Master Thesis

“It Wasn’t *He*”: Analyzing Gender Bias in Multilingual Machine Translation

Lena Cabrera Pérez

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Artificial Intelligence
at the Department of Advanced Computing Sciences
of the Maastricht University

Thesis Committee:

Dr. Gerasimos Spanakis
Dr. Nava Tintarev

External supervisor:
Prof. Dr. Jan Niehues
(Karlsruhe Institute of Technology)

Maastricht University
Faculty of Science and Engineering
Department of Advanced Computing Sciences

April 17, 2023

Abstract

Neural machine translation (NMT) models often suffer from gender biases that harm users and society at large. In recent years, several debiasing approaches have been proposed, including debiasing the data before model training, the models during training, or post-processing their outputs; however, it has yet to be explored how limited availability of large amounts of data affects gender bias in multilingual NMT, specifically for zero-shot directions. In this thesis, we compare pivot-based and zero-shot translation and study the influence of the bridge language—the language participating in all language pairs during training—and the effect of language-agnostic hidden representations—proven to achieve zero-shot performance gains—on models' ability to preserve the feminine and masculine gender equally well in their outputs. We test three bridge languages with differently pronounced gender-inflectional systems—English, German, and Spanish— as well as three different model modifications that encourage language-agnostic representations: *i*) removing a residual connection in a middle Transformer encoder to lift positional correspondences to the input tokens; *ii*) promoting similar source and target language representations through an auxiliary loss; and *iii*) joint adversarial training penalizing successful recovery of source language signals from the hidden representations. We find that language-agnostic representations reduce the gap between better masculine and worse feminine results (i.e., masculine bias) found throughout all models' outputs; they improve zero-shot models' performances to the point where they can outperform pivoting when bridging via English, a low gender-inflected language. With increased levels of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairer gender preservation when gender inflection depends on the speaker's gender; otherwise zero-shot translation performs slightly better regarding balanced gender preservation for feminine and masculine words.

Contents

List of Figures

List of Tables

List of Abbreviations

AI artificial intelligence.

ANN artificial neural network.

BPE byte pair encoding.

CE cross entropy.

CLM conditional language model.

LM language model.

LSTM long short-term memory.

ML machine learning.

MNMT multilingual neural machine translation.

MT machine translation.

NLP natural language processing.

NMT neural machine translation.

RBMT rule-based machine translation.

RNN recurrent neural network.

SGD stochastic gradient descent.

SMT statistical machine translation.

1

Introduction

Machine translation (MT) is the process of using artificial intelligence (AI) to automatically translate text from one language to another without human involvement. As such, it is a powerful instrument for communicating and exchanging information across linguistic barriers; it reduces both cost and time, enabling the translation of millions of words almost instantaneously, often supporting up to a hundred languages. Modern MT goes beyond simple word-to-word translation to accurately convey the original meaning of the source text in the target language. Most importantly, the rise of artificial neural networks within the field of AI has paved the way for the now prevailing approach to MT, termed neural machine translation (NMT). The great success of NMT is based on its "end-to-end" learning approach, having a single model learn all the steps between the initial input phase and the final output result, including sentence representation, determining semantically correct word translations, and generating grammatically correct and fluent target language sentences.

Contemporary NMT systems, which typically support translation between multiple languages, are trained on large amounts of parallel data consisting of pairs of sentences and their translations. Increased amounts of training data generally result in improved translation quality. Unfortunately, many languages lack parallel digital data, and are thus referred to as resource-poor languages. Building NMT systems for these languages is a great challenge due to the data sparseness problem, the phenomenon of not observing enough data, if any, to model language accurately. A promising line of research addressing this issue is to leverage resource-rich languages, such as English, in order to improve the translation of resource-poor languages. In the extreme case, there are no resources available, making the ability to translate between language pairs unseen during training, i.e., zero-shot translation, an important aspect in multilingual neural machine translation (MNMT). There are two prevalent approaches to the idea of leveraging available language resources to perform translation for unknown translation directions: *pivot-based* translation through explicit bridging and *zero-shot* translation performing implicitly-learned bridging.

In pivot-based translation, a high-resource language is used as a pivot language to build an explicit bridge between low-resource language pairs. Specifically, pivoting is performed by first translating to a common language, like English, and then translating from the common language into the desired target language. Meanwhile, zero-shot translation refers to direct translation between language pairs never explicitly seen by the system during training. The idea behind this approach is that an end-to-end system can learn a shared semantic representation for all languages it is trained on that subsequently can be used as an implicit bridge to perform translation between new language pairs. The advantage of zero-shot translation over pivoting is that the direct translation circumvents error propagation, reduces computation time, and is easily extendable to support a growing number of languages. However, achieving high-quality zero-shot translation is still a challenging task.

1.1 Motivation

As we employ a growing number of machine-based solutions relying on sophisticated machine learning (ML) algorithms, we see human problems occurring within them. Arguably one of the biggest concerns at present is the societal impact of bias and discrimination, which have appeared repeatedly in various technological applications. One of many examples is the bias found in numerous of the MT applications in use today, such as Google Translate, which alone serves millions of people daily. With the growing impact of translation technologies on our everyday lives, fears of machine bias and discrimination have also increased within and beyond the academic community.

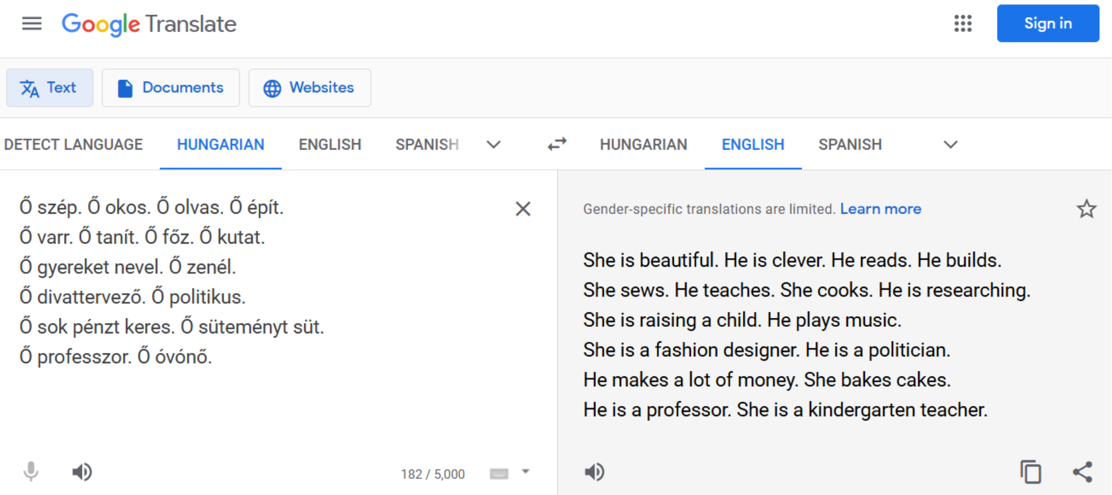
In 2021, a Twitter user deemed Google Translate's Hungarian-English translations as a portrayal of "everyday sexism" (Vargha, 2021). A screenshot attached to her tweet[1] showed that Hungarian sentences – lacking any gender specification as Hungarian is a genderless language with no gendered pronouns – were translated into their English counterparts by seemingly reflecting stereotypical gender roles. The Twitter user's observation, preceding a series of similar issues reported by other users for different languages, reveals the bias in multilingual MT systems and serves as a prime example of the phenomenon of bias in ML-based decision-making at large. While technology is often perceived by the public to be untouched by human bias, the reality is that the technologies shaping our world are deeply shaped by the humans who create them. In the case of ML, we are feeding historical data to have a model learn how to derive decisions in the future. While this data-driven approach can, in some cases, already reach human parity, we must acknowledge that patterns of humans' past discriminatory decision-making practices are built into these models, inevitably producing biased systems.

Gender bias, in particular, is linked to the prevalence of gender stereotyping, namely the preconception about attributes, characteristics, or the roles that are or ought to be possessed by, or performed by, women and men. An example for stereotyping is the association between the male gender and prestigious STEM professions, referring to careers related to science, technology, engineering, and mathematics. In the now infamous case of Google Translate, MT has been proven to hallucinate the gender of a person, based on stereotypes, when the source sentence does not provide any gender information, as illustrated in the Hungarian-English translations in Figure 1.1 (e.g., the gender-neutral sentence "Ő professzor" being translated to "*He* is a professor").

These stereotypical biases are harmful since they perpetuate and perhaps even amplify inequalities. For instance, given that a profession appears in a masculine context, i.e., associates with a male 70% of the times it is mentioned in texts used to train an MT system, the system may recognize and follow this pattern or even reflect it 100% of the times as male. As a result, women in that profession are under-represented, or worse, are essentially "wiped out of existence" altogether, revealing the far-reaching nature of the problem.

The phenomenon of machine bias does not exist due to active discrimination, as the bias is largely unconscious. In the case of Google Translate, the algorithm defaults to the masculine pronoun disproportionately more often, presumably because "he" more commonly appears in available utilized training resources, e.g., texts on the web, than "she"; despite half of the population being female. This example highlights the root cause of the problem of gender-biased MT:

[1] https://twitter.com/DoraVargha/status/1373211762108076034



Ő szép. Ő okos. Ő olvas. Ő épít.
Ő varr. Ő tanít. Ő főz. Ő kutat.
Ő gyereket nevel. Ő zenél.
Ő divattervező. Ő politikus.
Ő sok pénzt keres. Ő süteményt süt.
Ő professzor. Ő óvónő.

She is beautiful. He is clever. He reads. He builds.
She sews. He teaches. She cooks. He is researching.
She is raising a child. He plays music.
She is a fashion designer. He is a politician.
He makes a lot of money. She bakes cakes.
He is a professor. She is a kindergarten teacher.

Figure 1.1: Translating sentences from a genderless language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in MT. The screenshot shows how Google Translate (as of August 2022) interprets occupations from traditionally male-dominated fields, such as "politician", as male while interpreting "kindergarten teacher" as female. Moreover, adjectives and verbs such as "beautiful" and "cooking" are translated with the female pronoun, whereas "clever" and "reading" are attributed to the male pronoun. Acknowledging the issue, Google points to limitations in gender-specific translations.

Whenever data is collected and analyzed without accounting for imbalanced representations of gender within the data, models reflecting or amplifying these imbalances in their outputs are the result.

Some people have argued that the main danger of it is that technology is often assumed to be neutral when, in fact, these systems are built under the guise of objectivity—for the least apparent bias is often the most dangerous. After all, we cannot prevent the consequences of bias or even acknowledge its impact if we are not aware of it. As tech practitioners and developers broaden and learn new techniques for data science, ML, and AI, a deep understanding of bias and how to identify and prevent it is vital to creating a world that is positively – rather than negatively – shaped by technology. The general goal of this thesis is to shed more light on the undesired phenomenon, as combating it will make translation technology more inclusive and less discriminatory.

1.2 Problem Statement

Despite the importance of making translation services fairer and more equitable, investigating gender bias in MT is still a relatively new field of inquiry. To facilitate the evaluation of models' gender biases, some of the earliest works have published gender-sensitive benchmarks, including WinoMT (Saunders et al., 2020; Stanovsky et al., 2019) and MuST-SHE (Bentivogli et al., 2020). Concerning the mitigation of gender bias, so far, most research has concentrated on manipulating the training data to reduce unfair gender treatment. Different strategies include the integration of additional information, for instance through gender tags (Vanmassenhove et al., 2018) or context information (Basta et al., 2020). Other works have examined the effect of debiasing the representations of words in the form of pre-trained word embeddings (Escudé Font & Costa-Jussa, 2019) or fine-tuning models with gender-balanced datasets featuring an equal amount of feminine and masculine references (Costa-jussà et al., 2020).

