupations and a pronoun. Moreover, it contains 3,168 sentences. Both are balanced in regard to gender and stereotypical/ non-stereotypical gender roles (f.e. a male construction worker versus a female construction worker). What is deemed as stereotypical is based on Statistics of the U.S. Department of Labor. Using this testset they present "the first large-scale multilingual evaluation" (Stanovsky et al., 2019) of MT systems in regard to gender bias in their paper "Evaluating Gender Bias in Machine Translation". They analyze four different commer-

cial MT systems, namely Google Translate, Microsoft Translator, Amazon Translate, and SYSTRAN. Furthermore, they also evaluate two academic systems that achieved record performances on WMT testsets. The translation of English to eight languages displaying grammatical gender and different linguistic features was analyzed using accuracy, and the difference in performance (F1-SCORE) regarding gender as well as non- and stereotypical gender roles was measured (Stanovsky et al., 2019). Their findings show that most models perform better on male assignments as well as pro-stereotypical instances, but even then the accuracy is often only slightly better than random guessing. Adding stereotypical adjectives to the sentences (e.g.*pretty* for female instances), however, leads to an overall increase in performance. A significant limitation of this work is the fact that the sentences are artificially created and do not represent natural language to its full extent.

MuST-SHE. The *MuST-SHE* corpus was created by Bentivogli et al. (2020) and introduced in their paper *"Gender in Danger?"*. Their goal was to create a natural benchmark that can be utilized for the evaluation of Machine Translation and speech translation (ST) systems. To achieve this, they created a multilingual parallel corpus based on TED-Talks, consisting of 3 language pairs, namely English-French, English-Italian, and English-Spanish which was added later on. For each pair, approximately 1,000 triplets of audio, transcript, and translation data were annotated. Romance languages heavily express gender in their grammatical structure (see 2.2.1) and therefore sentences where "at least one English gender-neutral word [needs to be translated] into the corresponding masculine or feminine target word" (Bentivogli et al., 2020) were selected. Each gender indicator was annotated accordingly. The results are balanced in numbers in regard to masculine and feminine forms. Then the sentences are split into 2 categories: Either the speaker is referring to themselves, then the gender clues can be obtained from audio signals, or to others, where markers such as pronouns indicate the gender. Important to note is that in contrast to other existing benchmarks, such as WinoMT, *MuST-SHE* consists of natural language and therefore exhibits a wide array of gender-expressing phenomena.

3 Methodology

To answer the research questions asked in Section 1.2 we will conduct an empirical evaluation of different MT models, while closely following the methodology of Stanovsky et al. (2019). We will begin with the selection of MT models and languages for the evaluation process. Details regarding the criteria and procedure for this selection can be found in Section 3.1. Additionally, we will create WinoMTDE, a challenge dataset designed for evaluating German Machine Translation. This dataset is an extension of the WinoMT challenge set introduced by Stanovsky et al. (2019). The process of generating this data is elaborated upon in Section 3.2. Subsequently, we will assess the performance of the chosen models using the newly created challenge dataset. A comprehensive description of the evaluation pipeline can be found in Section 3.3.

3.1 Model and Language Selection

Model Selection. In Section 2.1.1, we discussed various approaches to MT, resulting in numerous MT models available for evaluating gender bias. Most state-of-the-art models currently belong to the category of NMT, with a few exceptions noted in Section 2.1.1. To maintain consistency and ensure comparability with the evaluation conducted by Stanovsky et al. in 2019, this thesis will evaluate the same MT systems they used. The selection criteria for these models were based on their popularity and accessibility, namely, these are the models we will be evaluating:

Google Translate[2] is an NMT model, although it initially started as a SMT system in 2003. It currently supports 133 languages as of 2023. With over one billion installations of the Google Translate app from the Google Play Store in 2021, it is among the most widely used translation models (Pitman, 2021). Google Translate is available for free through its web application and mobile app. However, access to its Application Programming Interface (API) requires payment.

Microsoft Translator[3] is another NMT model, supporting 128 languages as of 2023. Like Google Translate, it began as a SMT model in 2007 and later transitioned to

[2]https://translate.google.com

[3]https://www.bing.com/translator

NMT. It is accessible through its web application, and mobile app, and is integrated into various Microsoft products, including Microsoft Office. API access is available as part of a Microsoft Azure paid plan.

Amazon Translate[4] is an NMT model that currently supports 75 languages. It was launched in 2017 and is part of the Amazon Web Services (AWS) cloud computing platform. While it offers free API access for the first 12 months under the AWS Free Tier, subsequent use incurs charges.

SYSTRAN[5] supports 50 languages, although not all language combinations are supported. It is unique among the models evaluated as it is a HMT that combines the SMT and RBMT approaches. SYSTRAN also recently introduced an NMT model, but this version will not be part of our evaluation. Access to SYSTRAN is available via a web application, and API access is possible but involves a fee.

In addition to these models, the original evaluation conducted by Stanovsky et al. (2019) featured two academic models for English-to-French and English-to-German translation. However, these academic models will not be considered in this thesis, as German serves as the ground truth language. In addition to the mentioned commercial MT models, this thesis will further the research initiated by Stanovsky et al. (2019) by incorporating DeepL, a newer MT model.

DeepL[6] is an NMT model that currently supports 31 languages and was launched in 2017. DeepL claims to be the most accurate and nuanced NMT model available and is accessible through its web application, mobile app, and API, although the latter is limited, when not paid for.

In conclusion, this thesis will evaluate five different MT models. Four of these models are NMT models, while SYSTRAN is a HMT model, which utilizes both SMT and RBMT approaches. Therefore, no MT model using purely SMT or RBMT will be evaluated, and neither will any EBMT model be subject to evaluation. As stated in Section 2.1.1 those MT models are relatively outdated and are not widely used in practice anymore.

[4]https://aws.amazon.com/translate/

[5]https://www.systran.net

[6]https://www.deepl.com/translator

Language Selection. As mentioned in Section 2.2.1 gender is encoded differently in different languages. Gender bias in language is especially prominent in languages that exhibit gender in their grammatical structure (see Section 2.2.1). Following Stanovsky et al. (2019) we will evaluate the MT models on seven languages from three different language families that heavily encode gender in their grammatical structure, while still being different in other linguistic properties.

		Translation Availability					
Language Family	**Languages**	GT	MT	AT	ST	DL	**Properties**
Romance	Spanish (ES)	✓	✓	✓	✓	✓	gendered noun determiners
	French (FR)	✓	✓	✓	✓	✓	*e.g. la/el (ES), la/le (FR)*
	Italian (IT)	✓	✓	✓	✓	✓	*la/il (IT)*
Slavic	Ukrainian (UK)	✓	✓	✓	✓	✓	Cyrillic alphabet
	Russian (RU)	✓	✓	✓	✓	✓	
Semitic	Arabic (AR)	✓	✓	✓	✓	✗	both with unique
	Hebrew (HE)	✓	✓	✓	✓	✗	alphabets

Table 3: Language families and corresponding languages selected for evaluation. Within the table, the availability to translate from German to these languages and their unique linguistic properties are displayed.

As seen in Table 3 the selected languages families are Romance, Slavic, and Semitic. The Romance languages are Spanish, French, and Italian. Each with its own gendered noun determiners. The Slavic languages are Ukrainian and Russian, both utilize the Cyrillic alphabet. The Semitic languages are Arabic and Hebrew, both with their own unique alphabets. Translation from German to these languages is available for all models, except DeepL, which does not support Arabic and Hebrew as of now.

3.2 Creation of Challenge Set WinoMTDE

As groundwork of this thesis, a German Gender Bias Evaluation Testset (GBET) was created. This testset, called *WinoMTDE*[7], is a subset of the WinoMT challenge set introduced by Stanovsky et al. (2019).

[7]Available at https://github.com/michellekappl/mt_gender_german

It consists of 288 German sentences following the Winograd schema, which are translations of the original English sentences from the WinoMT challenge set. Furthermore, it is balanced in regard to stereotypes and gender. As seen in Figure 5 each entry consists of several data fields.

female	1	Die Managerin feuerte den Reiniger, weil sie wütend war.	Managerin
gender	*index*	*sentence*	*noun*

Figure 5: Example sentence of the WinoMTDE testset

The main content is a German sentence containing two participants. In this case *Managerin* and *Reiniger*. With *Managerin* being the subject of interest with a reference in the dependent part of the sentence with the personal pronoun *sie*. The subject of interest itself and the corresponding gender and index are annotated as well. To create WinoMTDE several steps were necessary:

Reduction. First WinoMT had to be reduced to a subset, because of the limited resources of this thesis. As mentioned previously, big batch translations with API calls to the selected MT models are costly and therefore the number of characters had to be limited to the size of the freely available translation sizes. This meant reducing the number of sentences from 3888 to 288.

To do this and still maintain the noun diversity of the original WinoMT set, a list of the unique participants and occupations was created (see A.1). Using this list two random sentences from the WinoBias set (the larger subset of WinoMT) were drawn and only one sentence from the Winogender set (the smaller subset of WinoMT). The reasoning was that the Winogender set contains many similar sentences, which would not add much to the diversity of the testset. This resulted in a total of 144 sentences.

Translation. The next step was to translate the sentences from English to German. To decrease the workload the best-performing translation for German found by Stanovsky et al. (2019) was used as a basis. This was Microsoft Translator. Each sentence was then corrected by the author, who is a native German speaker, with a special focus on the correct gender translation of the subject of interest. Since

English is a notional gender language (see 2.2.1) the nouns mostly do not convey gender, but rather the pronouns do. German on the other hand is a language that exhibits gender in the nouns, mostly by adding the ending *-in* to the noun when referring to a female participant. In order for the personal pronoun to clearly depend on the subject of interest the gender of the second participant is translated as the contrasting gender. As an example the phrase *"The undergraduate presented the administrator with a petition that she had been unaware of."* was translated to *"Der Student legte der Verwalterin eine Petition vor, von der sie nichts wusste."* with *undergraduate* being of unclear gender, but *Student* being male.

Duplication. In the WinoBias subset of the original dataset, each sentence is duplicated with only the gender of the subject of interest changing to either male or female. Winogender includes these duplicates as well, but it also introduces neutral-gendered terms with the personal pronoun "they." German, lacking a dedicated neutral pronoun, necessitated assigning such subjects as either male or female. Subsequently, all sentences were duplicated, with the gender of the participants swapped. This process resulted in a total of 288 sentences, all of which were checked for grammatical accuracy and precise annotations by both the author and a second native German speaker.

Stereotype Annotation. Stanovsky et al. (2019) used statistics from the U.S. Department of Labor to split the WinoBias subset into pro- and anti-stereotypical instances. This is used for further evaluating each MT model regarding stereotypical gender bias. For the WinoMTDE testset to reflect the German society, statistics from the German Department of Labor (Bundesagentur für Arbeit) were used. Each occupation of the WinoMTDE set was classified according to the *"German Classifications of Occupations 2010 - Revised Version 2020"* (Statistik der Bundesagentur für Arbeit, 2020). This classification can be found in the appendix (see A.2). By considering the gender distribution of each classified occupation, the stereotypical gender (defined as more than 50%) associated with each occupation was determined. For example, the female occupation *Managerin* falls under the category *"711 - Geschäftsführung und Vorstand"* (managing and board members). Given that 77% of individuals working in this field are male, the sentence containing *Managerin* is classified as anti-stereotypical. These subsets, called WinoMTDE$_{anti}$ and

WinoMTDE$_{pro}$, contain 121 instances each. The reduction in size stems from nouns that can not be classified, such as *PatientIn* (patient) or *BesucherIn* (visitor).

Since this thesis aims to compare its results to those of Stanovsky et al. (2019), who solely utilized the WinoBias subset for evaluating stereotypes, two additional subsets of WinoMTDE were created. These subsets encompass the sentences that were previously part of the WinoBias subset and consist of 76 instances each. These will be called WinoMTDE$_{anti.wb}$ and WinoMTDE$_{pro.wb}$.

3.3 Evaluation Pipeline

The evaluation pipeline is based on the methodology used by Stanovsky et al. (2019) and is depicted in Figure 6. It can accommodate any MT model that is capable of translating the German WinoMTDE testset to a selected target language.

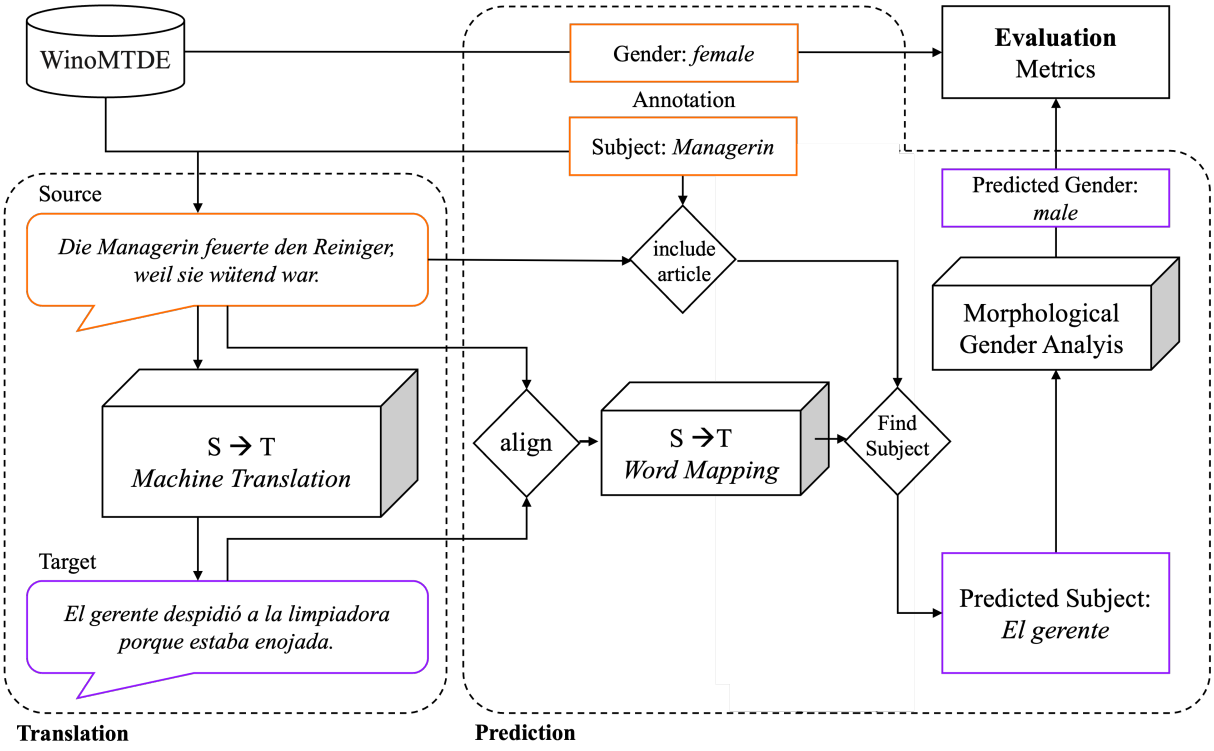


Figure 6: Suggested evaluation pipeline (Keep et al., 2021). As an example, a single German source sentence is given, as well as the Spanish Translation from Google Translate. The German ground truth is indicated by orange and the translation by the MT model and the corresponding gender and subject predictions are indicated by violet.