While existing works have presented initial and partially effective approaches to address the problem of gender-biased MT, a critical aspect of the problem has so far received little to no attention: Since multilingual translation systems, translating between multiple source and target languages, are today's de facto standard, it is vital to study gender bias in a *zero-shot* translation setting specifically. Translating between unknown language pairs creates a challenge for multilingual MT due to the morphological and structural diversity of different languages, including differences in gender inflection, which affect gender translation, particularly translation models' ability to *preserve gender*. It stands to reason that these linguistic aspects affect the emergence of gender biases in multilingual translation systems.

Altogether, the primary objective that practitioners and researchers ultimately aspire to accomplish is to *reduce and, wherever possible, eliminate gender bias* in *high-quality multilingual MT* while *expanding models to include new languages.* Motivated by this objective, in this thesis, we aim to investigate gender biases in zero-shot translation, an up-and-rising approach to achieve wide translation coverage, including many of the world's languages.

1.3 Research Questions

Our investigation into gender bias in multilingual MT is guided by three research questions. All questions address the different aspects of the research field's primary objective: achieving high-quality translations with minimized gender bias and extensible multi-language support. In terms of extensible multi-language support, we want to focus on the *zero-shot translation setting* and models capable of zero-shot translation specifically, which have achieved impressive results since they first emerged.

Despite their success, we acknowledge that high-quality zero-shot translations for a variety of different languages have yet to be accomplished. In light of this, regarding gender biases, we aim to compare zero-shot translation to pivoting, another multilingual translation approach widely used to overcome language resource limitations. In view of this, the first research question to be answered in this thesis is:

Research Question 1: How do zero-shot and pivot-based translation compare regarding gender-biased outputs in a zero-shot translation setting?

Given that the lexical encoding of gender relations and identities can vary from one language to another, it is unsurprising that multilingual MT relying on pivot languages bears the risk of losing gender information along the source-pivot-target translation pipeline. Zero-shot translation does not face this issue as it circumvents the explicit bridging step of pivoting, relying on implicitly-learned universal semantic representations of the source information. In light of this, we put forward the following hypothesis, which we test as part of our efforts to address the first research question:

On average, zero-shot translation generates fewer gender-biased outputs than pivot-based translation, where gender bias is conceived as the systematic and unfair discrimination against a group of individuals of the same sex, here either women or men, in favor of the other gender group, while maintaining comparable translation quality.

The results obtained during our investigation will either reject or fail to reject this central hypothesis. Along the way, we aim to study the influence of the training data and different model modifications on zero-shot and pivot-based translations of gender expressions.

Since languages express gender differently, in less and more obvious ways, we believe that the bridge language—the language that is present in all language pairs included during training—plays an important role in an MT model's ability to provide correct and unbiased translations of gender-specific words. We imagine training models using a highly gender-inflected bridge language, such as Spanish, yields better gender translations and reduces gender biases than a largely gender-neutral language, such as English, which is typically the most common language in training data. In light of this, the second research question is:

Research Question 2: Does the bridge language affect the gender biases perpetuated by zero-shot and pivot-based translations?

Besides studying the impact of the bridge language, we want to explore different model modifications to mitigate gender-biased model behavior. We aim to use methods that have previously shown to achieve zero-shot performance gains, to test whether translation quality and gender biases are negatively correlated. Accordingly, we formulate the third research question:

Research Question 3: Do translation quality improvements of zero-shot models reduce their gender biases?

1.4 Thesis Outline

Apart from this introductory chapter, the thesis comprises five additional Chapters. Chapter 2 establishes a theoretical base and provides the context for the coming chapters of research addressing the problem statement. The chapter introduces the task of MNMT at large and describes existing techniques to detect, assess, and mitigate gender bias in MT in detail. Chapter 3 describes the design of the experiments performed to answer the research questions; this includes the research procedure, techniques, and materials used to conduct the experiments and to assess the results. Chapter 4 outlines technical details required to ensure better comprehensibility and reproducibility of the experiments. Chapter 5 presents the results of the experiments and the discussions thereof. Chapter 6 summarizes the key findings of the experiments in answers to the research questions and concludes by pointing out avenues for potential future work.

2

Background & Related Work

This chapter sets the thesis in the context of existing literature related to the thesis topic. Section 2.1 reviews concepts and methods of NMT, the currently prevailing approach to MT. First, we introduce the basics of NMT according to the predictive pipeline: learning vocabularies and representing them using word embeddings, modeling translation using the encoder-decoder framework, model training, and inference. Moreover, we describe the state-of-the-art NMT model architecture and discuss approaches to multilingual NMT, including pivoting and zero-shot translation. Section 2.2 provides a brief overview of related works on the manifestation of gender bias in MT and its evaluation and mitigation.

2.1 Neural Machine Translation

MT is the task of translating natural language sentences using computers, a concept first put forward shortly after the development of the first programmable, general-purpose digital computer in the mid-1940s. Since then, MT has been considered one of the most challenging tasks in the field of natural language processing (NLP). Early approaches to MT relied heavily on hand-crafted translation rules and linguistic knowledge; however, since natural languages are inherently complex, covering all language irregularities with manual translation rules proved challenging.

With the availability of large-scale parallel corpora, corpus-based approaches that learn linguistic information and translation knowledge from data have gained increasing attention. Unlike rule-based machine translation (RBMT), statistical machine translation (SMT) (Brown et al., 1990; Koehn et al., 2003) learns intrinsic language patterns such as word alignments or grammar directly from parallel corpora. Therefore, SMT is less dependable on manual feature engineering and human linguistic labor than RBMT. However, incapable of modeling long-distance dependencies between words, the translation quality of SMT is far from satisfactory. With recent advances in deep learning—part of a broader family of machine learning methods, using artificial neural network (ANN)[1]—NMT (Bahdanau et al., 2014; K. Cho et al., 2014b; Kalchbrenner & Blunsom, 2013; Sutskever et al., 2014) emerged as a new paradigm and quickly replaced SMT as the mainstream approach to MT.

[1] A brief overview of the basic principles of deep learning can be found in the appendix A.1.

2.1.1 Vocabulary Building & Embeddings

As a data-driven approach, NMT starts with gathering a sufficiently large parallel text corpus, typically aligned on the sentence level, i.e., a sentence paired with its translation(s). Before training, the raw text corpus is preprocessed, starting with tokenization, during which sentences are split into smaller units called tokens, which can be either words, characters, or subwords. The discrete tokens are represented as vectors using continuous embedding: Mathematically, an embedding is a mapping from one set to another—it is continuous if the mapping is continuous. Embedding layers in neural networks use this concept to map a token into a continuous vector $\mathrm{e} \in \mathbb{R}^D$ of size D to carry out semantically founded numerical operations on the inputs. Embeddings are fed to the lower network layers for fine-grained feature extraction.

The set of unique tokens appearing in the text corpus is the vocabulary. In NMT, embedding vectors corresponding to the tokens in the vocabulary are typically gathered into matrices which are treated as a set of parameters of the translation model. Large models, i.e., neural networks with many parameters, are computationally expensive as they require large memory and small batch sizes, which leads to noisier gradients, slower training, and worse model performance. Since the size of the embedding matrices correlates with the vocabulary size, it is problematic to build vocabularies covering all words in a language; instead, it is desirable to avoid model inflation by reducing the vocabulary size to a moderate size.

Since a majority of the words rarely appear in the training corpus,[2] it is favorable to represent the most common words in the vocabulary as single tokens while breaking down rare words into two or more subword tokens. Common methods for subword tokenization are word piece models (Wu et al., 2016) and byte pair encoding (BPE) (Sennrich et al., 2016). BPE initializes the set of available subwords with the character set of the training data and extends it iteratively in subsequent merge operations. During a merge, two units with the most frequent co-occurrence in the text are combined. The operation is repeated until the desired pre-defined vocabulary size is reached. Following this procedure, the vocabulary is first built from the training data and then used to apply the same tokenization to the validation and test data.

2.1.2 Modeling Translation

Modeled at the sentence level, translation can be categorized as a *sequence-to-sequence* learning problem involving two sequences: a source language sentence $\mathrm{x} = (x_1, ..., x_n, ..., x_N)$ and a target language sentence $\mathrm{y} = (y_1, ..., y_t, ..., y_T)$. The lengths of these sequences can be different. In translation, the goal is to find the *most probable* target sequence y given input x; formally, the target sequence that maximizes the *conditional* probability $P(\mathrm{y}|\mathrm{x})$:

$$P(\mathrm{y}|\mathrm{x}) : \mathrm{y}* = argmax_y P(\mathrm{y}|\mathrm{x}). \tag{2.1}$$

NMT learns a function $P(\mathrm{y}|\mathrm{x}, \theta)$ with a set of parameters θ to find the translation y that maximizes the conditional probability for a given input x:

$$\mathrm{y}^* = argmax_y P(\mathrm{y}|\mathrm{x}, \theta). \tag{2.2}$$

In sequence-to-sequence learning, this is considered a conditional language model (CLM). A language model (LM) learns to estimate the unconditional probability $P(\mathrm{y})$ of a target sequence

[2] This can be derived from Zipf's law which, in the context of linguistics, states that the frequency of any word occurring in a corpus is inversely proportional to its rank in the frequency table. Thus, the most frequent word occurs approximately twice as often as the second most frequent word, three times as often as the third most frequent word, and so on.

of tokens y = $(y_1, y_2, ..., y_T)$, formally defined as:

$$P(\mathrm{y}) = \prod_{t=1}^{T} P(y_t | y_{<t}), \tag{2.3}$$

where $y_{<t}$ is the partial target sequence up until position $t{-}1$ $(y_1, y_2..., y_{t-1})$. Accordingly, each multiplication term is the probability of the token at position t given its previous context. A CLM operates similarly to an LM, but additionally receives source information x:

$$P(\mathrm{y}|\mathrm{x}) = \prod_{t=1}^{T} P(y_t | y_{<t}, \mathrm{x}). \tag{2.4}$$

The standard modeling paradigm for sequence-to-sequence tasks, including translation, is the *encoder-decoder* framework (K. Cho et al., 2014b). This framework consists of two key components: the *encoder* that reads the source sequence and produces a representation of it, and the *decoder* that uses the source representation from the encoder to generate the target sequence. While over the years, different encoder-decoder networks were proposed, they all share the same high-level translation pipeline: The source and previously generated target tokens are fed into the network; for the source and previous target, the network retrieves a vector representation from the decoder, from which it predicts a probability distribution for the next token in the target sequence. The probability distribution is a vector of size $|V|$ for a vocabulary V of target language tokens. The network uses a linear layer to transform the decoder vector representation of dimensionality d into a $|V|$-sized vector z. Afterward, the *softmax* operation is applied; it ensures that the final outputs are positive numbers that sum up to 1 so that they can be treated as probabilities. Formally, the softmax operation is defined as:

$$softmax(\mathrm{z}) = \frac{exp(\mathrm{z})}{\sum_{i=1}^{|V|} exp(z_i)}, \tag{2.5}$$

where z_i denotes the i-th component in z. Finally, the token corresponding to the highest probability among the probabilities produced by the softmax function is the final output token.

The simplest encoder-decoder model consists of two recurrent neural networks (RNNs), e.g., two long short-term memory (LSTM) networks: one for the encoder and another for the decoder. An RNN encoder reads the source sentence by sequentially processing the tokens of a sequence while maintaining an internal hidden state that encodes information about the tokens it has seen so far. In general, the hidden state at any position n can be computed based on the current input and the previous hidden state:

$$h_n = RNN_{enc}(x_n, h_{n-1}). \tag{2.6}$$

The final state is used as the initial state of the decoder RNN. The idea is that the final encoder state—referred to as *context vector*— encodes all information about the source that the decoder needs to generate the target sequence. The decoder then repeatedly retrieves a state vector for each position in the target sequence, compressing the decoding history into a state vector at any given time step (i.e., position) t:

$$s_t = RNN_{dec}(y_{t-1}, s_{t-1}). \tag{2.7}$$

2.1.3 Training

Generally, the training objective of a neural network is to minimize the *loss*—a metric that assesses the network's prediction error between the true output (the target) and the predicted output on the training data. At each training step, the model tries to maximize the probability it assigns to the correct next token and adjusts its parameters θ according to the error. The standard loss function in NMT is the *cross entropy (CE)*; expressed as the sum over the vocabulary V at position t for a sentence of length T, it is computed as:

$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = -\frac{1}{T}\sum_{j=1}^{|V|} y_{t,j}\, log(\hat{y}_{t,j}), \tag{2.8}$$

where $\hat{y}_{t,j}$ is the softmax output probability distribution over the vocabulary V at time step t. The goal is to find a set of model parameters θ for which the CE error is minimal. In a loss landscape, projecting error values across a parameter space, a model moves toward a local minimum using *gradient descent* or a variant such as stochastic gradient descent (SGD) (Kiefer & Wolfowitz, 1952). The gradient measures the direction of the steepest ascent in the loss landscape given the current parameters. Gradient descent, an iterative optimization algorithm, takes a step in the direction of the negative gradient of the loss function to reduce the loss as quickly as possible. The gradient information is used to repeatedly adjust the model parameters as per the following update equation:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha\Delta_{\theta} J^{(t)}, \tag{2.9}$$

where α is the *learning rate*—the parameter in any optimization algorithm that determines the step size at each iteration while moving toward a minimum of the loss function. By virtue of the *backpropagation* algorithm (Rumelhart et al., 1986), we can efficiently compute the gradient of $J^{(t)}$ with respect to θ. Instead of computing gradients on the full training set, SGD computes the loss function and gradients on a subset—a *minibatch*—of the training set. With a well-chosen learning rate α, the parameters θ are guaranteed to converge into a local optimum over time. In practice, instead of the plain SGD optimizer, adaptive learning rate optimizers such as *Adam* (Kingma & Ba, 2014) are found to greatly reduce the training time.

2.1.4 Inference

Decoding the most likely output sequence involves searching through all the possible output sequences based on their likelihood. Since the vocabulary typically spans hundreds of thousands of words, the search space is very large, namely exponential in the length of the output sequence; thus, finding the one translation with the highest probability among all possible translations is impractical.

In practice, heuristic search methods are used to return an approximate that is a "good enough" decoded output sequence for a given prediction. This is typically done by scoring alternative candidate sequences of words based on their likelihood, using a *greedy search* or a *beam search* algorithm to locate candidate sequences of text. The candidates represent parallel searches—i.e., beams—through the sequence of probabilities; the number of beams is controlled by a parameter called beam width k. Larger beam widths result in better performance of a model as multiple candidate sequences increase the likelihood of better matching a target sequence. However, this improved performance results in a decrease in decoding speed. At each time step t during inference, beam search expands all possible translations and keeps only the k best candidates with the highest probability as the most likely possible translations. The search process can terminate for each candidate separately by reaching a maximum length, an end-of-sequence token, or a threshold likelihood.

2.1.5 State-of-the-Art Architecture

A fundamental problem in NMT is the translation of long sentences due to a weakness of the fixed-length context vector (K. Cho et al., 2014a): In RNNs, the further back we go in the sequence, the more computation steps there are between an input and an output vector; this causes RNNs to "forget" information conveyed by words that occur early in the sequence. This issue gave rise to a different concept to form the context vectors, first introduced by Bahdanau et al. (2014), known as *attention*.

Attention

Attention is the concept of selectively concentrating on certain aspects of information while ignoring other perceivable information. In deep learning, attention provides a network with a "memory" containing facts that can later be exploited to sequentially perform a task, at each time step having the ability to put attention on different elements in the memory depending on their importance indicated by a weight. Instead of compressing an entire sequence's information into a single vector, attention produces representations for all tokens.

At each decoding step, the attention mechanism involves the following operations. First, it computes attention scores from the current decoder state s_t, and all encoder states $h_1, h_2, ..., h_N$. For each encoder state h_n, attention computes its "relevance" for the current decoder state s_t. Specifically, it applies a scoring function that receives one decoder state and one encoder state and returns a scalar value called *attention score*. A popular choice of scoring function is the dot-product operation which gives an intuition about how much two vectors point in the same direction; therefore, it computes a similarity score. Next, attention applies the softmax operation to the attention scores to obtain a probability distribution, typically referred to as *attention weights*. Finally, the *attention output* is computed as the weighted sum of all encoder states (i.e., multiplied by the attention weights).

The introduction of attention is considered a milestone in NMT architecture research; since then, it has become a vital part of various NMT architectures, culminating in the Transformer architecture (Vaswani et al., 2017).

Transformer

In 2017, Vaswani et al. (2017) proposed a novel neural network architecture, the Transformer, which relies only on attention mechanisms without any recurrences. Besides better translation quality, the model is faster to train and has become the de facto standard for many sequence-to-sequence tasks. Whereas RNNs process sequences sequentially, in a Transformer encoder and decoder, tokens interact with each other, all at once, regardless of their position in the sentence, using the concept of *self-attention*.

Since words (e.g., "can") often have more than one meaning represented in their embedding, it is crucial to leverage context information (from other words) to recognize which meaning the model should pay attention to. To highlight different aspects encoded within the embeddings, they experience three separate linear transformations[3] resulting in three vectors: a *query*, a *key*, and a *value* vector. The query-key-value concept of self-attention is analogous to a retrieval process. A *query* is seeking information to gain a better semantic understanding of itself and therefore consults a memory of key-value pairs. Each *key* responds to a query's request and is used to compute an attention weight; the weights represent similarity scores between a query and the keys. The *values* are used to compute the attention output, delivering their information

[3] These transformations occur when the embeddings are multiplied by three different weight matrices, which are learned through training.

to any query that inquires about it. Hence, for each query vector, an attention output vector is computed as a weighted sum of the value vectors.

Under the assumption that query, key, and value vectors have the same dimension d, they can be stacked into matrices $Q \in \mathbb{R}^{n \times d}$, $K \in \mathbb{R}^{m \times d}$, and $V \in \mathbb{R}^{m \times d}$ (for n queries and m key-value pairs). In their paper, Vaswani et al. (2017) proposed the novel *scaled dot-product attention*, where they scaled down the dot-product similarity scores for numerical stability before passing them through the softmax function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.10}$$

with a scaling factor $\sqrt{d_k}$ and K^T being the transpose of K. The resulting matrix holds contextual information for each token embedding.

Often, a word will have to pay attention to multiple other words, which a single self-attention operation may not be capable of. *Multi-head attention* expands the model's ability to focus on different positions in the text, performing H self-attention operations in parallel. It uses not one but H attention heads and H sets of weight matrices for queries, keys, and values. After training, each set is used to project the input embeddings, or vectors from lower encoders or decoders, into a different representation subspace. The output of multi-head attention is the concatenation of the outputs of all attention heads multiplied with yet another weight matrix W^O, ensuring the resultant vector is of the same size as the input embedding vector. The dimensionality of the attention heads is usually divided by H to avoid increasing the number of model parameters. Formally, multi-head attention can be expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_H)W^O, \tag{2.11}$$

with weight matrix $W^O \in \mathbb{R}^{d \times d}$, where

$$\text{head}_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V), \tag{2.12}$$

with weight matrices $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times \frac{d}{H}}$ for $h \in [1, H]$.

Multi-head attention occurs in three ways in the Transformer model. First, in the encoder, multi-head self-attention produces context-sensitive word representations, which depend on the whole source sentence. Second, in the decoder, *masked* multi-head self-attention is used to condition the outputs on the current translation history. At each position, the decoder is allowed to attend to all decoder positions up to and including that position but nothing beyond. With the entire target sequence being fed to the decoder during training, the model needs to prevent decoder positions from attending to subsequent positions. Therefore, scaled dot-product attention masks all values in the input of the softmax corresponding to future positions in the sentence. Third, multi-head *cross*-attention is used to condition the decoder on the source sequence, such that every position in the decoder attends to all positions in the input sequence. Here, attention is computed using the queries from previous decoder layers and the keys and values from the output of the encoder.

The architecture of the Transformer proposed by Vaswani et al. (2017) is depicted in Figure 2.1. In their paper, they use an encoder and a decoder composed of a stack of N=6 identical layers and H=8 attention heads. Each encoder layer consists of two sub-layers, a multi-head self-attention mechanism, and a simple position-wise fully connected feed-forward network. In addition to the two sub-layers in encoder layers, decoder layers include a third sub-layer, which performs multi-head cross-attention over the output of the encoder stack. In the encoder and decoder, the authors further employ a *residual connection* (He et al., 2016) followed by *layer*

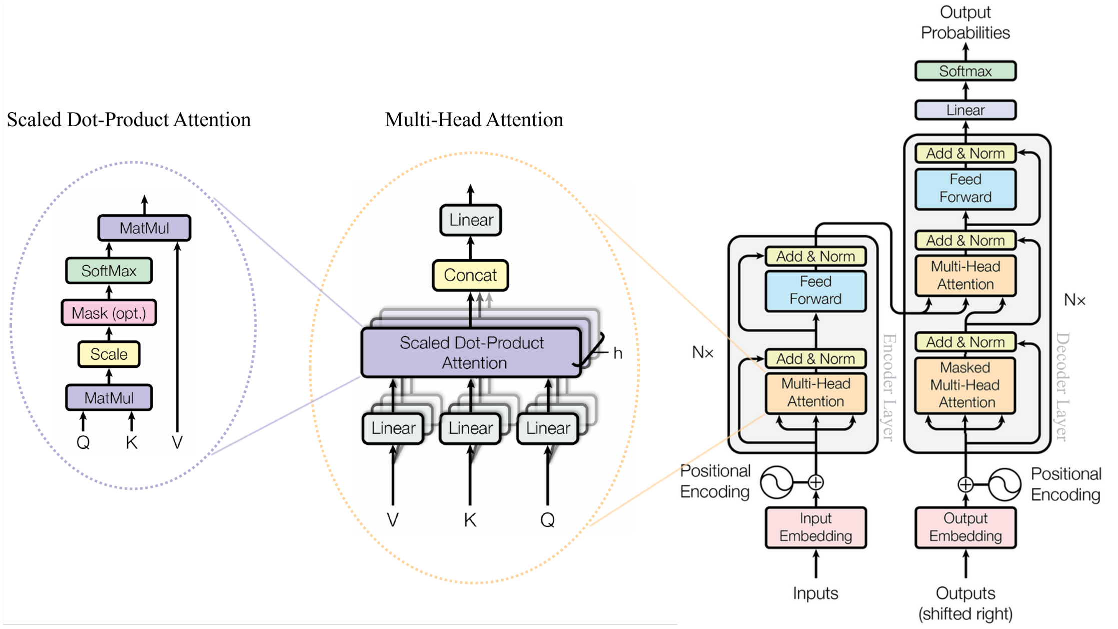


Figure 2.1: Scaled dot-product attention (left) and multi-head attention (middle) in the Transformer (right) proposed by Vaswani et al. (2017).

normalization (Ba et al., 2016) to counteract vanishing gradients, which effectively prevent the model weights from changing during training, and to improve performance and training time, respectively. To introduce a notion of order into the model, Vaswani et al. (2017) add stacked sine and cosine functions of different frequencies as *positional encodings* to the input embeddings at the bottoms of the encoder and decoder stacks.