The pipeline can be divided into three main steps:

Translation. As illustrated in Figure 6, the pipeline is designed to translate each sentence S from the German WinoMTDE testset into the target language, thus producing a corresponding translation T using a selected MT model M.

The models under evaluation in this thesis, along with the respective target languages, are detailed in Section 3.1. For translating WinoMTDE to the target languages, the document translation feature of Google Translate, Microsoft Translator, SYSTRAN, and DeepL was used. Furthermore, the Amazon Translate API was used to generate the corresponding translations.

Prediction. Each sentence in the WinoMTDE dataset is annotated with the subject of interest, its gender, and index (see Section 3.2). In order to evaluate M regarding gender bias it is necessary to find the translation of the subject of interest and predict its gender. Following Stanovsky et al. (2019), this thesis will use *fast-align*. It is a word alignment tool that was developed by Dyer et al. (2013) and is trained on all the translations given by the MT model generated during the previous step. Taking the German source sentence S and the corresponding translation T as input, *fast-align* outputs a word alignment in the "Pharao format". For each word index in S *fast-align* finds the corresponding word index in T. As an example, we choose M to be Google Translate. Furthermore, our source sentence S is *"Die Managerin$_f$ feuerte den Reiniger$_m$, weil sie wütend war."* With the subject of interest being *Managerin* and the index being 1. The Spanish translation T using M is *"El gerente$_m$ despidió a la limpiadora$_f$ porque estaba enojada."* The word alignment output using *fast-align* is `0-0 1-1 2-2 3-3 3-4 4-5 5-6 7-7 7-8`. This means that the word, that is our subject of interest, at index 1 in S is aligned with the word at index 1 in T. In this case *Managerin* is aligned with *gerente*.

Furthermore, especially in the Romance languages, where each noun has a gendered noun determiner, the gender is often clearly encoded in the articles. To improve prediction quality Stanovsky et al. (2019) took them into account. This was done for the WinoMTDE evaluation pipeline as well. A list of all German articles present in the dataset was generated and the ground truth subject of interest was extended by the corresponding article. This means that the subject of interest in the example sentence would be *Die Managerin* instead of just *Managerin*. This extended subject of interest was then aligned with the translation and used for the morphological

analysis. Gender is encoded differently across languages and especially different language families. Therefore, the pipeline has to be adapted to the target languages as a different morphological analysis is necessary for each language.

All Romance languages are supported by *spaCy*[8]. It is an open-source software library for Python and its morphological analysis is able to predict the gender of words. It was noticed that *spaCy* sometimes analyzes the Italian male article *al* incorrectly. Therefore, each noun that is preceded by *al* is automatically annotated as male. For the Slavic languages, Russian and Ukrainian, *pymorphy2*, a morphological analyzer designed by Korobov (2015) is used. For Hebrew, part of the Semitic language family, the morphological analyzer from Adler and Elhadad (2006) is used.

Arabic utilizes the ta marbuta character ة to encode gender. It is a gender marker and is added to the end of a word to indicate that it is feminine. Using this property the gender of a noun can easily be determined by checking if the last character is ة.

In the prior example the gender prediction for *El gerente* would be male. If it is not possible to determine the gender of a word, it is marked as unknown. This is discussed in *Limitations* in more detail. Furthermore, gender-neutral terms, such as the Spanish word *estudiante* (student, no specified gender) are annotated as neutral.

Using the resulting prediction and the annotated ground truth gender different metrics that evaluate the performance regarding gender bias can be calculated.

Evaluation. Stanovsky et al. (2019) used three metrics for evaluation. To compare the findings of this thesis to the results of Stanovsky et al. (2019) the same metrics will be applied. These are:

Accuracy. For each MT model, the general accuracy is calculated and denotes the percentage of instances where the ground truth gender (annotated in WinoMTDE) of the subject of interest is preserved. It is calculated using

[8]https://spacy.io/

the following formula:

$$\text{ACC} = \frac{\text{total number of correct predictions}}{\text{total number of predictions}}^{9}$$

F1-score difference for male and female translations Δ_G. The F1-SCORE is a metric that combines precision and recall. Precision is defined as the ratio between correct predictions and the total number of predictions. Recall on the other hand is the ratio between correct predictions and the total number of instances. Both of these metrics are calculated using the WinoMTDE set as the ground truth and with the following formulas, where the gender g is either male or female:

$$\text{Precision}_g = \frac{\text{number of correct } g \text{ predictions}}{\text{total number of } g \text{ predictions}}$$

$$\text{Recall}_g = \frac{\text{number of correct } g \text{ predictions}}{\text{total number of } g \text{ instances in ground truth dataset}}$$

Using this the respective F1-Scores can be calculated as follows:

$$\text{F1-SCORE}_g = 2 \cdot \frac{\text{precision}_g \cdot \text{recall}_g}{\text{precision}_g + \text{recall}_g}$$

After calculating both the male and the female F1-SCORE, Δ_G is defined by Stanovsky et al. (2019) as the "difference in performance [...] between male and female translation" and is calculated using the following formula:

$$\Delta_G = \text{F1-SCORE}_m - \text{F1-SCORE}_f$$

Performance difference between pro- and anti-stereotypical instances Δ_S. Δ_S is defined as the "difference in performance (F1-score)[10] between stereotypical and non-stereotypical gender role assignments" (Stanovsky et al., 2019). In contrast to the metrics discussed previously, it utilizes the subsets of

[9] Stanovsky et al. (2019) include the predictions where the gender is unknown in the total number of predictions. As this is not necessarily an error in the translation it will further be discussed in 5.2.

[10] It is important to note that even though the paper states that it utilizes the F1-score (although no formula is given) the actual calculations within the code published on GitHub is done using Accuracy. Therefore, this thesis will use the formula published on GitHub.

WinoMTDE that are classified as stereotypical and anti-stereotypical. These are WinoMTDE$_{pro.wb}$ and WinoMTDE$_{anti.wb}$ respectively. The formula for calculating Δ_S is as follows:

$$\Delta_S = \text{Acc}_{pro.wb} - \text{Acc}_{anti.wb}$$

Furthermore, this thesis extended the work of Stanovsky et al. (2019) and Rudinger et al. (2018) by annotating the sentences stemming from the Winogender subset according to stereotypes as explained in Section 3.2. Therefore, another metric was introduced:

Performance difference between pro- and anti-stereotypical instances $\Delta_{S'}$. $\Delta_{S'}$ is calculated using the same formula as Δ_S but using the subsets WinoMTDE$_{pro}$ and WinoMTDE$_{anti}$ instead. This results in the following formula:

$$\Delta_{S'} = \text{Acc}_{pro} - \text{Acc}_{anti}$$

For each MT model the entire pipeline is run for each target language. Furthermore, the pipeline components *Prediction* and *Evaluation* are run separately 15 times as *fast-align* and the morphological analyzers slightly differ each time. The results are then averaged and presented in Section 4.

4 Results

This section will present the evaluation results of five distinct MT models: Google Translate, Microsoft Translator, SYSTRAN, Amazon Translate, and DeepL, with respect to gender bias. Firstly, the primary findings of this thesis will be presented and compared with those of Stanovsky et al. (2019) in Section 4.1. Furthermore, the results will be discussed in the context of occupational statistics in Section 4.2.

4.1 Main Results in Comparison to Stanovsky et al. (2019)

For each MT model, four metrics, namely *Accuracy* (Acc), Δ_G, Δ_S, and Δ'_S, are utilized for evaluation. These metrics are explained in greater detail in Section 3.3. The results are presented in Table 4 and compared with the findings of Stanovsky et al. (2019) in Table 5. The average performance results for all models and for each language pair in this thesis, as well as in the research of Stanovsky et al. (2019), are displayed in Table 6. The Subsections 4.1.1, 4.1.2 and 4.1.3 will discuss the results of this thesis (WinoMTDE) in detail and compare them to the findings of Stanovsky et al. (2019) (WinoMT) for each metric.

	Google Translate				**Microsoft Translator**				**Amazon Translate**				**SYSTRAN**				**DeepL**			
Languages	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$
DE→ES	66.8	11.9	6.5	15.6	62.0	16.8	2.1	11.1	72.7	5.2	-1.7	6.8	**94.1**	0.1	2.4	6.6	83.1	6.4	5.9	5.6
DE→FR	64.2	12.1	4.3	16.2	69.2	6.2	7.8	20.9	68.0	5.7	3.7	24.5	80.6	1.5	-5.0	-2.7	**83.3**	0.4	-4.9	-2.3
DE→IT	52.0	26.2	6.7	14.2	51.8	31.8	4.7	14.4	58.9	16.8	5.5	13.2	**70.9**	7.7	-0.1	6.0	61.9	15.8	7.7	13.7
DE→UK	46.5	14.7	-4.6	11.6	48.2	18.8	-9.1	4.0	41.4	27.4	-4.4	8.0	38.2	27.4	-14.2	-8.2	**54.7**	8.2	0.2	11.7
DE→RU	42.7	19.4	-7.2	6.4	46.4	15.6	-7.3	8.2	**47.3**	15.6	-7.3	8.2	37.0	22.5	-6.0	6.9	42.3	15.5	-16.0	-3.0
DE→AR	55.2	18.3	7.3	9.0	54.0	20.8	-2.5	9.2	**59.2**	15.3	1.0	7.5	51.5	24.3	9.2	10.9	-	-	-	-
DE→HE	64.5	3.8	14.8	17.5	**65.4**	1.9	16.1	20.9	60.3	10.0	13.1	18.7	44.6	16.1	15.6	18.4	-	-	-	-

Table 4: Results of this thesis for all language pairs[11]. Languages are grouped into their respective language families: Romance, Slavic, and Semitic. The highest accuracy result for each language pair (row-wise) is highlighted in bold, while the best result for each MT model (column-wise) is underlined. DeepL is unable to translate German to either Arabic or Hebrew, which is why the corresponding cells are left empty.

[11] A vertical version can be found in A.3

	Google Translate				Microsoft Translator				Amazon Translate				SYSTRAN			
Languages	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$
EN→ES	53.1	23.4	21.3	-	47.3	36.8	23.2	-	**59.4**	15.4	22.3	-	54.6	46.3	15.0	-
EN→FR	**63.6**	6.4	26.7	-	44.7	36.4	29.7	-	55.2	17.7	24.9	-	45.0	44.0	9.4	-
EN→IT	39.6	32.9	21.5	-	39.8	39.8	17.0	-	**42.4**	27.8	18.5	-	38.9	47.5	9.4	-
EN→UK	38.4	43.6	10.8	-	**41.3**	46.9	11.8	-	-	-	-	-	28.9	22.4	12.9	-
EN→RU	37.7	36.8	11.4	-	36.8	42.1	8.5	-	**39.7**	34.7	9.2	-	37.3	44.1	9.3	-
EN→AR	48.5	43.7	16.1	-	47.3	48.3	13.4	-	**49.8**	38.5	19.0	-	47.0	49.4	5.3	-
EN→HE	**53.7**	7.9	37.8	-	48.1	14.9	32.9	-	50.5	10.3	47.3	-	46.6	20.5	24.5	-
EN→DE	59.4	12.5	12.5	-	**74.1**	0.0	30.2	-	62.4	12.0	16.7	-	48.6	34.5	10.3	-

Table 5: Results from Stanovsky et al. (2019) for all language pairs grouped by language family. The highest accuracy result for each language pair is highlighted in bold, and the best result for each MT model is underlined. The column for $\Delta_{S'}$ remains empty as this metric is an addition of this thesis.

	Acc		Δ_G		Δ_S		$\Delta_{S'}$	
Languages	DE	EN	DE	EN	DE	EN	DE	EN
S→ES	75.7	51.4	8.1	30.5	3.0	20.5	9.1	-
S→FR	73.1	52.1	5.2	26.1	1.2	22.7	11.3	-
S→IT	59.1	40.2	19.7	37.0	4.9	16.6	12.3	-
S→UK	45.8	36.2	19.3	37.6	-6.4	11.8	5.4	-
S→RU	43.1	37.9	17.7	39.4	-8.8	9.6	5.1	-
S→AR	55.0	48.2	19.7	45.0	3.8	13.5	9.2	-
S→HE	58.7	49.7	8.0	13.4	14.9	35.6	18.9	-

Table 6: Performance average of all models for each language pair grouped by language family compared to the findings of Stanovsky et al. (2019). 'S' denotes the source language, which can be either DE (results of this thesis) or EN (results of Stanovsky et al. (2019)). The column for Δ'_S is left blank as it is not available for the results of Stanovsky et al. (2019). German as the target language is excluded since comparable results from this thesis are unavailable.

4.1.1 Accuracy (Acc)

Acc represents the percentage of instances where the model preserves the original gender. The optimal value is 100% and 50% denotes random guessing.

WinoMTDE. The accuracy results of this thesis, as shown in Table 4, range from 37.0% to 94.1% for all models and language pairs, indicating significant variability. As depicted in Table 4, the accuracy of MT models concerning gendered instances mostly remains below 60%, with an average of 58.8%. Notably, Google Translate consistently performs worse, never achieving an accuracy exceeding 67% and consistently being outperformed by other models. When comparing the accuracy of the NMT models, DeepL, claiming to surpass Google Translate, Microsoft Translator and Amazon Translate usually performs the best. It outperforms the other NMT models in four out of five cases. SYSTRAN, which combines SMT and rules, excels in two out of three cases within the Romance language family and is the sole model that achieves an accuracy above 70% in the DE-IT case. Furthermore, it reaches an accuracy of 94.1% in the DE-ES case, which is 10% higher than any other recorded accuracy. In contrast, its performance within the Slavic language family is the worst among all models.