2.1.6 Multilingual Translation

Since end-to-end translation systems rose to the state of the art, researchers have been seeking to develop a single model for translation between as many languages as possible through the effective use of available linguistic resources. Naturally, the lack of training data among resource-poor languages poses a key problem in multilingual translation. So far, existing works have pursued different lines of research to address this problem. Intuitively, collecting more training data and making full use of the potential of that data is one strategy to follow. Compared with parallel corpus collection, it is easier to obtain a large amount of monolingual data, which can be used for training data augmentation. A widely used method is *back-translation* (Sennrich et al., 2015), in which the main idea is to first train a standard NMT model on a small parallel corpus, and then use the model to translate a large quantity of monolingual data, to generate a *pseudo bilingual corpus* that can be used to retrain the model.

Another promising line of research is to leverage resource-rich languages, such as English, to improve the translation of resource-poor languages. There are two different ways of achieving this: pivot-based translation and zero-shot translation.

In *pivot-based translation*, a high-resource language is used as a pivot language to build a bridge between low-resource language pairs. The simplest pivot-based translation method uses two cascaded models, where the first model translates the source sentence into the pivot language,

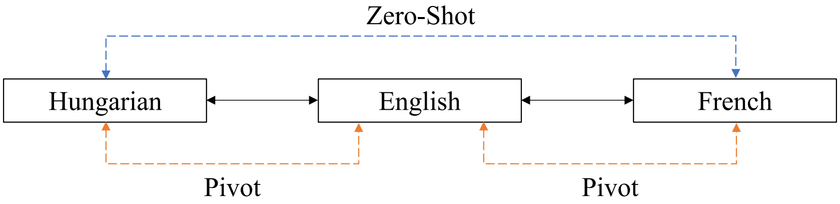


Figure 2.2: Multilingual translation setting with parallel training data for four translation directions: Hungarian↔English and English↔French. Hungarian↔French translations are inferred directly (zero-shot) or by translating via English (pivoting).

and the second model subsequently translates the pivot sentence into the target sentence. Figure 2.2 illustrates pivot-based Hungarian-French translation via English. Pivot-based translation is widely used in practical systems as it is easy to implement. However, the downside of this method is that the cascaded system suffers from error propagation: a translation error in the source-pivot translation will be propagated down the line to the pivot-target translation.

A different approach uses implicitly-learned bridging to translate between low-resource languages: The idea is to encapsulate several translation directions in a single model, where all parameters are shared implicitly by all language pairs modeled; this forces the model to generalize across language boundaries during training. These models are capable of *zero-shot translation*, that is translation between language pairs the system has never seen in this combination during training. For example, a multilingual NMT model trained with Hungarian→English and English→French examples can generate reasonable translations for Hungarian→French, despite never having seen any data for that language pair. When language pairs with little available data and language pairs with abundant data are mixed into a single model, translation quality on the resource-poor language pairs is, in some cases, significantly improved because the model acquires extra knowledge from the other languages (Shatz, 2017; Zoph & Knight, 2016). Compared to pivot-based translation, direct translation circumvents error propagation and reduces computation time. However, achieving high-quality zero-shot translation is a challenging task as prior works have shown that standard systems tend to generate poor outputs.

Existing works suggest that this difficulty arises from models' hidden language representations explicitly tailored to the language pairs included during training. As such, they are not well suited to model new, unfamiliar languages, as is the requirement in zero-shot translation, especially when they are very different concerning linguistic features (e.g., morphology and syntax).

Facilitating Zero-Shot Translation

Existing work (Liu et al., 2021) suggests that the difficulty in high-quality zero-shot translation arises from models' hidden language representations explicitly tailored to the language pairs included during training. As such, they are not well suited to model new, unfamiliar languages, as is the requirement in zero-shot translation, especially when they are very different concerning linguistic features (e.g., morphology or syntax).

Removed Residual Connection Liu et al. (2021) demonstrate that a main factor causing language-*specific* representations is the strong *positional correspondence* of encoder representations to input tokens. Because sentences and their translations are often of varying lengths and word order, the same semantic meaning gets encoded into different representations, i.e., different hidden state sequences, rather than into language-*agnostic* representations based on which the decoder can translate into any target language required. In light of this, the authors relax this structural constraint of the typical Transformer and offer the model some freedom for word reordering in the encoder. To achieve this, they remove residual connections in a middle encoder layer. Their results demonstrate that removed residual connections can facilitate zero-shot translation, whereas it worsened pivot-based translation (Liu et al., 2021).

Similarity Regularizer Another approach shown to facilitate zero-shot translation introduces *auxiliary training objectives* to *encourage similarity between representations* of different languages (Arivazhagan et al., 2019; Liu & Niehues, 2021; Pham et al., 2019): The core idea is that sentences that are semantically similar should also be represented similarly, regardless of the source language. Besides the translation loss, Liu and Niehues (2021) encourage the model to minimize the Euclidean distance between the (mean-pooled) source and target sentence embeddings. A drawback of the similarity regularizer is the need to select pooling methods, distance metrics, and a suitable scaling factor.

Adversarial Language Classifier A different strategy to *remove source language signals* in encoder representations is *adversarial training* (Ganin et al., 2016): Similar to Arivazhagan et al. (2019), the idea is to train an adversarial classifier—jointly with the translation model—that aims to identify the source language given a representation. High accuracy of this adversarial language classifier implies unequivocal signals within encoder representations hinting at the source language. To discourage source language-specific representations, the loss function is designed to penalize high language classification accuracy:

$$\mathcal{L}_{enc\_dec} = \mathcal{L}_{MT} + \mathcal{L}_{adv\_classifier}, \text{ with} \tag{2.13}$$

$$\mathcal{L}_{adv\_classifier} = \sum_{c=1}^{C} y_c \cdot log(1 - p_c), \tag{2.14}$$

where C is the number of classes to predict, y_c is a binary indicator whether the true language label is c, and p_c is the probability predicted for language c.

2.2 Gender Bias in Machine Translation

Bias and discrimination is one of the most challenging ethical issues at the center of attention in the field of AI. Some of the early pioneers on the topic of bias in computer systems, Friedman and Nissenbaum (2017), use the term bias to refer to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others". Interest in understanding, assessing, and mitigating bias is steadily growing within the NLP community, as it is widely recognized that NLP tools encode and reflect controversial social asymmetries and have the potential to propagate them further for many seemingly neutral tasks, MT included (Savoldi et al., 2021).

One of the most widespread societal asymmetries that exist in such systems is gender bias, which can be defined as the preference or prejudice toward one gender over the other. Naturally, gender in language is signaled through grammatical gender (e.g., *el* sol$_M$ versus *la* luna$_F$) or semantic gender in the form of gender identity (e.g., rooster$_M$ versus hen$_F$). In society, human bias introduces another category of semantic gender associations that is unwarranted and factitious, as it is not founded in linguistic systems' rules (e.g., doctor$_M$ versus nurse$_F$). This category present in natural language training examples causes NLP models to learn and reflect human gender biases.

Gender bias can manifest itself in different parts of the predictive pipeline of NLP systems, such as in the raw training data, word representations, and language modeling algorithms. NLP systems exhibiting biases in any of these parts can produce gender-biased predictions and even amplify biases present in the training data. One example of such discriminatory behavior can be observed in translation models; they are highly susceptible to gender biases because they are typically trained on parallel corpora featuring social bias and demographic misrepresentation across a spectrum of different languages.

2.2.1 Manifestation

In a large-scale analysis of the plethora of existing research addressing gender bias in MT, Savoldi et al. (2021) categorize them based on two conceptualizations of the problem: research works focusing on the weight of *prejudice and stereotypes* in MT, and studies assessing whether *gender is preserved* in translation.

Gender Stereotyping

The problem of stereotyping in MT can appear in different translation settings. An obvious case is the translation of phrases in gender-neutral languages, such as Hungarian or Korean, into languages that encode gender markings (e.g. English or French). The propagation of stereotypes is a widely researched form of gender bias in MT, one that, so far, has often been narrowed down to *occupational* stereotyping (Savoldi et al., 2021): This type of stereotyping has been investigated through the translation of occupations from gender-neutral languages such as Hungarian or Korean into English (W. I. Cho et al., 2019; Prates et al., 2020). A common finding is that large-scale MT systems yield a *masculine default*, showing strong a bias toward predicting male pronouns. A typical translation setting of this form, exemplifying ambiguity due to gender-neutrality in the source sentence, is illustrated in Figure 2.3.

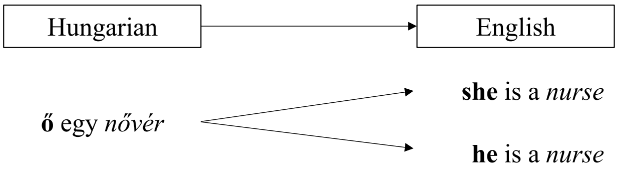
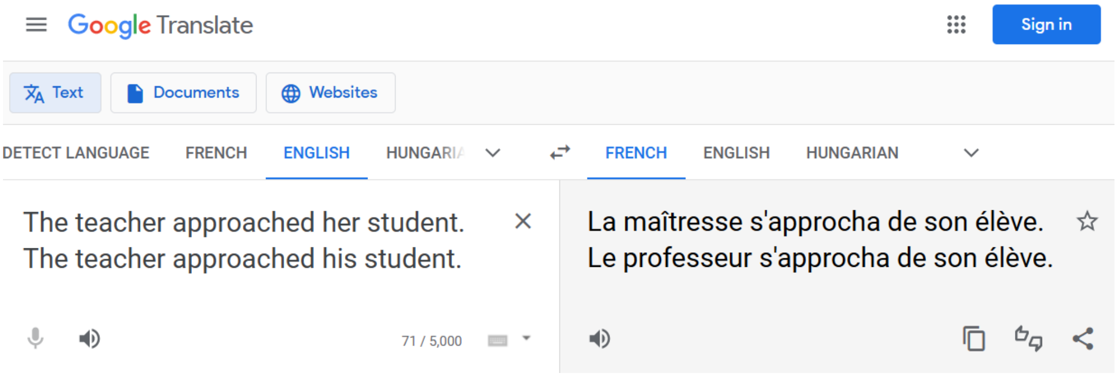


Figure 2.3: Illustration of the setting in which gender stereotyping in translation can occur: While the Hungarian source sentence is gender-neutral, the English translation requires specification of the gender of the referenced person. A biased system might mirror imbalances in the training data and therefore reflect stereotypical gender roles in its translations (i.e., in the given example output "*she* is a nurse").

The harmful effect that this can have is revealed when translating a simple sentence as "The teacher approached *her* student" from English to French, as can be seen in Figure 2.4. Google Translate turns the profession of a female teacher into that of a *mistress*, revealing a sexist association of the expressed teacher-student relationship in combination with the feminine possessive determiner *her*. In contrast, the same sentence with just the possessive determiner changed to *his* instead of *her* results in a more accurate translation (i.e., "le professeur"). While this translation behavior in itself is very concerning, MT has also proven to exacerbate gender disparities. By comparing the proportion of pronoun predictions against the real-world proportion of men and women employed in the respective professions, Prates et al. (2020) demonstrated



The teacher approached her student.
The teacher approached his student.

La maîtresse s'approcha de son élève.
Le professeur s'approcha de son élève.

Figure 2.4: English→French translation generated by Google Translate, where the word "teacher" referring to a female is translated to "maîtresse", which also translates back to the English word "mistress". For the masculine counterpart sentence, an evidently less sexist translation is provided.

that Google Translate underestimates feminine frequency at a greater rate than occupation data alone suggests.

Gender Preservation

A different scenario proven challenging for MT systems is preserving gender when translating to gender-inflected languages, such as most Romance, Germanic, Slavic, and Semitic languages. Figure 2.5 illustrates an exemplary translation setting where this challenge occurs.

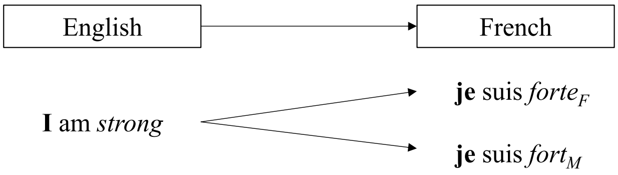
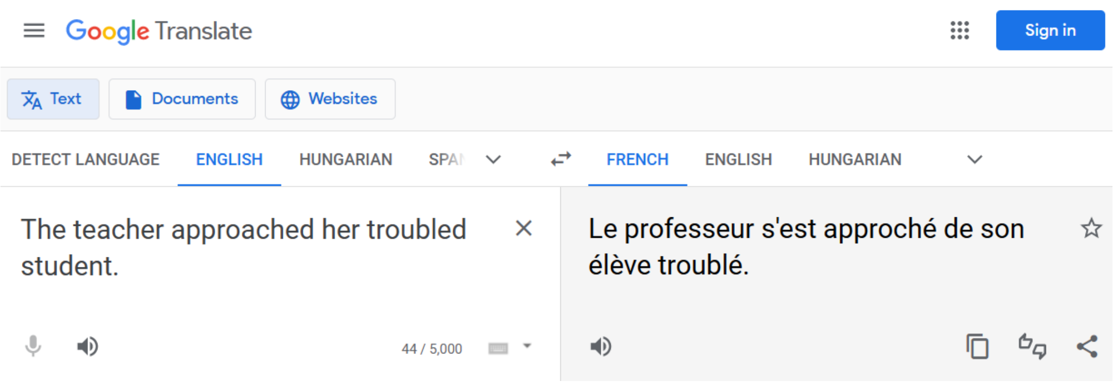


Figure 2.5: Illustration of the challenge of gender preservation in English→French translation, as the French translation is different for female and male speakers. A biased system might mirror imbalances in the training data and therefore reflect stereotypical gender roles in its translations (i.e., output the translation "je suis $fort_M$" for a male speaker).

The English source sentence "I am strong" gives no indication about the gender of the entity articulating the sentence. As French includes gender inflection within adjectives, a translation must agree with the gender of the author of the English source sentence. As there is no gender information in the source, the MT system must choose a gender; again, a biased system would base the decision on gender stereotypes (i.e., as men are typically assumed stronger than women, it outputs the male translation "je suis $fort_M$").

This ambiguity problem does not persist when the gender of the entity is included. Nevertheless, correctly resolving pronouns to multiple entities occurring in a translation scenario remains a challenge in translation to gender-inflected languages. For instance, in the teacher-student example, the pronominal adjective *her* referring to the teacher explicitly marks the teacher's gender as female; accordingly, there is enough information in the English source sentence to preserve the gender in the French translation. Nevertheless, as Figure 2.6 shows, Google Translate generates a translation referring to a male teacher.



The teacher approached her troubled student.

Le professeur s'est approché de son élève troublé.

Figure 2.6: English→French translation generated by Google Translate in which a female teacher turns into a male ("le professeur") despite a pronominal gender marking ("*her* troubled student)" being present in the English source sentence.

Notably, the addition of the word *troubled* (cf. Figure 2.6) turns the profession *teacher* into the semantically correct French translation *professeur*—in contrast to the previous wrong translation to *mistress* (in Figure 2.4)—but the gender of the article in the translated sentence is still incorrect. As such, Google Translate not only produces erroneous translations of the profession of females but even overrules explicit gender signals when they contradict presumed gender stereotypes.

2.2.2 Evaluation

In response to the multifaceted occurrence of gender bias in MT, various methods have been proposed to measure and evaluate it. Similar to regular MT evaluation, gender bias evaluation is built on two core components: appropriate evaluation metrics and suitable evaluation data.

Evaluation Metrics

Prior works have utilized BLEU (Bentivogli et al., 2020; Moryossef et al., 2019; Vanmassenhove et al., 2018) and accuracy (Bentivogli et al., 2020) scores as qualitative and quantitative performance metrics for gender bias evaluation in MT.

BLEU Short for **bil**ingual **e**valuation **u**nderstudy, BLEU is a widely used metric for the evaluation of MT (Papineni et al., 2002). The basic idea behind BLEU is to assign a single numerical score to a translation that determines how good it is compared to a reference translation. The reference translation is often human-generated and, following the idea that the human-generated reference is the best possible output, BLEU treats translations that are closer (meaning on a word level they are more similar) to the reference as better; for this, BLEU compares the n-grams—sequences of words or subwords—of the generated translation with the n-grams of the reference translation(s). While there are constant research efforts toward developing metrics that better correlate with human judgements, BLEU is still the de facto standard in MT evaluation.

Accuracy Generally, the accuracy is a metric that measures the proportion of correct classifications in relation to the total number of predictions made by an ML model. In the context of gender bias evaluation, the metric is used to measure the accuracy in producing the gender-marked words in target sentences. More precisely, it can be computed as the number of properly translated gender-marked words in relation to the total number of gender-marked words in a sentence. Accuracies for feminine and masculine words, or for other gender-related phenomena, can then be compared to determine if the performance for one group performs better than for the other.

Evaluation Data

Arguably the most extensive annotated benchmark for gender-sensitive evaluation of MT is MuST-SHE[4] (Bentivogli et al., 2020)—a subset of the TED-based MuST-C (Di Gangi et al., 2019), short for **mu**ltilingual **s**peech **t**ranslation **c**orpus[5]—allowing for a fine-grained analysis of gender bias in text and speech translation. MuST-SHE comprises 3,367 sentence-level instances of (audio, transcript, translation) triplets, uttered by 295 different speakers. The collection of data is based on audio recordings of TED talks originally held in English that are manually transcribed and translated into three target languages: French, Italian, and Spanish. All three target languages are Romance languages that extensively express gender via feminine or masculine morphological markers on nouns, adjectives, verbs and other functional words such as articles and demonstratives.

Each MuST-SHE triplet includes at least one English gender-neutral word that requires translation into the corresponding feminine or masculine target word(s), where such formal distinction semantically conveys and conflates with an actual distinction of biological sex. For better analysis, the gender of all gender-marked words in a sentence is the same, resulting in exclusively feminine or exclusively masculine sentences. Besides the sentence gender, MuST-SHE triplets are annotated with further gender-related phenomena. Overall, the gender-related annotations comprise three distinguishing aspects, which can be thought of as triplet properties, each of which decomposes the corpus into two groups of instances, as illustrated in Figure 2.7:

(1) the gender expressed throughout the sentence, which is either *feminine* or *masculine*;

(2) the source of information necessary to disambiguate gender, which is either

    (a) the *audio recording* of speakers' tone of voice when gender-agreement only depends on speakers' gender, e.g., when speakers use a first person point of view, or

    (b) the *utterance content*, where contextual hints inform about the gender of the referent, such as "Mom", pronouns such as "she", "his", or proper nouns such as "Paul";

(3) the gender of the speaker, which is either *female* or *male.*

These different divisions of instances make MuST-SHE suitable for a comprehensive and multifaceted evaluation of gender bias in MT. For instance, it allows for individual evaluations of translations of sentences spoken by women and men so that differences, if they exist, can be detected. To prevent imbalanced gender representations, the two respective groups always include a similar number of instances, which ensures equal distribution of gender-phenomena in the corpus.

[4] https://ict.fbk.eu/must-she/

[5] https://ict.fbk.eu/must-c/

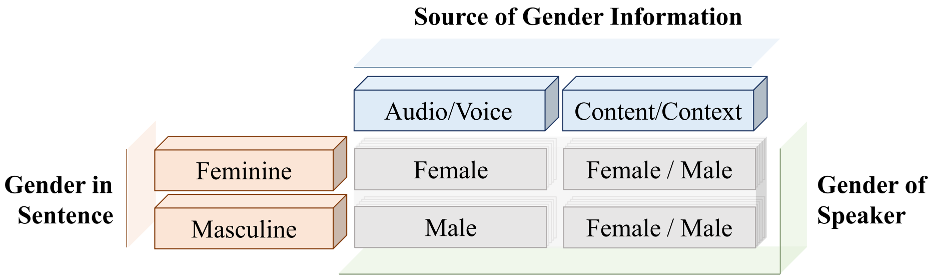


Figure 2.7: Illustration of the subdivision and interplay of three different categorizations of instances comprising MuST-SHE (English→{French, Italian, Spanish}). Instances can be grouped by (1) the gender of words in a reference (exclusively feminine or masculine), (2) the source of information to disambiguate gender (the speaker's tone of voice in the audio recording or the utterance content itself), and (3) the gender of the speaker (female or male).

Evaluation Procedure

With the two core components of suitable metrics and data at hand, the authors of the original MuST-SHE corpus (Bentivogli et al., 2020) propose a qualitative and quantitative evaluation strategy. They rightly point to the fact that the BLEU metric provides only a global score measuring the translation quality for sentences as a whole. Therefore, variations of BLEU scores are only a coarse and indirect indicator of better or worse overall performance of MT systems (Callison-Burch et al., 2006). This characteristic of BLEU hinders the evaluation of MT models' performance on individual phenomenons such as gender translation since, in language, gender is expressed and marked in only a few of the words occurring in a sentence. In light of this, Bentivogli et al. (2020) point toward the limitation of recent works (Moryossef et al., 2019; Vanmassenhove et al., 2018) relying on the results of a BLEU-based quantitative analysis alone to ascribe BLEU gains to better gender translation and thus to their bias mitigation strategies.

Bentivogli et al. (2020) continue to use BLEU, but they propose to make the resulting scores informative about a system's ability to generate the correct gender forms. To this aim, for each reference *C-REF* in the corpus, they create a "wrong" one, *W-REF*, that is identical to *C-REF* except for the morphological signals that convey gender agreement. More precisely, for each gender-inflected word in the target sentence, the correct translation is swapped into its opposite gender form. For instance, the French words with masculine inflection "un", "grands", and "innovateurs" are changed into the feminine-marked words "une", "grandes", and "innovatrices". Gender-swapping is performed for any gender-marked part of speech. The result is a new set of references that, compared to the correct ones, are "wrong" only with respect to the formal expression of gender. For better illustration, some examples extracted from MuST-SHE are presented in Table 2.1.