In general, it is evident that all models perform the least biased when translating from German to the Romance language family, with an average precision of 69.3%, while translations to Italian consistently yield the lowest accuracy. The mean accuracy for the Slavic family, where all MT models exhibit the most bias, is 44.5%, which is inferior to random guessing. Only one model, DeepL, manages to surpass random guessing with an accuracy of 54.7%. The average accuracy for the Semitic family is 56.9%, with slightly superior results for Hebrew (58.7%) compared to Arabic (55.0%).

WinoMT. The results presented in Table 5 by Stanovsky et al. (2019) exhibit similar patterns. However, the average accuracy is generally lower, with most of the top-performing results even falling below the percentage of random guessing. Within the Romance language family, translations into French and Spanish surpass Italian with a more pronounced difference than what this thesis' results indicate. Notably,

the average percentage of gender preservation during translations into Italian is even worse than that observed for Russian, which is the language where models performed the least effectively according to the findings of this thesis. Translations into the Slavic family perform strongly biased in the evaluation conducted by Stanovsky et al. (2019) as well, with an average accuracy of 37.1%. In contrast to the results displayed in this thesis, Stanovsky et al. (2019) find that the average translation accuracy for the Semitic family is the highest among the three language families. The average precision for Arabic and Hebrew is 48.2% and 49.6%, respectively. Stanovsky et al. (2019) also assessed the performance of the MT models when German is chosen as the target language. Nearly all models perform at their best when translating into German, with an average accuracy of 61.1%.

4.1.2 F1-Score Difference Δ_G between Male and Female Instances

Δ_G represents the difference in performance between female and male instances, with the optimal value being 0%. If Δ_G falls below zero, the MT model performed better when translating female instances and for a Δ_G above zero the performance was better for male instances.

WinoMTDE. The results of this thesis show that the performance of all models improves when translating male instances, with an average Δ_G of 14.0% across all languages and models. The best Δ_G is achieved by SYSTRAN, at 0.1%, when translating from German to Spanish. On the other hand, the worst Δ_G is recorded by Microsoft Translator, at 31.8%, when translating from German to Italian.

WinoMT. In contrast, the results of Stanovsky et al. (2019) find a significantly stronger male bias within the MT models. The average Δ_G across all languages is 32.7%, more than twice as high as the average Δ_G observed in the results of this thesis.

4.1.3 Acc Difference Δ_S and Δ'_S between Stereotypical and Anti-Stereotypical Instances

Δ_S is defined as the difference in accuracy between stereotypical and anti-stereotypical instances with the instances and ground truth stemming from WinoMTDE$_{anti.wb}$ and WinoMTDE$_{pro.wb}$ and the optimal value being 0%. If Δ_S falls below zero, the MT model

performed better on translating anti-stereotypical instances and for a Δ_S above zero the performance was better for stereotypical instances.

WinoMTDE. Within the Romance language family, most models perform better when translating stereotypical instances, with an average Δ_S of 3.0%. In the Slavic languages, the average difference is -7.6%, and all models, except DeepL for Ukrainian, perform worse on stereotypical instances. Translations to Hebrew and Arabic both perform better on stereotypical instances, with an average Δ_S of 14.9% and 3.8%, respectively.

WinoMT. As Table 6 shows, the results vary significantly between the two studies. In the evaluation of Stanovsky et al. (2019), all models perform dramatically better when translating stereotypical instances, with Microsoft Translator performing nearly 50% better when translating stereotypical gender-role assignments from English to Ukrainian and Arabic.

The definition of Δ'_S is the same as that of Δ_S, with the only difference being that the accuracies for stereotypical and anti-stereotypical stem from the subsets WinoMTDE$_{anti}$ and WinoMTDE$_{pro}$. More information on the subsets can be found in Section 3.2.

The results of this thesis show that most models perform better on stereotypical instances when confronted with the majority of the challenge set. SYSTRAN and DeepL are exceptions to this trend as their translation to French, SYSTRAN's translation to Ukrainian, and DeepL's translation to Russian perform better on anti-stereotypical instances. Δ_S is averaging 1.8% across all languages and models and Δ_S in comparison averages at 10.2%. Compared to the values of Δ_S of Stanovsky et al. (2019), the average of Δ'_S, even though being higher than the average Δ_S of this thesis, is closer to 0% as the research of Stanovsky et al. (2019) shows an average Δ_S of 18.6% across all languages and models.

4.2 Results in Relation to the Occupation Statistics

As discussed in Section 3.3 each job instance in WinoMTDE was classified using statistics from the German Department of Labor, which are displayed in Table 7. Each job instance is then assigned to an occupational group with a unique code (see A.2).

Table 7: Occupation Statistics of the German Department of Labor. All occupational groups present in the dataset are displayed. Code denotes the labeling of "Klassifikation der Berufe 2010 – überarbeitete Fassung 2020" with each occupation having a unique code for reference. Furthermore, the distribution of female and male persons working in each field is showcased. The German Department of Labor did not include additional statistics for persons of other genders. The prominent gender in each occupation is highlighted in italics.

Code	Occupational Group Name	Women (in %)	Men (in %)
111Lw	Landwirtschaft	31.48	*68.52*
223Ho	Holzbe- und -verarbeitung	8.19	*91.81*
251Ma	Maschinenbau- und Betriebstechnik	12.33	*87.67*
252FL	Fahrzeug-Luft-Raumfahrt-,Schiffbautechn.	5.66	*94.34*
262Et	Energietechnik	4.04	*95.96*
263El	Elektrotechnik	17.69	*82.31*
282Tv	Textilverarbeitung	*70.43*	29.57
292Le	Lebensmittel- u. Genussmittelherstellung	32.83	*67.17*
293Sp	Speisenzubereitung	*50.73*	49.27
311BA	Bauplanung u. -überwachung, Architektur	30.32	*69.68*
321Hb	Hochbau	1.79	*98.21*
332MB	Maler.,Stuckat.,Bauwerksabd,Bautenschutz	6.46	*93.54*
341Gt	Gebäudetechnik	10.22	*89.78*
342KS	Klempnerei,Sanitär,Heizung,Klimatechnik	1.16	*98.84*
413Ch	Chemie	34.66	*65.34*
434Pr	Softwareentwicklung und Programmierung	13.9	*86.1*
521Fa	Fahrzeugführung im Straßenverkehr	8.39	*91.61*
531OP	Obj.-,Pers.-,Brandschutz,Arbeitssicherh.	26.17	*73.83*
532Po	Polizei,Kriminald.,Gerichts,Justizvollz.	37.84	*62.16*
541Re	Reinigung	*77.31*	22.69
621Vo	Verkauf (ohne Produktspezialisierung)	*69.66*	30.34
632Ho	Hotellerie	*75.07*	24.93

Continued on next page

Code	Occupational Group Name	Women (in %)	Men (in %)
633Ga	Gastronomie	*61.86*	38.14
711GV	Geschäftsführung und Vorstand	22.21	*77.79*
714BS	Büro und Sekretariat	*77.66*	22.34
715PD	Personalwesen und -dienstleistung	*73.0*	27.0
721VF	Versicherungs- u. Finanzdienstleistungen	*54.89*	45.11
722Rw	Rechnungswesen, Controlling und Revision	*65.81*	34.19
731Rb	Rechtsberatung, -sprechung und -ordnung	*72.69*	27.31
732Ve	Verwaltung	*75.44*	24.56
733MD	Medien-Dokumentations-Informationsdienst	*75.45*	24.55
813GK	Gesundh.,Krankenpfl.,Rettungsd.Geburtsh.	*78.72*	21.28
814Hm	Human- und Zahnmedizin	*53.91*	46.09
815Tm	Tiermedizin und Tierheilkunde	*77.15*	22.85
816Py	Psychologie, nichtärztl. Psychotherapie	*81.11*	18.89
818Ph	Pharmazie	*82.84*	17.16
822EW	Ernährungs-,Gesundheitsberatung,Wellness	*76.16*	23.84
823Kp	Körperpflege	*83.97*	16.03
832Hw	Hauswirtschaft und Verbraucherberatung	*92.75*	7.25
841Le	Lehrtätigkeit an allgemeinbild. Schulen	*72.05*	27.95
842LA	Lehrt.berufsb.Fächer,betr.Ausb.,Betr.päd	*52.76*	47.24
843LF	Lehr-,Forschungstätigkeit an Hochschulen	47.32	*52.68*
924RJ	Redaktion und Journalismus	*55.14*	44.86
931PI	Produkt- und Industriedesign	38.51	*61.49*
947Mt	Museumstechnik und -management	*53.91*	46.09
0	Allgemein	49.6	*50.4*

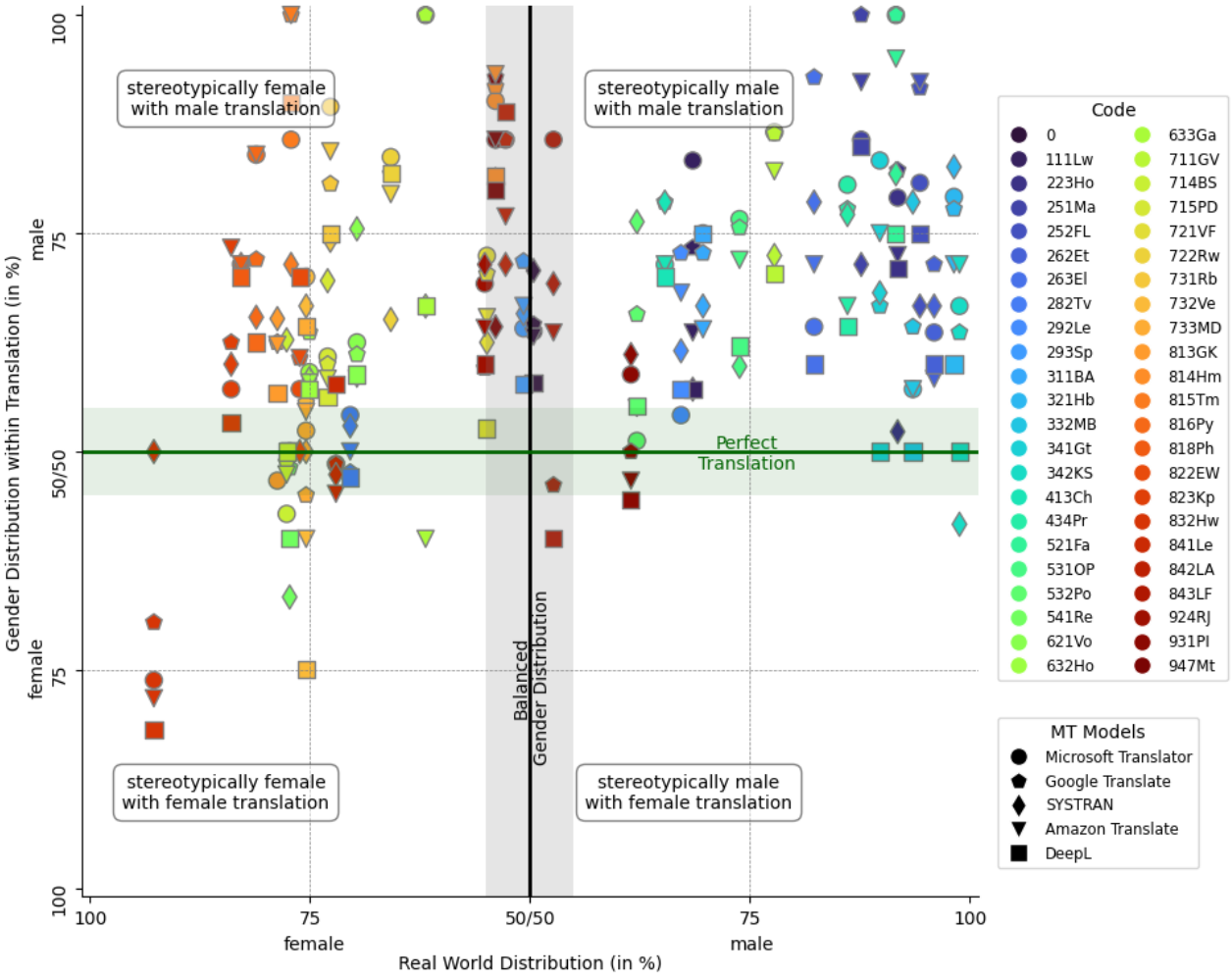


Figure 7: Gender distribution in relation to the real-world distribution of each occupation group. All gender predictions of a translation by an MT model across all languages for one occupation group were summed up and used for this Figure. Each color denotes a different professional class. With blue hues denoting agricultural, manufacturing and construction jobs. Turquoise hues are jobs in natural sciences, logistics, transportation and security. The green instances are professions in cleaning, tourism and trade in goods. Greenish-yellow occupations are managing positions, office and human resources jobs. Yellow hues represent jobs in accounting, finance and law. Orange hues are jobs in health care. Red denotes professions in education and social work. Dark red hues are jobs in media, journalism and design. The black instances are general job descriptions. Each symbol in the figure represents a different MT model. The x-axis corresponds to Table 7 and the real-world distribution of each occupation group, ranging from 100% female workers on the left to a 50% (50% male) balance in the middle, and finally to 0% (100% male) on the right. The grey vertical line marks occupations with minimal gender imbalance in the real world. The y-axis represents the gender distribution within the translated challenge set. An ideal translation would result in all markers aligning with the green horizontal line, indicating preserved original distribution.

Using the statistics displayed in Table 7 and the evaluations of the different MT models, relations to the job statistics can be drawn as illustrated in Figure 7. The gender distribution in relation to the real-world distribution of each occupation group is displayed in Figure 7 and can be split into four quadrants:

Stereotypically male with male translation. In the top-right quadrant, the observable instances are those, where the workforce predominantly comprises males, and MT models inaccurately translate most of them as male. This quadrant primarily encompasses professions in manufacturing, construction, engineering, natural sciences, and the occupational group *711GV*, which refers to managerial positions and board members. Google Translate's (indicated by the polygon) translations assigns a 100% male distribution to the occupational group *mechanical and industrial engineering* (251Ma), mirroring Microsoft Translator's translations (marked with a circle), which exhibit a 100% male workforce in the domain of *drivers of road vehicles* (521Fa). Notably, the majority of translations displaying over 75% male instances are produced by Amazon Translate (marked with a triangle), Microsoft Translator, and Google Translate. In contrast, SYSTRAN (represented by a diamond shape) and DeepL (referenced by a square) seldom exceed the 75% threshold. It's noteworthy that professions where the male gender stereotype is less pronounced, i.e., those closer to the vertical line, tend to be translated with higher accuracy, i.e. are closer to the horizontal line.