The underlying idea put forward by the authors is that, as the two reference sets differ only in the swapped gender forms, differences in results for the same set of translations generated by a given model can measure its capability to properly handle gender phenomena. Specifically, they argue that higher values on the wrong set signal gender-biased behavior. In all cases where the required gender realization is feminine, significantly higher BLEU scores computed on the wrong set would signal a bias toward producing masculine forms, and vice versa. In addition to the qualitative BLEU-based evaluation, they suggest performing a fine-grained quantitative analysis of the models' accuracy in producing the target gender-marked words. The accuracy scores are computed on both the correct and the wrong reference set for all instances, as well

Table 2.1: MuST-SHE annotated segments organized by category. For each source-translation pair in en-es, en-fr, and en-it, the correct reference translation (C-REF) shows the realization of target gender-marked forms (Fem./Masc.) corresponding to English gender-neutral words in the source (SRC). In the wrong reference translation (W-REF), Spanish, French, and Italian gender-marked words are swapped to their opposite gender form. The rightmost column of the table provides information about the speaker's gender (Female/Male).

Word Form		**Source of Gender Information**	**Speaker Gender**
		Category 1: *Gender information only in audio*	
Feminine	SRC	I felt **alienated**, **intimidated** and **judged** by many.	Female
	C-REF$_{es}$	Me sentí **alienada**, **intimidada** y **juzgada** por muchos.	
	W-REF$_{es}$	Me sentí **alienado**, **intimidado** y **juzgado** por muchos.	
	C-REF$_{fr}$	Je me sentais **isolée**, **intimidée** et **jugée** par beaucoup.	
	W-REF$_{fr}$	Je me sentais **isolé**, **intimidé** et **jugé** par beaucoup.	
	C-REF$_{it}$	Mi sentivo **esclusa**, **intimidita** e **guidicata** da molti.	
	W-REF$_{it}$	Mi sentivo **escluso**, **intimidito** e **guidicato** da molti.	
Masculine	SRC	I was **an obsessive compulsive student**.	Male
	C-REF$_{es}$	Yo era **un** estudiante **obsesivo compulsivo**.	
	W-REF$_{es}$	Yo era **una** estudiante **obsesiva compulsiva**.	
	C-REF$_{fr}$	J'étais **un étudiant obsessif compulsif**.	
	W-REF$_{fr}$	J'étais **une étudiante obsessive compulsive**.	
	C-REF$_{it}$	Ero **uno studente ossessivo compulsivo**.	
	W-REF$_{it}$	Ero **una studentessa ossessiva compulsiva**.	
		Category 2: *Gender information in utterance content*	
Feminine	SRC	She was **tough**, she was **strong**, she was **powerful**.	Male
	C-REF$_{es}$	Era **dura**, era fuerte, tenía poder.	
	W-REF$_{es}$	Era **duro**, era fuerte, tenía poder.	
	C-REF$_{fr}$	Elle était difficile, elle était **forte**, elle était **puissante**.	
	W-REF$_{fr}$	Elle était difficile, elle était **fort**, elle était **puissant**.	
	C-REF$_{it}$	Era **dura**, era forte, era potente.	
	W-REF$_{it}$	Era **duro**, era forte, era potente.	
Masculine	SRC	He was **one of the** worst **students** in class.	Female
	C-REF$_{es}$	Era **uno** de **los** peores de su clase.	
	W-REF$_{es}$	Era **una** de **las** peores de su clase.	
	C-REF$_{fr}$	C'était l'**un** des pires élèves de la classe.	
	W-REF$_{fr}$	C'était l'**une** des pires élèves de la classe.	
	C-REF$_{it}$	Era **uno degli studenti** peggiori della sua classe.	
	W-REF$_{it}$	Era **una delle studentesse** peggiori della sua classe.	

as per gender form (feminine or masculine). Furthermore, the corpus allows for an analysis of performance on sentences per category (gender information in audio or utterance content) and per speaker gender (female or male).

2.2.3 Mitigation

From the growing interest in exposing and actively fighting gender bias in recent years, a number of works on gender bias mitigation in MT have emerged, relying on a range of different techniques.

A recent comprehensive overview of different gender bias mitigation techniques in MT is provided by Savoldi et al. (2021). Following their classification, we can group mitigation techniques into two categories: debiasing models and debiasing through external components.

Model Debiasing

The first group of techniques focuses on architectural changes of general-purpose MT models or on dedicated training procedures to mitigate gender bias. One line of research studies the effect of gender tagging. Specifically, Vanmassenhove et al. (2018) leverage additional gender information by prepending a gender tag to each source sentence, both at training and inference time, to improve the generation of speaker's referential markings. While their approach proves useful to handle morphological agreement when translating from English into French, this solution requires additional (training) metadata regarding speakers' gender that might not always be feasible to acquire. Moreover, Saunders et al. (2020) argue that gender tagging introduces noise if applied to sentences with references to multiple participants, as it pushes translations toward the same gender. Avoiding additional information needed for training or inference, Basta et al. (2020) adopt a generic approach, concatenating each sentence with its predecessor; providing more context information in such a manner resulted in a slight improvement in gender translations.

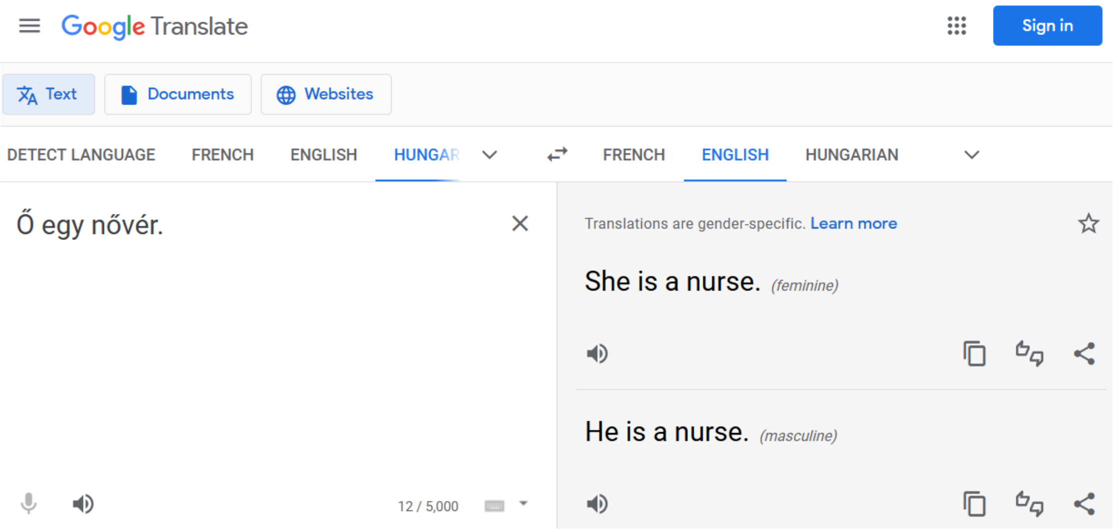
While the two previously mentioned mitigation strategies share the intent of supplying the model with additional gender knowledge, Escudé Font and Costa-Jussa (2019) follow a different approach, leveraging pre-trained debiased word embeddings of English gender-neutral words where gender associations are removed. The downside of this approach is that it is highly language-specific as translation of gender-marked languages requires the presence of gender associations in the embeddings. A different debiasing approach is to fine-tune models on gender-balanced datasets, featuring equal amounts of feminine and masculine references, which has shown to improve the generation of feminine gender overall (Costa-jussà & de Jorge, 2020). However, this approach does not mitigate gender stereotyping if it does not account for the qualitative different ways in which men and women are portrayed (Wang et al., 2022); equal representation of gender must also apply to equal representation among roles, occupations and other aspects.

Debiasing Through External Components

While the first group of mitigation strategies focus on directly debiasing the MT model, the second group intervenes in the inference phase with external dedicated components. Such approaches do not require retraining of the MT model. However, they introduce additional costs of maintaining separate modules and integrating them with the MT model. One strategy is to inject context information to disambiguate the translation of gender-neutral words. Moryossef et al. (2019) prepend a short phrase, such as "*she* said to *them*", to the source sentence, translate the sentence with the prefix, and afterward remove the prefix translation from the MT output. Specifying gender and numeral inflection (plural or singular) improves models' ability to generate feminine target forms. The drawback of this approach is that it relies on metadata about speakers and listeners, which is not always available.

A different approach is to post-process the MT output using counterfactual data augmentation, where sentences are transformed into a set of identical sentences differing only in terms of gender references. Saunders and Byrne (2020) use a lattice rescoring module which maps gender-marked words in the MT output to all possible inflectional variants. The module then rescores all paths in the lattice corresponding to the different sentences with a model that has been gender-biased at the cost of lower overall translation quality. Choosing the sentence with the highest score as the final translation results in increased accuracy of gender selection. A downside is that data augmentation is very demanding for complex sentences with a variety of gender phenomena, such as those typically occurring in natural language scenarios.

A different approach, avoiding this effort, is post-processing to re-inflect gender. For example, when provided with an indication of preferred gender, a model can re-inflect sentences in the desired form (Alhafni et al., 2020). This approach naturally requires metadata in the form of users' input, which limits the level of automation of translation. A different way of approaching re-inflection is to always produce both gender forms from an MT output. A leading example for this approach is Google Translate: Since 2019, Google Translate provides both feminine and masculine translations for single gender-neutral words or sentences, as illustrated in Figure 2.8. Unfortunately, this feature is currently limited to single sentences and supports only a limited number of language pairs.



Ő egy nővér.

Translations are gender-specific. Learn more

She is a nurse. (feminine)

He is a nurse. (masculine)

Figure 2.8: In an effort to reduce gender bias in its translations, Google Translate provides both feminine and masculine translations for single gender-neutral words or sentences. However, as of April 2023, the feature only supports a limited number of language pairs.

3

Experimental Design

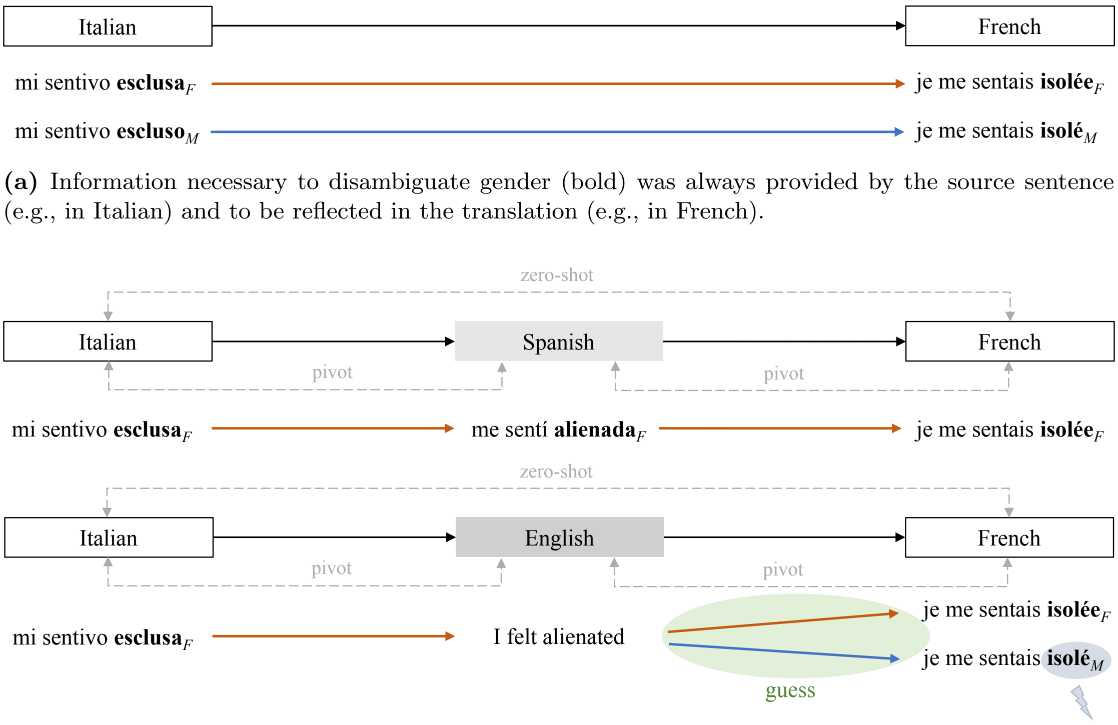
This chapter describes the experimental design and methods employed in this thesis and explains how the research proceeded to answer the research questions. We performed quantitative analyses through experiments in which we systematically varied one or more independent explanatory variables (introduced in Section 3.2) and analyzed their effect on the dependent response variable (introduced in Section 3.1) used to measure gender biases, to learn about their cause-effect relationship. In Sections 3.3 and 3.4, we describe our methods of measurement and the organization of the evaluation data used in our experiments, respectively.

3.1 Response Variable

In this thesis, we explore gender bias in multilingual MT in the context of *gender preservation* for translation directions that are unseen during training. Generally, translation into a gender-sensitive target language faces the task of gender preservation, where the gender information conveyed in the source language needs to be carried into the target language translation. Effectively, this involves words related to a noun — like determiners, pronouns or adjectives — to change their form (inflection) according to the gender of the noun they refer to (agreement). Whenever the source language is (largely) genderless, i.e., the gender of the noun is unspecified (cf. Figure 2.5), and context information is unavailable, gender preservation is a non-trivial task, for machines and humans alike, which inevitably boils down to guessing the gender.

In our study, however, information required to disambiguate gender was *always* provided to the model by the source sentence (cf. Figure 3.1a). In the scope of our inquiry, we generally view gender as dichotomous, conforming with the *gender binary*, i.e., the classification of gender into the opposite forms of feminine and masculine, when it comes to "classifying" the gender of people, which we consider indicative of a person's biological sex[1]. We examined translations in terms of differences in gender preservation between both genders which, if found, were evidence of gender-biased MT. If observed, we analyzed under which circumstances models exhibited gender-biased translation behavior. Accordingly, we used gender bias as the primary dependent response variable of interest in our experiments.

[1] While this view of gender or sex as a binary concept has started to change, and language is constantly evolving (i.e., to gender-neutral language), the gender binary persists in many languages used today.


Italian
French
mi sentivo **esclusa**$_F$
je me sentais **isolée**$_F$
mi sentivo **escluso**$_M$
je me sentais **isolé**$_M$
zero-shot
Italian
Spanish
French
pivot
pivot
mi sentivo **esclusa**$_F$
me sentí **alienada**$_F$
je me sentais **isolée**$_F$
zero-shot
Italian
English
French
pivot
pivot
mi sentivo **esclusa**$_F$
I felt alienated
je me sentais **isolée**$_F$
je me sentais **isolé**$_M$
guess

(a) Information necessary to disambiguate gender (bold) was always provided by the source sentence (e.g., in Italian) and to be reflected in the translation (e.g., in French).

(b) The richness of the gender-inflectional system of the bridge language, used to facilitate translation for unseen language pairs, affects models' ability to preserve the gender information from the source sentence. Scarcity of gender inflection in the bridge language (e.g., English) causes models to miss gender clues from the source and resort to guessing the gender; when making the wrong guess, i.e., choosing the opposite gender as presented in the source, the model exhibits gender hallucination.

Figure 3.1: Overview of our investigated translation scenario: At inference, we translated between unseen gender-inflected source-target language pairs (e.g., Italian→French) by bridging, implicitly (zero-shot) and explicitly (pivot-based), using bridge languages with different gender-inflectional systems (e.g., Spanish or English).

3.2 Explanatory Variables

Motivated by our research inquiry, we used three independent explanatory variables to better understand the cause-and-effect relationship of gender bias in MT. By considering only source sentences in which gender information is always present, we could focus our gender bias investigation on the effect of *bridging* on gender preservation in multilingual MT (of unseen language pairs), as illustrated broadly in Figure 3.1b. In particular, we focused on different aspects revolving around bridging with respect to their influence on gender preservation: the type of bridging performed (explicit or implicitly-learned); the choice of bridge language; and broadening and improving language coverage through language-agnostic model hidden representations.

3.2.1 Explicit & Implicitly-learned Bridging

To bridge the gap between an unknown source-target language pair at inference, we followed two different approaches (pictured in Figure 3.2), both using the same trained translation model.

For *pivot-based translation*, we cascaded the model to perform source→pivot and pivot→target translation. As such, pivoting uses the pivot language as an explicit bridge between the unknown language pair. For *zero-shot translation*, we used the model to translate directly between the unknown language pair, relying on the model's inherent, implicitly-learned bridge. In light of our inquiry, we analyzed each approach's ability to preserve gender, focusing on the comparison of the performances for the feminine and the masculine gender.

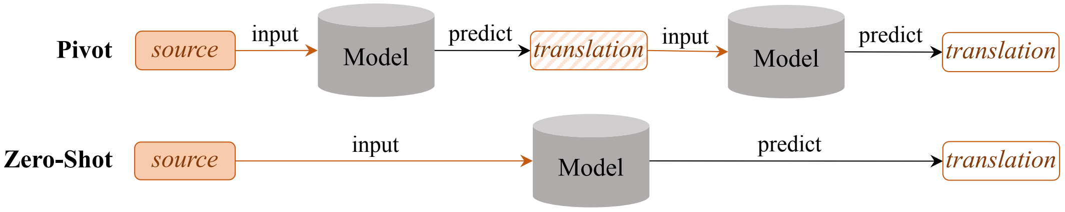


Figure 3.2: Illustration of pivot-based and zero-shot translation, whose performances regarding gender preservation we analyzed and compared.

3.2.2 Bridge Language

English is arguably the most resource-rich language in the context of NLP. For this reason, English often participates in most, if not all, language pairs present in a training corpus, making English the most reasonable choice for a bridge language. Considering the differences in gender-inflectional systems of languages, it is important to note that English is a relatively low gender-inflected language. In view of this, it is fair to assume that bridging with English results in the loss of gender information conveyed by the source sentence; how much explicit and implicitly-learned bridging are affected by this was to be examined in our study.

When translating to a genderless language, the loss of gender information is unproblematic as it is evidently without detrimental consequence. However, when translating to a language with a richer gender-inflected system than English, the loss of gender information poses a significant problem, since the information necessary to disambiguate gender is virtually no longer existent (cf. bottom in Figure 3.1b). As preserving non-existent gender information is inherently impossible, also for humans, it is fair to assume that translation models have difficulty when encountering this phenomenon of gender ambiguity. In cases of gender ambiguity, the simplest solution is to resort to *random guessing*, with a 50% chance of choosing one gender over the other. Any other gender distribution (≠ 50:50%) is not reflective of random guessing but instead indicative of *educated guessing*, guessing based on knowledge or observations *assumed* to be true that can, however, include biases.

Against this background, we explored the role of the bridge language in gender preservation, focusing on the gender bias differences between pivot-based and zero-shot translation using bridge languages with different gender-inflectional systems, including the two grammatical gender languages, Spanish and German, as well as English. German and English are both Germanic languages. Whereas in German, all noun classes require masculine, feminine, or neuter inflection[2], English lacks a similar grammatical gender system. In German, the gender of the noun is also reflected in determiners like articles (see Figure 3.3a), possessives (Figure 3.3b), and demonstratives (Figure 3.3c). Spanish, on the other hand, is a Romance language that has a binary grammatical gender system, differentiating masculine and feminine nouns; from a grammatical

[2] In German, neuter inflection does not apply to nouns identifying people.

	article	adjective	noun	possessive determiner	noun	demonstrative determiner	noun
English	the	good	teacher	my	teacher	this	teacher
German	**die**$_F$ **der**$_M$	gute gute	**Lehrer*in***$_F$ **Lehrer**$_M$	**mein*e***$_F$ **mein**$_M$	**Lehrer*in***$_F$ **Lehrer**$_M$	**dies*e***$_F$ **dies*er***$_M$	**Lehrer*in***$_F$ **Lehrer**$_M$
Spanish	**la**$_F$ **el**$_M$	**buen*a***$_F$ **buen**$_M$	**profesor*a***$_F$ **profesor**$_M$	mi mi	**profesor*a***$_F$ **profesor**$_M$	**est*a***$_F$ **est*e***$_F$	**profesor*a***$_F$ **profesor**$_M$

(a) (b) (c)

Figure 3.3: Languages differ regarding the word classes (i.e., noun, adjective, determiner) that are inflected for gender to comply with gender agreement. The above example sentences show no gender indication in English, while German and Spanish use grammatical gender and conforming gender inflection (bold).

point of view, there are no gender-neutral nouns. The gender of nouns agrees with (some) determiners and, unlike in German, adjectives (see Figure 3.3a), making gender a very pervasive feature in Spanish. We imagine training models using bridge languages with richer gender inflection, such as Spanish or German, yields better gender translations and reduces gender biases compared to a lower gender-inflected language, such as English.

3.2.3 Language-Agnostic Hidden Representations

Arguably, a model's hidden representations are tailored to the language pairs included during training, likely making them ill-suited for translation between new, unfamiliar language pairs. Since languages are characterized by different linguistic features, including those related to gender, it is reasonable to assume that language-*specific* representations *impair* gender preservation.

Figure 3.4 exemplifies some differences in morphological features of gender and syntax between English, French, and Czech, critical for gender preservation. The illustrated example reveals the need for zero-shot models to infer different and unfamiliar pathways of gender information flow, depending on the source-target language pair. Conceptually, hidden representations tailored to the source language appear inadequate for modeling the gender-related linguistic features of the target language, especially when both languages are very different in terms of gender inflection and gender agreement.

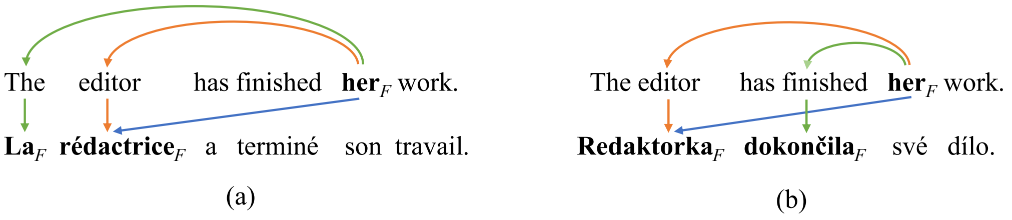

The editor has finished her$_F$ work.
La$_F$ rédactrice$_F$ a terminé son travail.
(a)
The editor has finished her$_F$ work.
Redaktorka$_F$ dokončila$_F$ své dílo.
(b)

Figure 3.4: Illustration of semantic pathways required to transfer gender information for English→French (a) and English→Czech (b) translation.

Because of this, in the context of gender preservation, we explore the effect of three modifications to (the training of) a baseline Transformer *to promote language-agnostic hidden representations*, which previously have caused performance gains for zero-shot translation (for details, see Section 2.1.6):

- we removed a residual connection in a middle Transformer encoder to *lessen positional correspondences* to the input tokens (R),
- we encouraged *similar (i.e., "closer") source and target language representations* through an auxiliary loss (AUX_{SIM}), and
- we performed joint adversarial training *penalizing recovery of source language signals* in the representations (ADV_{LAN}).