Stereotypically female with male translations. The top-left quadrant encompasses all occupation groups that are stereotypically associated with female persons, yet the MT model translates more than 50% of them as male. Predominantly, jobs in this quadrant belong to the general fields of health care, accounting, law, and education. The blue occupation corresponds to the manufacturing profession of *textile processing* (282Tv), while the green instance represents roles in the tourism and trade in goods sector. Once again, it is evident that Amazon, Google, and Microsoft Translator's translations frequently exceed the 75% threshold. Notably, Google Translate assigns a 100% male distribution to professions in *gastronomy* (633Ga) and *veterinary medicine* (815Tm), as do Amazon Translate and Microsoft Translator. Occupations that exhibit a more balanced gender distribution in the

real world, falling within the grey area, tend to be translated as more male, such as *education in training* (842LA), *human and dental medicine* (814Hm), and *museum technology and management* (947Mt).

The green horizontal line denotes translations, where the gender distribution is balanced as within WinoMTDE. DeepL's and SYSTRAN's translations are the ones that preserved the gender correctly the most.

Stereotypically female with female translation. The bottom-left quadrant encompasses occupations that are stereotypically associated with females, and the translations by the MT models also feature a higher proportion of female gendered individuals. It is evident that there are significantly fewer instances in this quadrant compared to the upper two. Professions in this quadrant predominantly belong to the fields of cleaning, housekeeping, administration, and education. There is no occupation groups, where all models translated more females than males. Only *housekeeping* (832Hw) exhibits a female bias, with four out of five models generating more female than male translations. Notably, Google Translate and SYSTRAN display the least female bias in their translations, as none of their translations surpass the 75% female threshold. Once again, it is observed that the higher the proportion of female workers in a profession, the more instances are translated as female.

Stereotypically male with female translation. In the bottom-right quadrant, the fewest instances can be observed. These are cases where the stereotypical workforce is male, but MT models have translated the majority as female. Only five instances belonging to three occupation groups are displayed: *product and industrial design* (931PI), *museum technology and management* (947Mt), and one instance of jobs in *plumbing, sanitary, heating, and air conditioning technology* (342KS). Notably, the latter occupation group exhibits the most significant gender disparity in the real world, with only 1.1% female representation. It's important to note that none of the translations in this quadrant originate from Microsoft Translator, and the majority are from DeepL. Generally, all instances are much closer to the horizontal line than those in the other quadrants.

The occupation group with the largest range of differing gender translations across all MT models is *gastronomy* (633Ga). The one with the smallest disagreement is *textile process*

(282Tv), which is also the one where the gender distribution within the translations is closest to 50%.

4.3 Detailed Gender Distributions of Translations

This Section will further discuss the gender distribution within the translations of Microsoft Translator, DeepL and SYSTRAN.

Gender Distribution within Translations across all Language Families. Figure 8 displays the detailed gender prediction distribution of translations by Microsoft Translator, DeepL and SYSTRAN to all languages. The results of Google Translate and Amazon Translate as well as versions suited for colorblindness can be found in the Appendix A.4 and A.5.

Microsoft Translator is the model that performed best in the evaluation of (Stanovsky et al., 2019). It can be examined that this MT model generally translates much fewer female instances than the actual 50%. For two professions, namely *gastronomy* (633Ga) and *drivers of road vehicles* (521) a female percentage of 0% was recorded, with circa 25% of ground truth female instances being translated as male and the rest being either unknown or neutral. The only two occupations where the female share reaches 50% are *housekeeping* (832Hw) and *cleaning* (541Re). Overall the share of female instances stays under 25% for ca 50% of occupations. The share of male translations is largest within the profession *human and dental medicine* (814Hm) and smallest within *housekeeping* (832Hw). The occupation where the translated genders are closest to the ground truth is *Police, Criminal Office, Law* (532Po). Nearly 57% of occupations include a false female part, while all include a false male share, which is generally higher.

DeepL is an NMT model that this thesis evaluated additionally, and it displays a similar trend. Generally the share of female translations is significantly lower than the share of male translations. Circa 35% of occupations display a female percentage below 25%. In contrast to Microsoft Translator, each occupation group still includes some share of female translations, but only 5 professions reach a percentage of 50% female

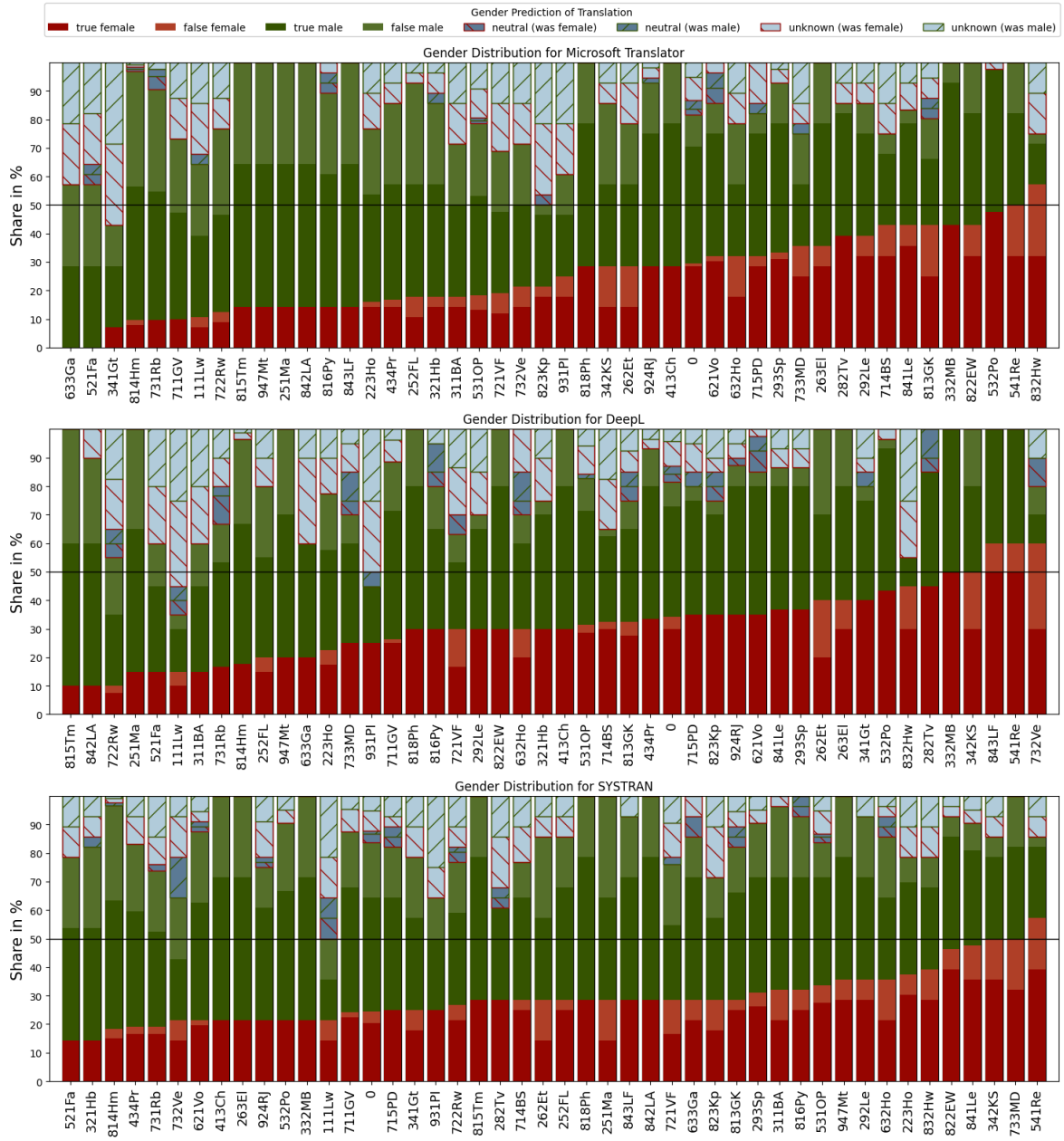


Figure 8: Gender Distribution of Translations by Microsoft Translator, DeepL and SYSTRAN. For all occupation groups (x-axis) the distribution (in %) of female (red), male (green), neutral (blue) and unknown (light blue) instances is displayed. The darker red shade corresponds to the true female translation, i.e. female ground truth terms, where the gender was preserved. The light red denotes the share of false female translation, i.e. a male instance being translated to a female instance. The dark and light green bar indicate the percentage of true and false male translations. Within the neutral part a red, left hatch denotes the percent of instances originally of female gender contributing to the neutral percentage. A green, right hatch corresponds to a male ground truth gender. The same symbolism is used to display the distribution within the unknown instances. The horizontal line displays the ground truth distribution with the perfect translation consisting of 50% male and 50% female instances.

translations, namely *painting, plastering and building protection* (332MB), *plumbing, sanitary, heating and air conditioning technology* (342KS), *cleaning* (541Re), *education and research at a university* (843LF) and lastly *administration* (732Ve), with the latter 3 crossing the 50% mark.

The share of male instances is largest within the profession *veterinary medicine* (815Tm) with DeepL translating 90% of instances to male. Once, all gender assignments were preserved, this was during the translation of occupations within the group *painting, plastering and building protection* (332MB). Circa 90% of occupations groups include false male instances, whereas only 40% include false females. The mean of the share of false males and false females of DeepL is the lowest within all evaluation models with 12.6% and 3.7% respectively.

SYSTRAN is the only HMT model that was evaluated in this thesis. Similarly to the other two models less female instances are translated correctly in comparison to male subjects. Approximately 30% of occupations display a female percentage less than 25%, but each occupation group has at least a female share of 15% which is the highest out of the all evaluated models. A female percentage within the translations of at least 50% was generated thrice, namely in the professions *cleaning* (541Re), *plumbing, sanitary, heating and air conditioning technology* (342KS) and *media and documentation information service* (733MD). The professions where the male share is the largest are *human and dental medicine* (814Hm), *chemistry* (413Ch) and *electrical engineering* (263El). A 50/50 translation was made for the occupation group *media and documentation information service* (733MD), but it is to be seen that not all translations were correct, but instead circa 20% false female and false male translations. Nearly all translation within an occupation group include false male translations, while only 30% display some (generally lower) share of false female translations, with a respective mean of 19.1% and 5.5%.

In conclusion, all MT models translate female instances correctly less often, as the share of females is much lower in contrast to the share of males. Furthermore, it is observable that the share of translations where the gender can not be predicted, i.e. is unknown, is higher within the NMT models. Reasons for this are further discussed in Section 5.2.

Gender Distribution within Translations to the Romance Language Family. As seen in Table 4 all evaluated MT models achieved their best score within the Romance language family. The gender distribution within translations to these languages are displayed in Figure 9. A version suited for colorblindness is appended in A.6.

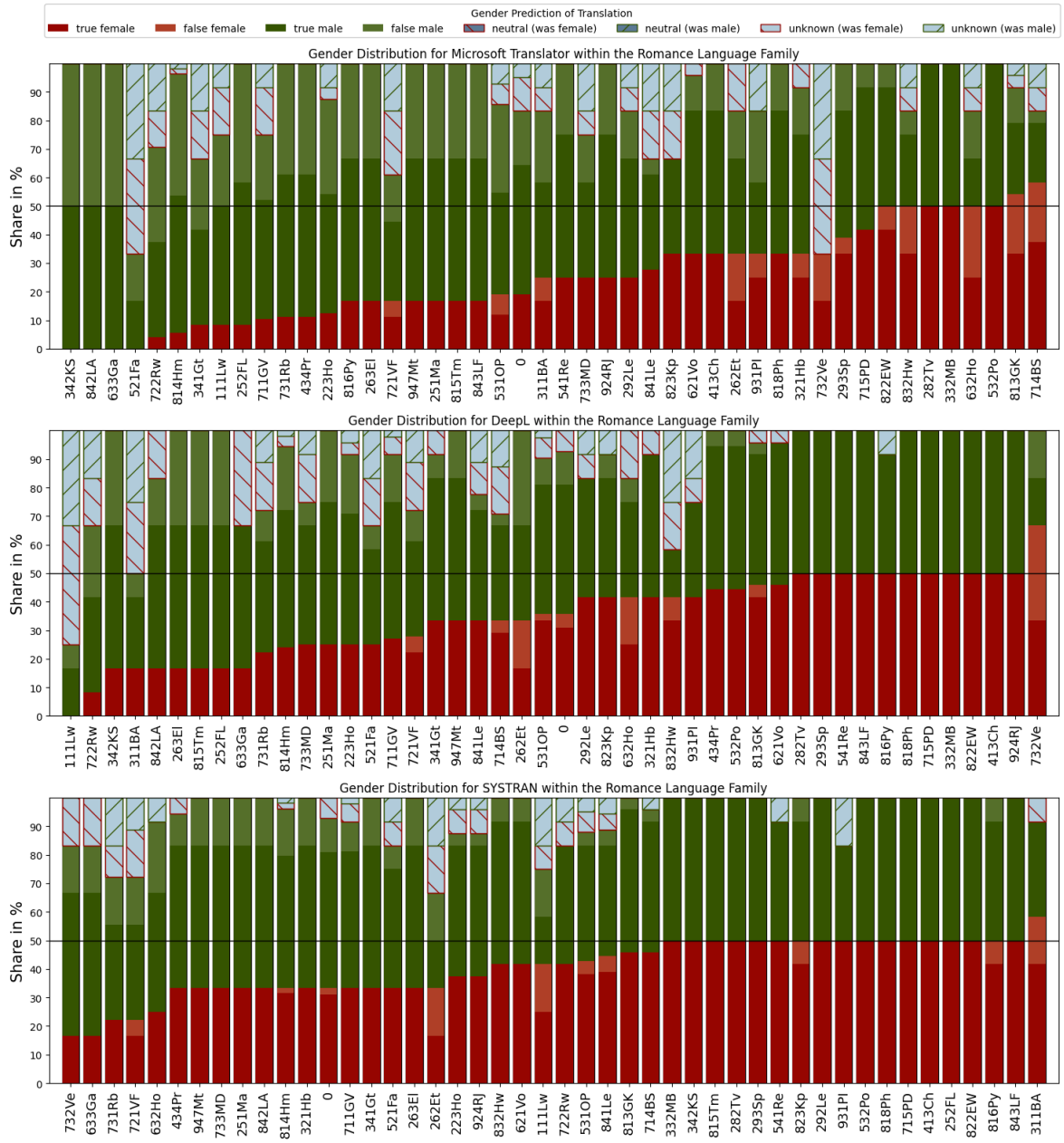


Figure 9: Gender Distribution by Microsoft Translator, DeepL and SYSTRAN within the Romance language family. For all occupation groups the percentage distribution of all possible gender predictions is displayed. These are only the results of the Romance language family summed up. Red hues denote true and false female instances. Green hues true male and false male instances. Blue and light blue refer to the share of unknown and neutral instances, where the hatching indicates the ground truth gender.

The observed trend is similar to that of the gender distribution across all languages.

Microsoft Translator does not translate any share of females in four occupation groups, but in nearly 18% a female share of 50% or more was translated. A perfect translation was achieved three times.

DeepL clearly performs better within the Romance language family. Ten perfect translations were achieved, but in one instance, in the professional group *111Lw*, no females were recorded.

SYSTRAN performed the best within the Romance language family as seen in Table 4 and this is further proven within Figure 9. 30% of translations are perfect and there is no occupation group were only males were translated correctly. The false translation share is the lowest with a mean of 1.9% for females and 7.8% for males. Furthermore, the mean of females within all occupations is generally the closest to the ground truth with nearly 41%.

An overall improvement within the translation quality regarding gender bias can be observed in Figure 9. This supports the result that the MT models perform best within the Romance language family. Furthermore, it is to be observed that SYSTRAN, which is the only HMT model, performs best within the Romance language family, even considering the percentage of unknown predictions as Figure 9 shows.

5 Discussion and Limitations

This Chapter discusses the results of this thesis in Section 5.1 and the limitations of the methodology used within Section 5.2. Furthermore, Section 5.3 gives an outlook on future work.

5.1 Discussion

Previous work on gender bias in MT has shown that different models create biased translations and gain performance when translating male instances (see Section 2.3.2) in comparison to translating female subjects. The results of this thesis suggest the same for German Machine Translation. First, the thesis results will be examined independently. Then a comparison to the results of Stanovsky et al. (2019) will be conducted. Lastly, the discussion will focus on the results in relation to occupational statistics.

5.1.1 Discussion of Evaluation Results using WinoMTDE

Language Families. It can be seen that MT models generally accomplish similar results within the same language families. This could be due to shared characteristics and consequently the ability of the MT model to pick up similar amounts of semantic gender hints. Moreover, as discussed in Section 2.1.1, translation quality for both NMT and SMT models relies on a vast amount of parallel text corpora.

Translations into Romance languages typically perform best, likely because of the geographic proximity of Spain, Italy, and France as well as their membership in the European Union, which provides parallel corpora between German and Romance languages. In contrast, translations to Slavic and Semitic languages exhibit more bias, possibly due to grammatical structure differences and the geographical distance of native-speaking countries from Germany.

Variations within language families may be linked to factors like the number of speakers. For example, Italian consistently lags behind possibly due to its smaller speaker base (67.9 million), leading to a smaller parallel corpus. However, the number of speakers alone doesn't explain everything, as seen in the Semitic language family, where Arabic (274

million speakers) performs worse than Hebrew (9.3 million speakers). This discrepancy could be related to the popularity of Christianity in Germany and the availability of abundant parallel data for the German-Hebrew pair in the form of theological texts, as the Old Testament is originally written in Hebrew.

Nonetheless, this can only be speculated as the exact composition of the parallel corpora used by the evaluated MT models is unknown, and multiple complex factors are likely to influence their performance.

MT models. The MT models evaluated in this thesis are widely used translation tools and are expected to perform well. However, the results of this thesis show that MT models generate translations that exhibit gender bias.

Accuracy is a measurement to evaluate the general translation quality regarding gender bias and varies greatly across the assessed MT models. Especially interesting are the results of Google Translate and SYSTRAN, which will further be discussed in the following:

Google Translate. The MT model that is always outperformed by others in the metric Acc is Google Translate, which is among the most popular translation tools (see Section 3.1). Google Chrome for example utilizes this translation tool to prompt users to automatically translate webpages into their preferred language. An example is given in Table 8.

Original Text	**Translated Text**
Angela Merkel wurde am 17. Juli 1954 in Hamburg geboren. Vom 22. November 2005 bis zum 8. Dezember 2021 war *die CDU-Politikerin*$_f$ Bundeskanzlerin der Bundesrepublik Deutschland und damit die mächtigste *Regierungschefin*$_f$ in Europa. Beiträge und Hintergründe rund um *die Altkanzlerin*$_f$ finden Sie auf unserer Themenseite.	Angela Merkel nació el 17 de julio de 1954 en Hamburgo. Del 22 de noviembre de 2005 al 8 de diciembre de 2021, *el político*$_m$ de la CDU fue canciller de la República Federal de Alemania y, por tanto, *el jefe*$_m$ de gobierno más poderoso de Europa. Puede encontrar artículos e información general sobre *el ex canciller*$_m$ en nuestra página temática.

Table 8: Example of a translation by Google Translate of a webpage via Google Chrome. The original text is in German and the translated text is in Spanish. The text is taken from https://www.tagesspiegel.de/politik/themen/angela-merkel. The gender of the occupation referrals to the subject of interest that are gendered in both languages are marked with a subscript.

When examining this example gender bias within Google Translate becomes obvious. The female ex-chancellor of Germany Angela Merkel gets misgendered to the extent, that, without further gender information from either name or pre-knowledge, it must be presumed that the subject of interest within this text is male.

SYSTRAN. The only HMT model that this thesis evaluated is SYSTRAN, which utilizes SMT and RBMT to generate a translation. It performed best within all metrics in the Romance language family. This is especially interesting as it indicates that using set grammatical rules could be a possibility to minimize gender bias within MT from German to Romance languages. However, SYSTRAN performed worse than the other MT models within the Slavic and Semitic language families. This could be due to the fact that the grammatical rules to translate to those languages are more complex and therefore harder to implement.

Δ_G, Δ_S and $\Delta_{S'}$ measure the difference in performance between either male and female instances (Δ_G) or stereotypical and anti-stereotypical instances (Δ_S, Δ'_S). An improvement within translations by all MT models when confronted with male instances can be observed within the results of this thesis. Similarly, the performance of a MT model increases when translating stereotypical instances as opposed to anti-stereotypical occupation titles. Both these phenomena can be explained within the NMT models by how these models generate a translation.

NMT models are trained on parallel corpora and therefore learn to translate a sentence by predicting the next word in the target language given the previous words in the source language. If these corpora mainly consist of male instances, the learned probability that a subject of interest is of male gender is higher as the model learned that male instances are generally more likely.

Furthermore, as explained in Section 2.1.1 these models use word embeddings to generate a translation. Occupation titles, where the workforce is stereotypically female are more likely to be used within the context of subjects of female gender. Therefore, the word embedding reflects this and produces stereotypical translations. This becomes especially apparent within the results of $\Delta_{S'}$, which is calculated us-

ing the translation accuracies of WinoMTDE$_{pro.wb}$ and WinoMTDE$_{anti.wb}$. These subsets include all job titles that are included in the WinoMT challenge set as opposed to the subsets used to calculate Δ_S, where the stereotypical bias is not as pronounced. The Winogender subset included more unusual job titles such as "firefighter", "pathologist" or "specialist" in comparison to the WinoBias subset, which included more common job titles such as "teacher" and "doctor". Since the Winogender job titles are likely to be less frequently used within the training data, it is probable that the MT models are less familiar with those job titles and therefore perform worse on them in regard to stereotypes. Moreover, the sentences within the Winogender subset are longer, and as explained in Section 2.1.1 especially NMT models are prone to make more errors on longer sentences.

Subsequent harms. Within the results of this thesis gender bias is clearly observable and negative consequences in the form of representational harm emerge. Savoldi et al. (2021) defines two types of representational harm: *Underrepresentation* (see 2.3 and reflected in Δ_G) and *Stereotyping* (see 2.3 and displayed in Δ_S and $\Delta_{S'}$). Both types of representational harm can be observed in the translation of the WinoMTDE challenge set and examples are given in Table 9.

Original Sentence	**Translated Sentence**	**Observed Phenomena**
Der Arbeiter$_m$ sagte der Fußgängerin$_f$, sie solle den geschlossenen Teil der Straße meiden.	El trabajador$_m$ le dijo al peatón$_m$ que evitara la parte cerrada de la carretera.	*Underrepresentation* The female pedestrian gets translated to male pedestrian. This contributes to a general reduced visibility of women without considering occupation titles.
Die Managerin$_f$ feuerte den Reiniger$_m$, weil sie wütend war.	El gerente$_m$ despidió a la limpiadora$_f$ porque estaba enojada.	*Stereotyping* The gender of the subject of interest "manager" gets translated as male incorrectly. Furthermore, although not included in the statistics, the gender of cleaner changes from male to female as well. Therefore, contributing to harmful occupational stereotyping.

Table 9: Examples of representational harm observed within translations by Microsoft Translator. With the original sentence denoting the German ground truth, the translated sentence and the observed phenomena. Furthermore, the gender of each noun is marked with a subscript.

This thesis solely evaluated the MT models regarding gender bias, but as laid out in Section 2.1.2 there are other translation qualities to consider as well. Nevertheless, this thesis advocates for making gender bias an important quality feature when creating and evaluating MT models.

5.1.2 Discussion of Results in Relation to Stanovsky et al. (2019)

In comparison to the results of Stanovsky et al. (2019) all evaluated MT models perform significantly less gender biased. A probable improvement reason is that the ground truth language is changed from English to German, a grammatical gender language. Those languages embed gender in their grammatical structure more frequently than English. All subjects of interest within WinoMTDE encode gender in the noun itself additionally to personal pronoun Stanovsky et al. (2019) used. Therefore, it is easier for the MT models to pick up the gender correctly. Considering this it is surprising that the mean accuracy of the MT models is still below 60%.
This could be due to the fact that some MT models, namely Google Translate and Microsoft Translator, are known to utilize English as an interlingua (see Section 2.1.1) and therefore translate from German to English and then to the target language. This could lead to a loss of gender information and therefore a worse translation. Another reason for the improved performance on the other hand could be the fact that there is a four-year time difference between the evaluation of Stanovsky et al. (2019) and this thesis and development in the field of MT is rapid.

In conclusion the results are able to answer R2 with yes, there is an observable improvement in comparison to the results of Stanovsky et al. (2019).

5.1.3 Discussion of Results in Relation to Occupation Statistics

When digging deeper into the gender distribution in relation to occupational statistics (see Section 4.2) the occupational stereotyping and underrepresentation of females becomes even more apparent. Translations with a female subject of interest are far more likely to not preserve the gender correctly. Furthermore, a slight correlation between the distribution within the real-world and the distribution within the translations can be observed. This could be explained by the fact, that the real-world distribution is embedded within

the training data and therefore reproduces gender discrepancies. However, this correlation is not as strong as expected. This could be due to the fact, that the influence of male bias is stronger than the influence of stereotypes. This is corroborated by Figure 7, where the influence of male bias clearly is visible. Nevertheless, a stereotypical bias is observable and is especially interesting when examining the different occupation groups. A high female percentage correlates with occupation groups such as *541Re* (cleaning). Whereas a high male percentage is found in medical professions such as *814Hm* (human medicine) and *815Tm* (veterinary medicine), even though the share of female gendered individuals within the workforce is above 50%. This phenomenon is further addressed within Section 5.3 as it indicates that the stereotypes are not solely based on the real-world distribution, but rather on perceived stereotypes.

5.2 Limitations

WinoMTDE is a German challenge set created for the purpose of evaluating gender bias in German MT. However, it is not without limitations.

Size. WinoMTDE only contains 288 sentences and was translated manually. Therefore, it is more error-prone in comparison to the English WinoMT challenge set, which contains 3168 automatically generated sentences.

Cultural Differences. A translated English dataset is not necessarily representative of the German culture and therefore some sentences do not make as much sense in German as they do in English. For examplification the phrase *"Der Zuschauer rief 911 an und sprach mit der Disponentin, die sagte, sie habe Hilfe geschickt."* can be examined. 911 is the emergency number in the United States, whereas in Germany the emergency number is 112. Therefore, it may be possible, that the MT models are not able to translate the phrase correctly, because it is not possible to link the emergency number to the occupation of the subject of interest.

Stereotype Annotation. After translating the English job titles each occupation was annotated using statistics from the German Department of Labor. However, the job title translation were sometimes unclear. For example the occupation *UntersucherIn* has different meanings in German. It can either denote a medical examiner or an

investigator, with each one being part of a different occupation group and therefore a different gender distribution. Therefore, the occupation group had to be determined by the context of the sentence. This was done by one person and therefore is prone to errors and bias. Moreover, while annotating a hard cut at 50% was made. This leads to different job titles being classified as either pro- or anti-stereotypical, even though they are close to the 50% mark, e.g. *RedakteurIn* is annotated as being stereotypically female, even though the real-world distribution is relatively neutral with 55% female workers. Besides that the division of occupations and therefore stereotypes is not very fine-grained. This becomes especially apparent within the occupation group *health and nursing care, emergency service and obstetrics*, where very different occupations, such as *nurse* and *dispatcher* are grouped together.

Other Forms of Bias. Stanovsky et al. (2019) stated that even their dataset is not able to encompass all forms of bias and neither can WinoMTDE. An example of this is semantic derogation, which is a form of stereotypical bias that is present within the evaluated translations, but not further examined in this thesis. For instance the two phrases *"Die Lehrerin$_f$ traf sich mit einem Schüler$_m$, um ihre Bewertungspolitik zu besprechen."* and *"Der Lehrer$_m$ traf sich mit einer Schülerin$_f$, um seine Bewertungspolitik zu besprechen."* got translated by Amazon Translate to *"La maestra$_f$ se reunió con una estudiante$_f$ para hablar sobre su política de evaluación"* and *El profesor$_m$ se reunió con un estudiante$_m$ para hablar sobre su política de evaluación.*. *La maestra* is commonly used to denote elementary and middle school teachers, whereas *el profesor* is used to denote university professors and high school teachers. Therefore, the translation of *teacher* to *maestra* is semantically derogatory in comparison to the translation of *teacher* to *profesor*.

Gender Diversity. WinoMT contains neutral sentences, which are not included in WinoMTDE as German has neither non-binary pronouns nor neutral singular nouns. Therefore, this thesis is treating gender as a binary term and is not able to evaluate gender bias to its full extent.

Methodology. This thesis used automatic alignment, automatic morphological analysis, and different metrics to evaluate the MT models. However, these methods are not flawless.

Metrics. The example used in 5.2 reveals another limitation of this thesis. The second person in the example sentence did not get translated correctly either. These errors are not reflected within the metrics used in this thesis. Therefore, when examining the entire sentence as an instance and not only the subject of interest, the accuracy regarding gender bias could be lower than the calculated accuracy.

Unknown Predictions. In comparison to the results of Stanovsky et al. (2019) a larger share of unknown predictions can be observed. The reasons for this can stem from different sources. First, the MT models evaluated in this thesis are not able to translate all job titles correctly, and therefore they get predicted as *unknown*. Furthermore, it can be observed that *fast-align* is not able to align all sentences correctly. This can be due to the fact that the German language is more complex than the English language and therefore the alignment is more difficult. This would also explain why the number of unknown predictions is lower within SYSTRAN, a HMT model, that utilizes rules to generate a translation. The sentence structure is presumed to be more similar to the source sentence and therefore *fast-align* works better. But as it is unclear whether the unknown predictions are correct, the accuracy of the MT models could be higher than the calculated accuracy. Table 10 displays the accuracy of the MT models when unknown predictions are removed, labeled as Acc'. This value is always higher than the accuracy presented previously as unknown predictions are not included. The disadvantage of using Acc' is that the ground truth gender distribution is not reflected correctly as female instances are more likely to be unknown predictions. Some first manual checks were already made and indicated that most unknown gender predictions are indeed incorrect. However, a manual check of all unknown predictions is not feasible within the scope of this thesis.

	GT		MT		AT		ST		DL	
Languages	Acc'	Acc	Acc'	Acc	Acc'	Acc	Acc'	Acc	Acc'	Acc
DE→ES	78.0	66.8	72.7	62.0	83.2	72.7	96.8	94.1	89.7	83.1
DE→FR	73.0	64.2	77.8	69.2	76.2	68.0	87.0	80.6	92.5	83.3
DE→IT	65.8	52.0	63.8	51.8	76.2	58.9	82.2	70.9	81.2	61.9
DE→UK	66.4	46.5	62.1	48.2	62.5	41.4	49.4	38.2	68.2	54.7
DE→RU	60.4	42.7	63.3	46.4	62.2	47.3	52.8	37.0	60.6	42.3
DE→AR	67.4	55.2	66.0	54.0	68.2	59.2	62.0	51.5	-	-
DE→HE	72.9	64.5	77.2	65.4	69.4	60.3	52.1	44.6	-	-

Table 10: Accuracy results of this thesis without unknown gender predictions. For each MT model and all languages grouped within their respective family, the accuracy is provided. The first column displays Acc', denoting the accuracy values that do not include unknown predictions, and the second column displays the accuracy presented previously, including all predictions.

5.3 Future Work

Addressing Limitations. Future work concerning gender bias in German Machine Translation (Machine Translation) should prioritize addressing the limitations of this thesis.

Dataset Expansion. It is advisable to focus on expanding the dataset WinoMTDE. This dataset should encompass a more extensive collection of sentences and job titles. Moreover, automating its creation would minimize human error and variations in translation styles. Additionally, this dataset should better represent the German culture and not merely be a translation of an English dataset.

Stereotype Annotation. When annotating stereotypes, the usage of detailed statistics and the division of occupations into smaller groups is recommended. Furthermore, this process should involve multiple annotators to mitigate bias. Moreover, instead of categorizing job titles as solely pro-stereotypical or anti-stereotypical, introducing

a neutral category could be beneficial. Additionally, the use of statistics based on perceived stereotypes as the foundation for annotation, rather than real-world occupation statistics should be explored as the results indicate that the stereotypical bias does not only stem from gender distribution within the real world.

Methodology. To reduce the number of unknown predictions, a manual review of all such cases is recommended, despite its time-consuming nature. For enhanced comparability with the findings of Stanovsky et al. (2019), it is suggested to re-evaluate the MT models using the WinoMT challenge set, thereby minimizing the effect of time.

Expanding Research of Gender Bias within German MT. The creation of WinoMTDE and the evaluation of five different MT models is solely a first step in researching gender bias within German MT. Future work should further expand on this foundation.

Other Forms of Bias. To fully understand the impact and extent of gender bias within German MT it is needed to assess MT models for other forms of bias, such as semantic derogation.

Natural Language Dataset. A curated dataset like WinoMTDE cannot fully encapsulate the complexities of the German language due to its limited grammatical variability. Thus, using a dataset consisting of instances of natural language rather than automatically generated phrases is recommended.

Model Selection. Expanding the evaluation of German MT should involve assessing a broader range of MT models. ChatGPT for instance showed promising first results as it was manually checked for translating DE-ES. Furthermore, evaluating MT models with known architectures and training data may provide further insights into the observed bias.

Expanding Research of Gender Bias within MT. Lastly, future work should extend the evaluation to other basis languages therefore laying a foundation for cross-language comparisons that are able to draw conclusions on the influence of the source

language on gender bias in MT models. Furthermore, considering the strong performance of the HMT models SYSTRAN within the Romance language family, an evaluation of the overall translation quality (including gender bias) of HMT models in comparison to Neural Machine Translation (NMT) models should be done. Combining both approaches may offer potential solutions for minimizing gender bias within the translations done by MT models.

6 Conclusion

The work of this thesis contributes to research within Machine Translation with a focus on evaluating gender bias. The state-of-the-art evaluation method by Stanovsky et al. (2019) was expanded to German by creating WinoMTDE, a German version of the WinoMT dataset. Additionally, each occupation within this GBET was annotated using statistics from the German Department of Labor. Using this, five different state-of-the-art Machine Translations to seven languages displaying grammatical gender were evaluated. Furthermore, a new metric Δ'_S was introduced to examine stereotypical bias more thoroughly.

The results of this thesis show a significant gender bias within German Machine Translation systems. It was observed that all evaluated models perform better on male instances. Moreover, a gender bias towards stereotypical occupations was observed in most systems as well. In comparison to the results of Stanovsky et al. (2019) an improvement is evident. Reasons for this could be the implementation of a new basis language, namely German, which displays gender in its grammatical structure, as opposed to English, which does not. The time difference between the two studies could also be a contributing factor, as research in the field of Machine Translation is progressing rapidly. When examining the results of this thesis in relation to the real-world distribution of gender within different occupation groups a slight correlation is noticeable.
Furthermore, findings suggest that using HMTs could improve the quality of translations regarding gender bias to a certain degree, especially within the Romance language family, but further research is needed to confirm this.

The used code and all evaluation results are available at https://github.com/michellekappl/mt_gender_german.

References

Adler, M., & Elhadad, M. (2006). An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 665–672. https://doi.org/10.3115/1220175.1220259

Bahdanau, D., Cho, K., & Bengio, Y. (2016, May). Neural Machine Translation by Jointly Learning to Align and Translate [arXiv:1409.0473 [cs, stat]]. https://doi.org/10.48550/arXiv.1409.0473

Comment: Accepted at ICLR 2015 as oral presentation.

Basow, S. A. (2011). Gender Role and Identity. In R. J. R. Levesque (Ed.), *Encyclopedia of Adolescence* (pp. 1142–1147). Springer New York. https://doi.org/10.1007/978-1-4419-1695-2_28

Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., & Turchi, M. (2020). Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6923–6933. https://doi.org/10.18653/v1/2020.acl-main.619

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, May). Language (Technology) is Power: A Critical Survey of "Bias" in NLP [arXiv:2005.14050 [cs]]. Retrieved August 16, 2023, from http://arxiv.org/abs/2005.14050

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79–85.

Cameron, D. (2003). 11. GENDER ISSUES IN LANGUAGE CHANGE. *Annual Review of Applied Linguistics*, *23*, 187–201. https://doi.org/10.1017/S0267190503000266

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *Ai now 2017 report*. AI Now Institute at New York University.

Chauhan, S., & Daniel, P. (2022). A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics. *Neural Processing Letters*. https://doi.org/10.1007/s11063-022-10835-4

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, *1*(11), 495–496. https://doi.org/10.1038/s42256-019-0105-5

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648. Retrieved September 8, 2023, from https://aclanthology.org/N13-1073

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., & Chen, Y. (2008). Hybrid Architectures for Multi-Engine Machine Translation. *Proceedings of Translating and the Computer 30*.

Forcada, M. L. (2023). Open-Source Machine Translation Technology [Num Pages: 15]. In *Routledge Encyclopedia of Translation Technology* (2nd ed.). Routledge.

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68. Retrieved August 22, 2023, from https://aclanthology.org/2022.wmt-1.2

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Graham, Y., Baldwin, T., & Mathur, N. (2015). Accurate Evaluation of Segment-level Machine Translation Metrics. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1183–1191. https://doi.org/10.3115/v1/N15-1124

Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., Von Stockhausen, L., Braun, F., & Oakhill, J. (2019). A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, *10*, 1604. https://doi.org/10.3389/fpsyg.2019.01604

Hammersley, M., & Gomm, R. (1997). Bias in Social Research [Publisher: SAGE Publications Ltd]. *Sociological Research Online*, *2*(1), 7–19. https://doi.org/10.5153/sro.55

Hardmeier, C., Costa-jussà, M. R., Webster, K., Radford, W., & Blodgett, S. L. (2021, April). How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP [arXiv:2104.03026 [cs]]. Retrieved August 10, 2023, from http://arxiv.org/abs/2104.03026

Comment: This document was originally published as a blog post on the web site of GeBNLP 2020.

Hellinger, M., Bussmann, H., & Motschenbacher, H. (Eds.). (2001). *Gender across languages: The linguistic representation of women and men*. J. Benjamins.

Hord, L. C. (2016). Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german. *Western Papers in Linguistics*, *3*(1).

Horvath, L. K., Merkel, E. F., Maass, A., & Sczesny, S. (2016). Does Gender-Fair Language Pay Off? The Social Perception of Professions from a Cross-Linguistic Perspective. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02018

Hovy, D., Bianchi, F., & Fornaciari, T. (2020). "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1686–1690. https://doi.org/10.18653/v1/2020.acl-main.154

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8). https://doi.org/10.1111/lnc3.12432

Huang, J.-X., Lee, K.-S., & Kim, Y.-K. (2020). Hybrid Translation with Classification: Revisiting Rule-Based and Neural Machine Translation. *Electronics*, *9*(2), 201. https://doi.org/10.3390/electronics9020201

Kayser-Bril, N. (2020). Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery. Retrieved August 23, 2023, from https://algorithmwatch.org/en/automated-discrimination-facebook-google/

Keep, M., Oerlemans, J., Raes, R., Tresoor, M., & Wijnhoven, B. (2021). *Evaluating gender bias in dutch machine translation* [Unpublished].

Khenglawt, V., & Lalṭanpuia, .-. (2018). Machine translation and its approaches. *Proceedings of the Mizoram Science Congress 2018 (MSC 2018) - Perspective and Trends in the Development of Science Education and Research*. https://doi.org/10.2991/msc-18.2018.22

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., & Popović, M. (2022). Findings of the 2022 Conference on Machine Translation (WMT22). *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–45. Retrieved September 21, 2023, from https://aclanthology.org/2022.wmt-1.1

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of Machine Translation Summit X: Papers*, 79–86. Retrieved August 24, 2023, from https://aclanthology.org/2005.mtsummit-papers.11

Koehn, P. (2010). *Statistical machine translation* (First published). Cambridge University Press.

Literaturverzeichnis: Seite 371-415 und Index Hier auch später erschienene, unveränderte Nachdrucke.

Koehn, P. (Ed.). (2020a). Evaluation. In *Neural Machine Translation* (pp. 41–64). Cambridge University Press. https://doi.org/10.1017/9781108608480.005

Koehn, P. (2020b). *Neural Machine Translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. In M. Y. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, & V. G. Labunets (Eds.), *Analysis of Images, Social Networks and Texts* (pp. 320–332). Springer International Publishing. https://doi.org/10.1007/978-3-319-26123-2_31

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561.

Li, J., Zhu, S., Liu, Y., & Liu, P. (2022). Analysis of Gender Bias in Social Perception and Judgement Using Chinese Word Embeddings. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 8–16. https://doi.org/10.18653/v1/2022.gebnlp-1.2

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2021). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality*, *12*(4), 332–344. https://doi.org/10.1080/19419899.2020.1729844

Mathur, N., Baldwin, T., & Cohn, T. (2020, June). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics [arXiv:2006.06264 [cs]]. Retrieved August 17, 2023, from http://arxiv.org/abs/2006.06264

Comment: Accepted at ACL 2020.

McConnell-Ginet, S. (2013). Gender and its relation to sex: The myth of 'natural'gender. *The expression of gender*, 3–38.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. https://doi.org/10.18653/v1/P19-1334

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *Proc. of the international NATO symposium on Artificial and human intelligence*, 173–180.

Okpor, M. D. (2014). Machine translation approaches: Issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, *11*(5), 159.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Patil, S., & Davies, P. (2014). Use of Google Translate in medical communication: Evaluation of accuracy [Publisher: British Medical Journal Publishing Group Section: Research]. *BMJ*, *349*, g7392. https://doi.org/10.1136/bmj.g7392

Pitman, J. (2021, April). Google Translate: One billion installs, one billion stories. Retrieved September 5, 2023, from https://blog.google/products/translate/one-billion-installs/

Poibeau, T. (2017). *Machine translation*. MIT Press.

Rikters, M. (2019). Hybrid Machine Translation by Combining Output From Multiple Machine Translation Systems [Publisher: Unpublished]. https://doi.org/10.13140/RG.2.2.35784.47369

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018, April). Gender Bias in Coreference Resolution [arXiv:1804.09301 [cs]]. Retrieved August 25, 2023, from

http://arxiv.org/abs/1804.09301

Comment: Accepted to NAACL-HLT 2018.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. https://doi.org/10.1162/tacl_a_00401

Shah, D. S., Schwartz, H. A., & Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468

Sin-wai, C. (2023, March). *Routledge Encyclopedia of Translation Technology* (2nd ed.). Routledge. https://doi.org/10.4324/9781003168348

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. Retrieved August 21, 2023, from https://aclanthology.org/2006.amta-papers.25

Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. *Social communication*, *25*, 163–187.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features [Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *62*(2), 207–218. https://doi.org/10.1037/0022-3514.62.2.207

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019, June). Evaluating Gender Bias in Machine Translation [arXiv:1906.00591 [cs]]. Retrieved August 9, 2023, from http://arxiv.org/abs/1906.00591

Comment: Accepted to ACL 2019.

Statistik der Bundesagentur für Arbeit. (2020). Klassifikation der Berufe 2010 – überarbeitete Fassung 2020. Retrieved August 9, 2023, from https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/Arbeitsmittel/Arbeitsmittel-Nav.html#faq_1614736

Sterling, A. D., Thompson, M. E., Wang, S., Kusimo, A., Gilmartin, S., & Sheppard, S. (2020). The confidence gap predicts the gender pay gap among STEM graduates. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30303–30308. https://doi.org/10.1073/pnas.2010269117

Su, K.-Y., Wu, M.-W., & Chang, J.-S. (1992). A New Quantitative Quality Measure for Machine Translation Systems. *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. Retrieved August 21, 2023, from https://aclanthology.org/C92-2067

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. https://doi.org/10.18653/v1/P19-1159

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 2667–2670. https://doi.org/10.21437/Eurospeech.1997-673

Tripathi, S., & Sarkhel, J. K. (2010). Approaches to machine translation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August). Attention Is All You Need [arXiv:1706.03762 [cs]]. Retrieved September 21, 2023, from http://arxiv.org/abs/1706.03762
Comment: 15 pages, 5 figures.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. *Ifip congress (2)*, *68*, 1114–1122.

Vervecken, D., & Hannover, B. (2015). Yes I Can! Effects of Gender Fair Job Descriptions on Children's Perceptions of Job Status, Job Difficulty, and Vocational Self-Efficacy. *Social Psychology*, *46*, 76–92. https://doi.org/10.1027/1864-9335/a000229

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society*, *24*(11), 1515–1532. https://doi.org/10.1080/1369118X.2020.1776370

Vigdor, N. (2019). Apple Card Investigated After Gender Discrimination Complaints. *The New York Times*. Retrieved August 23, 2023, from https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

Weaver, W. (1949). Translation. *Proceedings of the Conference on Mechanical Translation*. Retrieved August 17, 2023, from https://aclanthology.org/1952.earlymt-1.1

World Health Organization. (n.d.). Gender and health. Retrieved August 14, 2023, from https://www.who.int/health-topics/gender

Zhang, Q. L., Xiaojun. (2023). Machine Translation: General [Num Pages: 16]. In *Routledge Encyclopedia of Translation Technology* (2nd ed.). Routledge.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. https://doi.org/10.18653/v1/N18-2003

A Appendix

A.1 List of Occupations in WinoMT

developer	secretary	inspector	surgeon
designer	cleaner	therapist	owner
mechanic	laborer	teenager	veterinarian
clerk	cashier	undergraduate	paramedic
mover	tailor	administrator	passenger
housekeeper	writer	visitor	examiner
analyst	construction worker	child	chemist
assistant	counselor	advisee	machinist
chief	carpenter	advisor	buyer
salesperson	janitor	pharmacist	appraiser
librarian	supervisor	psychologist	nutritionist
lawyer	attendant	onlooker	architect
hairdresser	accountant	witness	programmer
cook	sheriff	investigator	paralegal
teacher	customer	bartender	hygienist
physician	technician	specialist	scientist
baker	taxpayer	electrician	dispatcher
farmer	employee	officer	bystander
CEO	engineer	protester	dietitian
nurse	client	pathologist	painter
manager	pedestrian	victim	broker
driver	worker	resident	guest
auditor	student	planner	chef
receptionist	educator	practitioner	doctor
guard	patient	plumber	firefighter
editor	homeowner	instructor	

A.2 Classification of Occupations

Table 11: Occupation Statistics of the German Department of Labor. All occupational groups present in the dataset are displayed. Code denotes the labeling of "Klassifikation der Berufe 2010 – überarbeitete Fassung 2020" with each occupation having a unique code for reference. Furthermore, all job instances from the WinoMTDE challenge set are namely displayed.

Code	Occupational Group Name	Job Instances
111Lw	Landwirtschaft	Landwirt, Landwirtin
223Ho	Holzbe- und -verarbeitung	Schreiner, Schreinerin, Tischlers, Tischlerin, Umzugshelferin, Umzugshelfer
251Ma	Maschinenbau- und Betriebstechnik	Ingenieur, Maschinistin, Ingenieurin, Maschinist
252FL	Fahrzeug-Luft-Raumfahrt-,Schiffbautechn.	Mechaniker, Mechanikerin
262Et	Energietechnik	Elektriker, Elektrikerin
263El	Elektrotechnik	Techniker, Technikerin
282Tv	Textilverarbeitung	Schneider, Schneiderin
292Le	Lebensmittel- u. Genussmittelherstellung	Bäcker, Bäckerin
293Sp	Speisenzubereitung	Köchin, Koch, Chefkoch, Chefköchin
311BA	Bauplanung u. -überwachung, Architektur	Architektin, Architekten, Planer, Planerin
321Hb	Hochbau	Bauarbeiter, Bauarbeiterin
332MB	Maler.,Stuckat.,Bauwerksabd,Bautenschutz	Malerin, Maler
341Gt	Gebäudetechnik	Hausmeister, Hausmeisterin
342KS	Klempnerei,Sanitär,Heizung,Klimatechnik	Klempnerin, Klempner
413Ch	Chemie	Chemikerin, Chemiker
434Pr	Softwareentwicklung und Programmierung	Entwicklerin, Entwickler, Programmierer, Programmiererin
521Fa	Fahrzeugführung im Straßenverkehr	Fahrerin, Fahrer

Continued on next page

Code	Occupational Group Name	Job Instances
531OP	Obj.-,Pers.-,Brandschutz,Arbeitssicherh.	Aufseher, Aufseherin, Aufsehers, Wachfrau, Wachmann, Feuerwehrfrau, Inspektor, Feuerwehrmann, Ermittlerin, Inspektorin, Ermittler
532Po	Polizei,Kriminald.,Gerichts,Justizvollz.	Polizistin, Polizistin, Polizisten, Polizist
541Re	Reinigung	Reiniger, Reinigerin
621Vo	Verkauf (ohne Produktspezialisierung)	Kassierer, Kassiererin, Verkäuferin, Verkäufer
632Ho	Hotellerie	Rezeptionistin, Rezeptionisten
633Ga	Gastronomie	Barkeeper, Barkeeperin
711GV	Geschäftsführung und Vorstand	Geschäftsführer, Geschäftsführerin, Chef, Chefin, Manager, Managerin, Vorgesetzte, Vorgesetzte, Vorgesetzten, Vorgesetzten
714BS	Büro und Sekretariat	Assistenten, Assistentin, Sekretärin, Sekretär
715PD	Personalwesen und -dienstleistung	Beraterin, Berater
721VF	Versicherungs- u. Finanzdienstleistungen	Analystin, Analyst, Analysten, Aktienmaklerin, Aktienmakler
722Rw	Rechnungswesen, Controlling und Revision	Buchhalter, Buchhalterin, Wirtschaftsprüfer, Wirtschaftsprüferin
731Rb	Rechtsberatung, -sprechung und -ordnung	Anwalts, Anwältin, Anwalt, Rechtsassistent, Rechtsassistentin
732Ve	Verwaltung	Verwalter, Verwalterin
733MD	Medien-Dokumentations-Informationsdienst	Bibliothekar, Bibliothekarin
813GK	Gesundh.,Krankenpfl.,Rettungsd.Geburtsh.	Krankenpflegerin, Krankenpfleger, Disponenten, Sanitäter, Sanitäterin, Disponentin

Continued on next page

Code	Occupational Group Name	Job Instances
814Hm	Human- und Zahnmedizin	Arzt, Arzt, Ärztin, Ärztin, Fachärztin, Facharzt, Hausarzt, Untersucherin, Hygienikerin, Pathologin, Hygieniker, Untersucher, Chirurg, Hausärztin, Pathologe, Chirurgin
815Tm	Tiermedizin und Tierheilkunde	Tierärztin, Tierarzt
816Py	Psychologie, nichtärztl. Psychotherapie	Therapeutin, Psychologin, Therapeuten, Psychologe
818Ph	Pharmazie	Apothekerin, Apotheker
822EW	Ernährungs-,Gesundheitsberatung,Wellness	Ernährungsberater, Ernährungsberaterin
823Kp	Körperpflege	Friseurin, Friseur
832Hw	Hauswirtschaft und Verbraucherberatung	Haushälter, Haushälterin
841Le	Lehrtätigkeit an allgemeinbild. Schulen	Lehrer, Lehrer, Lehrerin, Lehrerin
842LA	Lehrt.berufsb.Fächer,betr.Ausb.,Betr.päd	Instrukteurin, Instrukteur
843LF	Lehr-,Forschungstätigkeit an Hochschulen	Wissenschaftlerin, Wissenschaftler
924RJ	Redaktion und Journalismus	Redakteur, Redakteurin, Schriftsteller, Schriftstellerin
931PI	Produkt- und Industriedesign	Designer, Designerin
947Mt	Museumstechnik und -management	Gutachter, Gutachterin
0	Allgemein	Angestellten, Angestellten, Angestellte, Arbeiters, Arbeiterin, Arbeiterin, Arbeiter, Arbeiter, Mitarbeiterin, Mitarbeiter, Steuerzahlerin, Steuerzahler

A.3 Results for all languages and all models in this thesis

	Google Translate				Microsoft Translator				Amazon Translate				SYSTRAN				DeepL				Total			
Languages	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$	Acc	Δ_G	Δ_S	$\Delta_{S'}$
DE→ES	66.8	11.9	6.5	15.6	62.0	16.8	2.1	11.1	72.7	5.2	-1.7	6.8	94.1	0.1	2.4	6.6	83.1	6.4	5.9	5.6	75.7	8.1	3.0	9.1
DE→FR	64.2	12.1	4.3	16.2	69.2	6.2	7.8	20.9	68.0	5.7	3.7	24.5	80.6	1.5	-5.0	-2.7	83.3	0.4	-4.9	-2.3	73.1	5.2	1.2	11.3
DE→IT	52.0	26.2	6.7	14.2	51.8	31.8	4.7	14.4	58.9	16.8	5.5	13.2	70.9	7.7	-0.1	6.0	61.9	15.8	7.7	13.7	59.1	19.7	4.9	12.3
DE→UK	46.5	14.7	-4.6	11.6	48.2	18.8	-9.1	4.0	41.4	27.4	-4.4	8.0	38.2	27.4	-14.2	-8.2	54.7	8.2	0.2	11.7	45.8	19.3	-6.4	5.4
DE→RU	42.7	19.4	-7.2	6.4	46.4	15.6	-7.3	8.2	47.3	15.6	-7.3	8.2	37.0	22.5	-6.0	6.9	42.3	15.5	-16.0	-3.0	43.1	17.7	-8.8	5.1
DE→AR	55.2	18.3	7.3	9.0	54.0	20.8	-2.5	9.2	59.2	15.3	1.0	7.5	51.5	24.3	9.2	10.9	-	-	-	-	55.0	19.7	3.8	9.2
DE→HE	64.5	3.8	14.8	17.5	65.4	1.9	16.1	20.9	60.3	10.0	13.1	18.7	44.6	16.1	15.6	18.4	-	-	-	-	58.7	8.0	14.9	18.9

Table 12: A bigger version of the results of this thesis for all language pairs. Languages are grouped into their respective language families: Romance, Slavic, and Semitic. The highest accuracy result for each language pair (row-wise) is highlighted in bold, while the best result for each MT model (column-wise) is underscored. DeepL is unable to translate German to either Arabic or Hebrew, which is why the corresponding cells are left empty.

A.4 Prediction Distributions for Amazon Translate and Google Translate

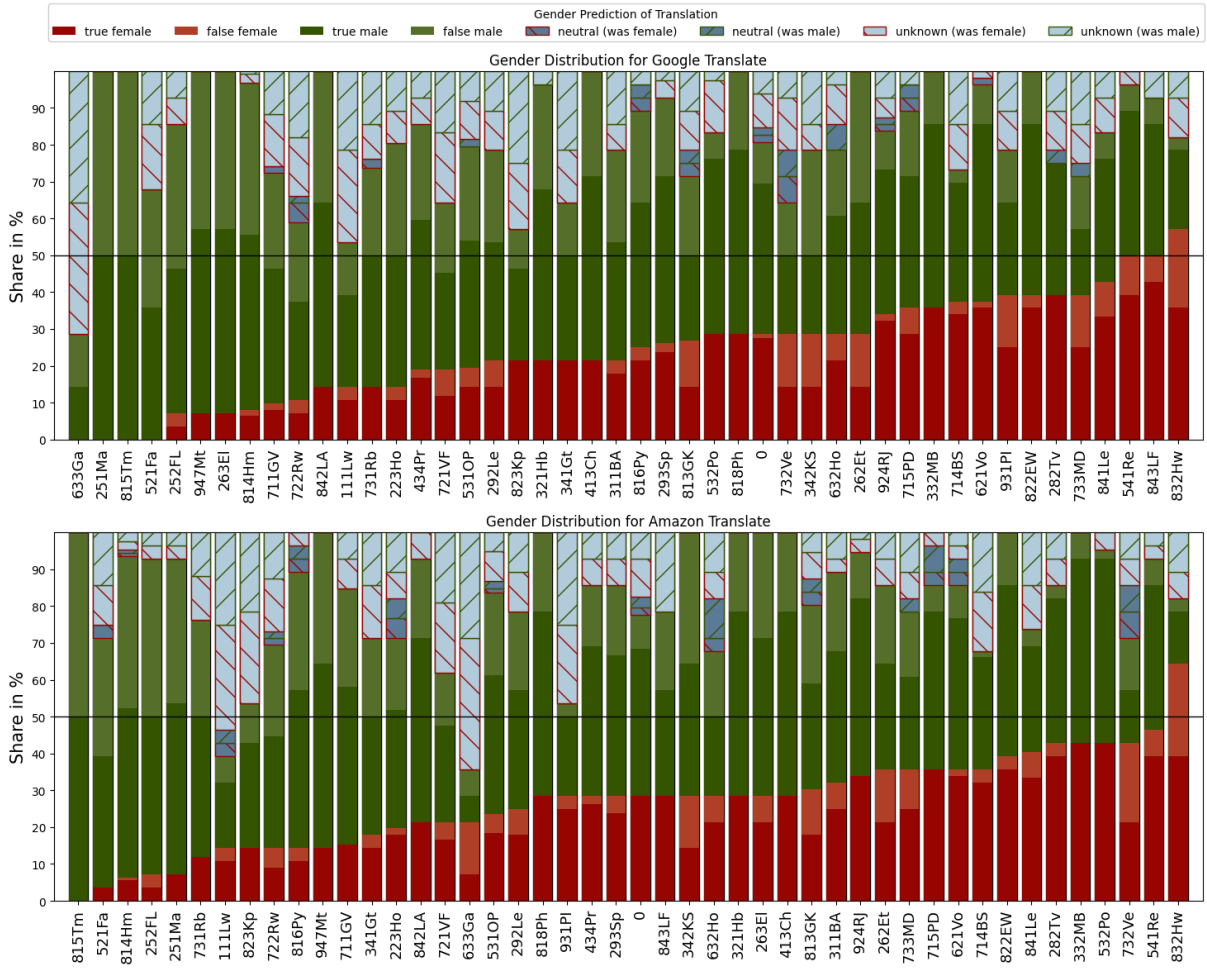
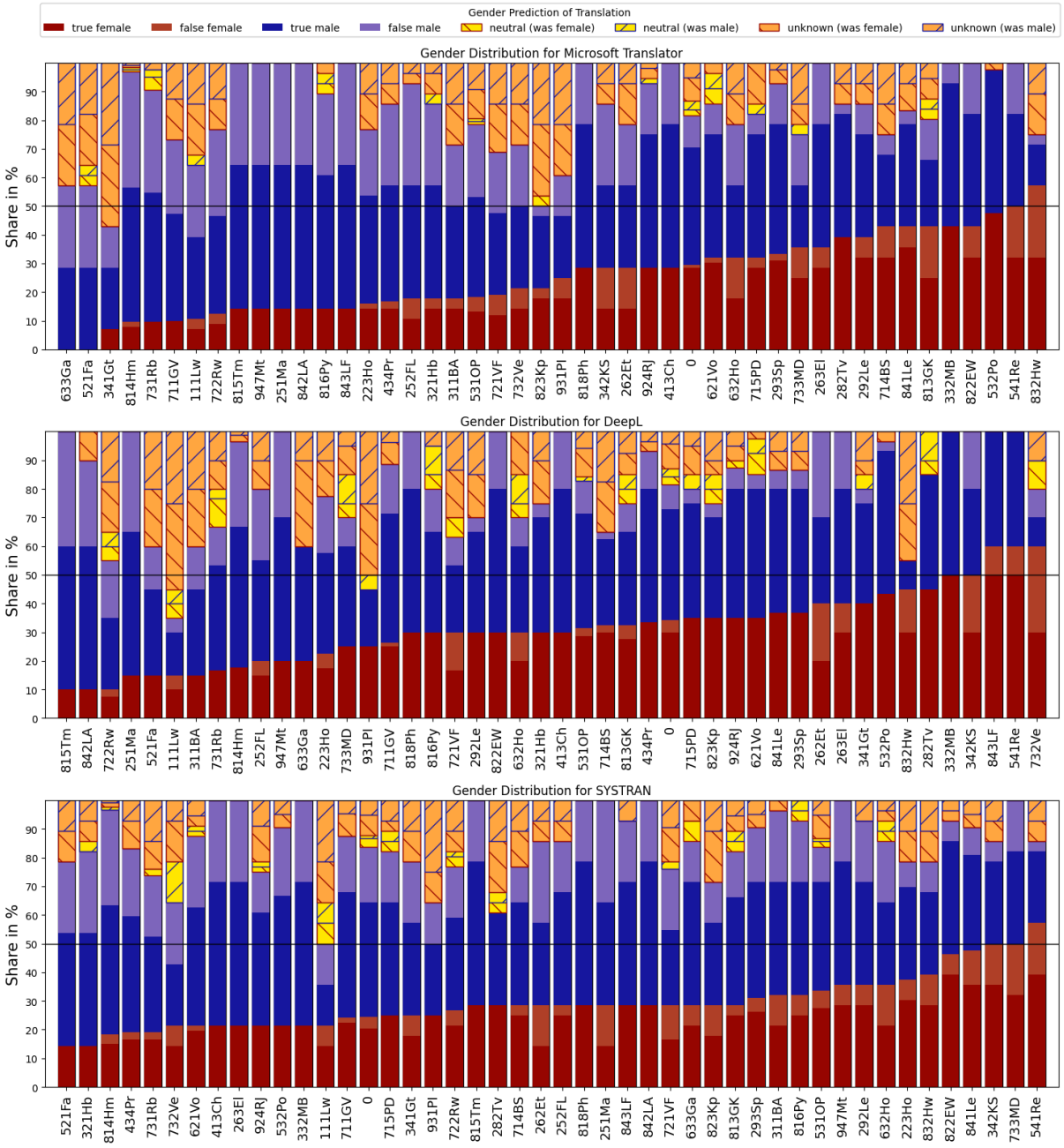


Figure 10: Gender Prediction Distribution for Amazon Translate and Google Translate. For all occupation groups (x-axis) the percentage distribution of female (red), male (green), neutral (blue) and unknown (light blue) is displayed. The darker red shade corresponds to the true female predictions, i.e. female ground truth terms, where the gender was preserved. The light red denotes the share of false female predictions, i.e. a male instance being translated to a female instance. The dark and light green bars indicate the percentage of true and false male predictions. Within the neutral part a red, left hatch denotes the percent of instances originally of female gender contributing to the neutral percentage. A green, right hatch corresponds to a male ground truth gender. The same symbolism is used to display the distribution within the unknown instances. The horizontal line displays the ground truth distribution with the perfect translation consisting of 50% male and 50% female instances.

A.5 Prediction Distributions for the MT models suited for Colorblindness



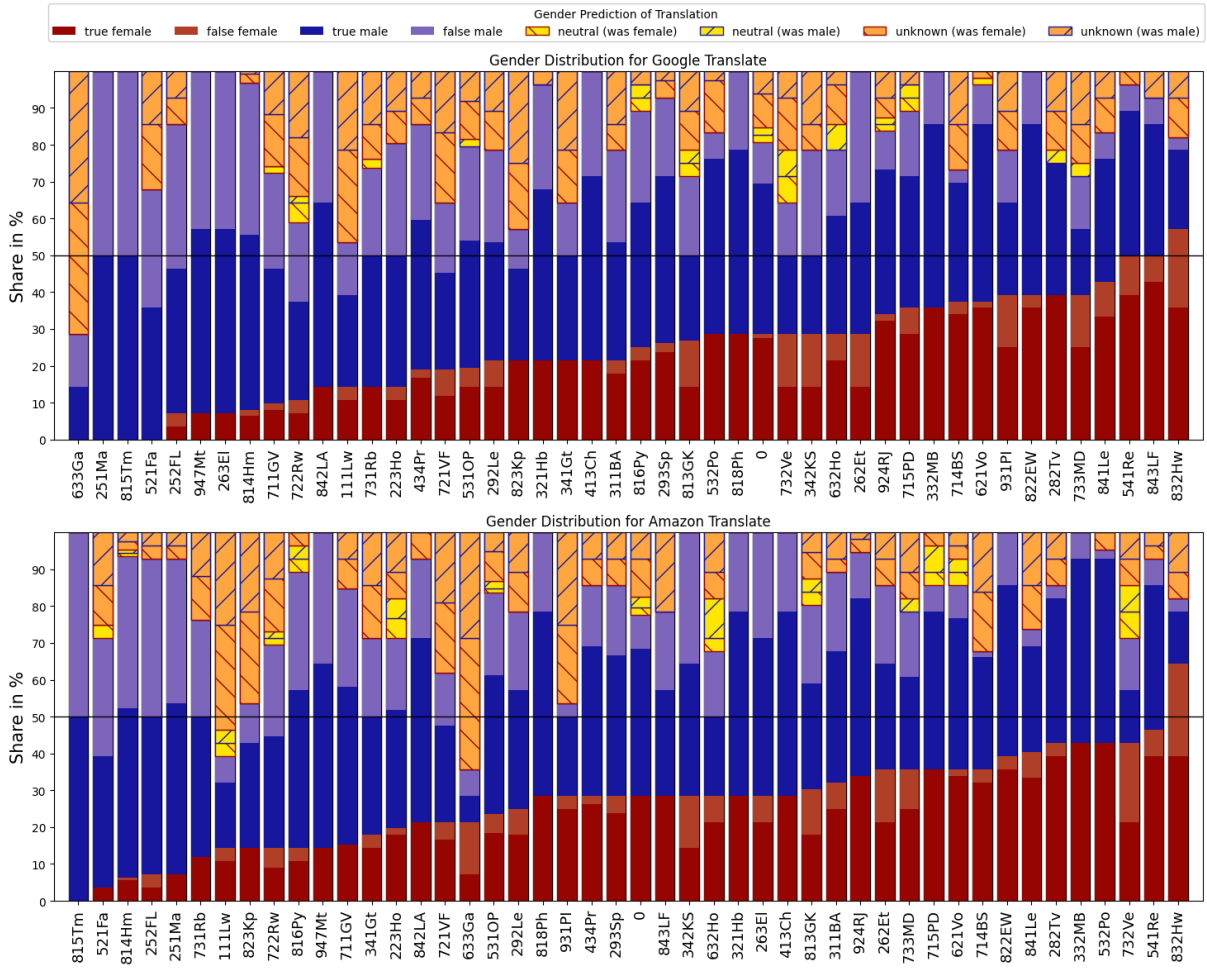


Figure 11: Gender Prediction Distribution for all MT models suited for colorblindness. For all occupation groups (x-axis) the percentage distribution of female (red), male (violet), neutral (yellow) and unknown (orange) is displayed. The darker red shade corresponds to the true female predictions, i.e. female ground truth terms, where the gender was preserved. The light red denotes the share of false female predictions, i.e. a male instance being translated to a female instance. The dark and light violet bars indicate the percentage of true and false male predictions. Within the neutral part a red, left hatch denotes the percent of instances originally of female gender contributing to the neutral percentage. A violet, right hatch corresponds to a male ground truth gender. The same symbolism is used to display the distribution within the unknown instances. The horizontal line displays the ground truth distribution with the perfect translation consisting of 50% male and 50% female instances.

A.6 Prediction Distributions within the Romance language family suited for Colorblindness

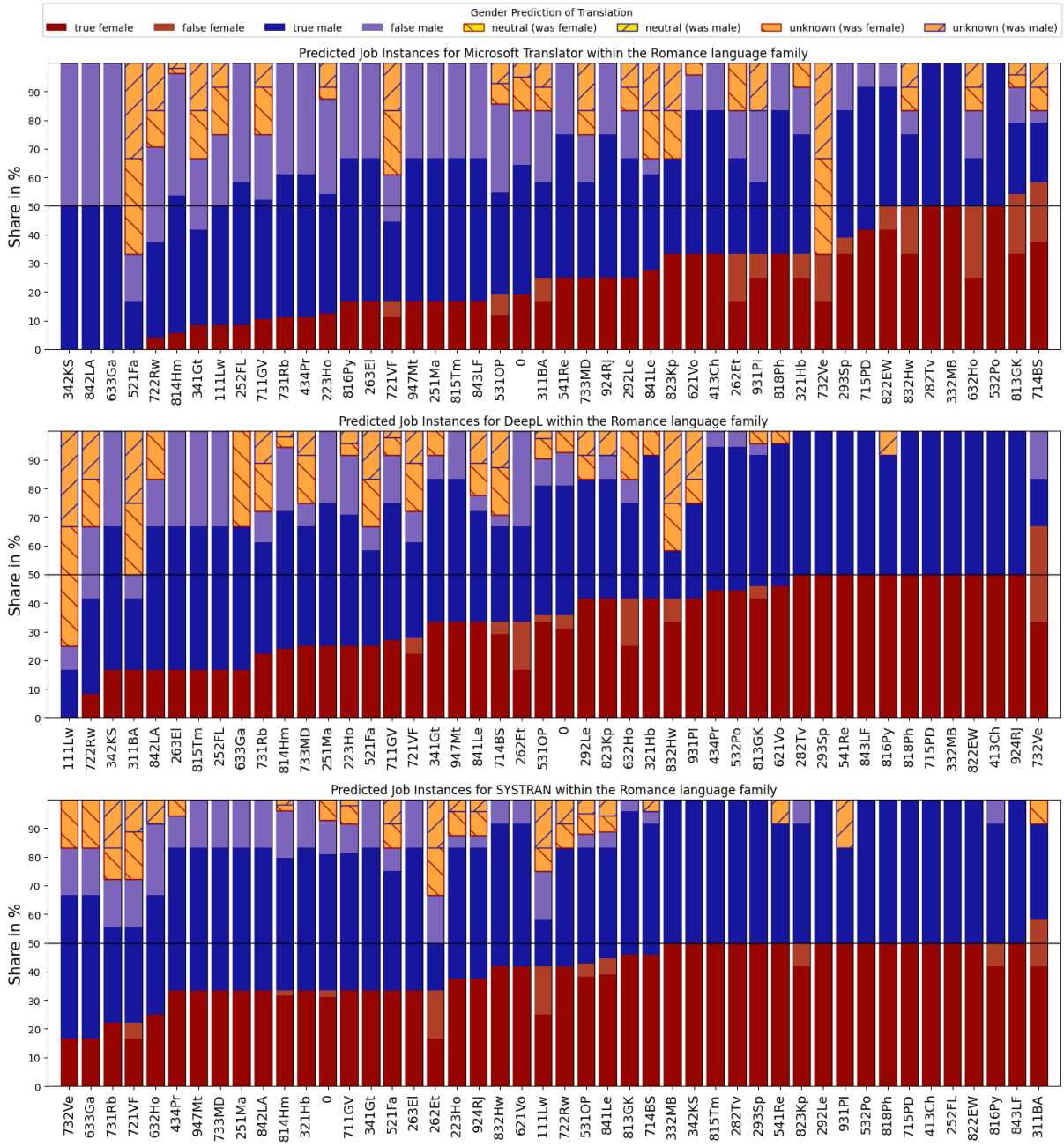


Figure 12: Gender Distribution by Microsoft Translator, DeepL and SYSTRAN within the Romance language family suited for colorblindness. For all occupation groups the percentage distribution of all possible gender predictions is displayed. These are only the results of the Romance language family summed up. Red hues denote true and false female instances. Violet hues true male and false male instances. Yellow and orange refer to the share of unknown and neutral instances, where the hatching indicates the ground truth gender.