In our experiments, we examined the effect of the three modifications in isolation and also tested combinations of some of the three; in total, we compared the gender preservation performance of five different models to that of our baseline (B) — which we refer to as B+AUX_{SIM}, B+ADV_{LAN}, R, R+AUX_{SIM}, and R+ADV_{LAN} — to determine whether they miextigated models' gender biases. The modifications affected model training but did not intervene in the inference phase and were equally applicable to zero-shot and pivot-based translation.

3.3 Methods of Measurement

We evaluated our models' performances by different means of measure: using the concept of "gender-swapping" to measure models' gender preservation for the two different genders; using the standard BLEU metric to verify reasonably good translation quality for gender-marked and gender-neutral words; and probing models' hidden representations for gender signals using a classifier.

3.3.1 Gender Bias

Abstractly speaking, in our gender bias evaluation, we measured how often a model preserved the gender compared to how often it produced the opposite gender form, thus opting for the wrong instead of the correct gender despite the lack of gender ambiguity. Opting for the wrong instead of the correct gender in numerous cases led us to conclude that the model is acting on gender biases.

Following this idea, models' generated translations of gender-marked words could be grouped into the following three different categories, which we exemplify using Figure 3.5. First, the *expected translation*, for which we measured how often the *correct* translation (ground truth) — specified by a reference translation *C-REF* — was produced (e.g., "isolée" in the exemplary model output in Figure 3.5). Second, the *gender-reversed translation*, for which we measured how often the translation was *wrong*, but only regarding the gender inflection of gender-marked words — specified by a reference translation *W-REF* — i.e., instead of the required correct gender realization as per ground truth (e.g., the feminine adjective "intimidée"), the model produced the opposite gender form (e.g., the masculine adjective "intimidé", as depicted in Figure 3.5). Third, a *translation different from both reference translations*, e.g., instead of "jugée" (C-REF) or "jugé" (W-REF), the model produced the adjective "condamnée", or any other word not matching C-REF or W-REF; in this case, we had no way of knowing whether the gender inflection, regardless of the predicted word base, is correct or wrong (because we had no reference to compare it to),

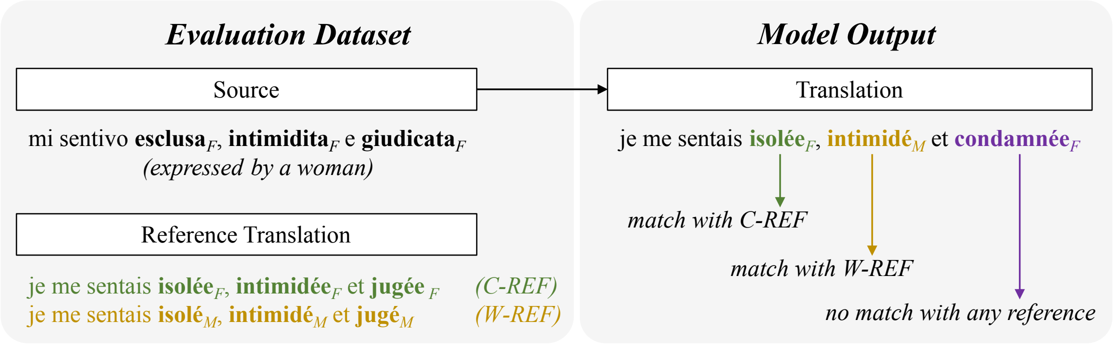

Evaluation Dataset
Model Output
Source
Translation
mi sentivo esclusa_F, intimidita_F e giudicata_F
(expressed by a woman)
Reference Translation
je me sentais isolée_F, intimidée_F et jugée_F (C-REF)
je me sentais isolé_M, intimidé_M et jugé_M (W-REF)
je me sentais isolée_F, intimidé_M et condamnée_F
match with C-REF
match with W-REF
no match with any reference

Figure 3.5: Illustration of the three possible translation outcomes of required gender preservation: The translation of a gender-inflected word either matched the correct reference translation *C-REF* (here, "isolée"), the wrong reference translation *W-REF* (here, "intimidé"), or neither (here, "condamnée").

forcing us to exclude these translations from our gender bias evaluation but not the assessment of models' translation quality (cf. 3.3.2).
Practically speaking, we compared a model's generated translations to the two sets of reference translations,

- the proper, *correct* reference translations (set of C-REF/*Correct* set), and
- references that were *wrong* regarding the gender-inflected words (set of W-REF/*Wrong* set), in that they were inflected for the opposite gender, i.e., feminine words were swapped with their masculine counterparts and vice versa,

measuring their similarity to each reference set using two different metrics.

BLEU-based Evaluation

The BLEU metric considers all words in a sentence, not just those that are inflected for gender. Therefore, it is worth emphasizing that the gender-marked words are the only distinguishing factor between the otherwise identical references. The idea behind gender-swapping (Bentivogli et al., 2020) is that a difference between the BLEU scores for the correct and the wrong references is attributed only to the gender-marked words, which were the focus of our gender preservation evaluation. If, for instance, a model wrongly generates a masculine translation for an originally feminine gender-inflected word, then this translation contributes to a higher BLEU score for the *Wrong* set. For this evaluation to work, all gender-inflected words appearing in the same sentence must agree with the same gender, i.e., all of them are inflected for the feminine gender, or all are inflected for the masculine gender.

Higher scores for the *Correct* set are better, while for the *Wrong* set, they should be as low as possible, since high *Wrong* scores indicate gender-biased behavior attributed to the inability to preserve gender, resulting in unfounded frequent gender confusion. It is worth noting that there is a partial correlation between scores for the *Correct* and *Wrong* set, as better translation of words not expressing gender, which are always part of a sentence, increases both the *Correct* and *Wrong* score; therefore, when aiming to improve the *Correct* BLEU scores, *Wrong* scores will always be greater than zero. Nevertheless, a greater difference between *Correct* and *Wrong* scores for the two reference sets is considered better, as it signals more correct than wrong gender translations (i.e., proper gender preservation).

To compute BLEU scores, we used sacreBLEU[3] (Post, 2018), which provides a fair and reproducible evaluation, as it operates on detokenized text.

Accuracy-based Evaluation

Since BLEU is designed to consider all words and not just those inflected for gender, we wanted to rely on an additional metric that only accounts for gender-inflected words. The metric measures how close a given set of gender-marked words produced by a model are to their target values, similar to the accuracy, but we had two sets of target values, the correct and the wrong gender-inflected target words, which we wanted to consider in relation to each other to assess gender biases. Therefore, to compute the metric, we compared the set of gender-marked words produced by a model to the two sets of target values and counted the number of matches with either set. In this way, we could evaluate gender preservation for each gender, feminine and masculine words, measured by how often a model was able to produce the correct gender (C) for those words that matched either the correct or the wrong reference set ($C + W$); we refer to this as *gender preservation* ($\alpha_{correct}$).

As mentioned before, we were forced to exclude those words that did not match any of the two reference sets (N) from our gender bias evaluation and only rely on "correct and wrong matches" ($C + W$); naturally, the larger this set was (i.e., the larger the sample size), the more significant our findings would be. To favor larger numbers of translations that matched either the correct or the wrong reference ($C + W$), we weighted $\alpha_{correct}$ by the size of C+W in relation to the number of all translations ($C + W + N$), matching a reference ($C + W$) or not matching any reference (N); we refer to this weighting factor as *sample size* (ρ).

Formally, we defined the metric $\gamma_{correct}$ to measure the *gender preservation performance weighted by the sample size*, that can be calculate for either gender (feminine/masculine) or both combined (feminine & masculine), as follows:

$$\gamma_{correct} = \underbrace{\frac{C}{C+W}}_{\alpha_{correct}} \cdot \underbrace{\frac{C+W}{C+W+N}}_{\rho} = \frac{C}{C+W+N} \tag{3.1}$$

Equation 3.1 shows that $\gamma_{correct}$ is essentially the accuracy, meaning the fraction of the total model translations inflected for the correct gender, which, if higher, is better. The equivalent calculation of γ_{wrong} can be made using $\alpha_{wrong} = \frac{W}{C+W}$ to measure how often the model produces the opposite, wrong gender instead.

To compare the performances for the two genders, we computed the *gender gap* $\delta_{correct}$ between results for the feminine (F) and the masculine (M) gender as formalized in Equation 3.2:

$$\delta = 1 - \frac{\min(\gamma^F_{correct}, \gamma^M_{correct})}{\max(\gamma^F_{correct}, \gamma^M_{correct})} \tag{3.2}$$

As a reflection of gender biases, gender gaps should be as small as possible and, ideally, zero, due to minimal differences between the results for the feminine and the masculine gender; in particular, we examined the *referent gender gap* $\delta_{referent}$ between results for feminine and masculine words (cf. Section 3.4.1 for details on referent gender) and the *speaker gender gaps* $\delta^F_{speaker}$ and $\delta^M_{speaker}$ between feminine and masculine results for female and male speakers (cf. Section 3.4.3 for details on speaker gender).

[3] https://github.com/mjpost/sacrebleu

3.3.2 Translation Quality

While the evaluation of gender bias in MT naturally focuses on the translation of gender-specific words, the overall translation quality of *both* gender-marked and gender-neutral words should not be ignored. A model that only performs well on gender-marked words but not on gender-neutral words is not desirable as it can not be deployed for general application. To substantiate the relevance of our bias evaluation, we complemented our results by assessing the overall translation quality of the different models using BLEU (i.e., sacreBLEU).

3.3.3 Probing Model Hidden Representations For Gender Signals

In principle, zero-shot translation balances abstraction and generalization, simplifying language representations to depict essential linguistic properties while resorting to common essential properties among different languages. Arguably, a model's ability to disambiguate gender in translation depends on preserving and rendering gender clues in language representations. Accordingly, when encouraging language-agnostic representations for the betterment of zero-shot translation, it is crucial to *maintain gender information.*

To validate whether improvements in gender translation indeed stemmed from better gender preservation — i.e., gender clues better captured in hidden representations — we assessed the difficulty of recovering gender-specific information before and after applying the different model modifications encouraging language-agnostic representations (cf. Section 3.2.3) for the considered bridge languages. Specifically, we trained a *classifier* on the output of a model's encoder *to predict the gender conveyed by a source token or the entire source sentence.* Similar prediction tasks have been used to analyze linguistic properties encoded in representations (Adi et al., 2017; Liu et al., 2021). The classifier operated on each time step (token level) or a combination of them (sentence level) using a linear projection from the embedding dimension to the number of classes C representing the genders.

Both, token-level and sentence-level classification, were interesting to us because removing residual connections in the encoder lessened the positional correspondence between encoder representations and tokens and, thereby, the correspondence between gender clues and individual words, or tokens for that matter; similar motivation held for the other modifications as they conceptually promote the generalizability of hidden representations. The idea behind token-level and sentence-level gender labels is illustrated in Figure 3.6.

For token-level gender classification, we assigned each token one of three labels (C=3), *feminine*, *masculine*, or *neuter.* For sentence-level classification, we first transformed the token embeddings by average pooling to obtain a sentence embedding; this embedding was fed to the classifier to predict the sentence gender, labeled as either *feminine* or *masculine* (C=2) depending on the gender-marked words that occurred in the sentence, which are exclusively associated with one of the two genders. Both, the token-level and the sentence-level classifier minimized the cross-entropy loss:

$$\mathcal{L}_{classifier} = -\sum_{c=1}^{C} y_c \cdot log(p_c), \tag{3.3}$$

where y_c is the true gender label, p_c is the probability predicted for the c^{th} gender class, with $C = 2$ for sentence-level and $C = 3$ for token-level classification.

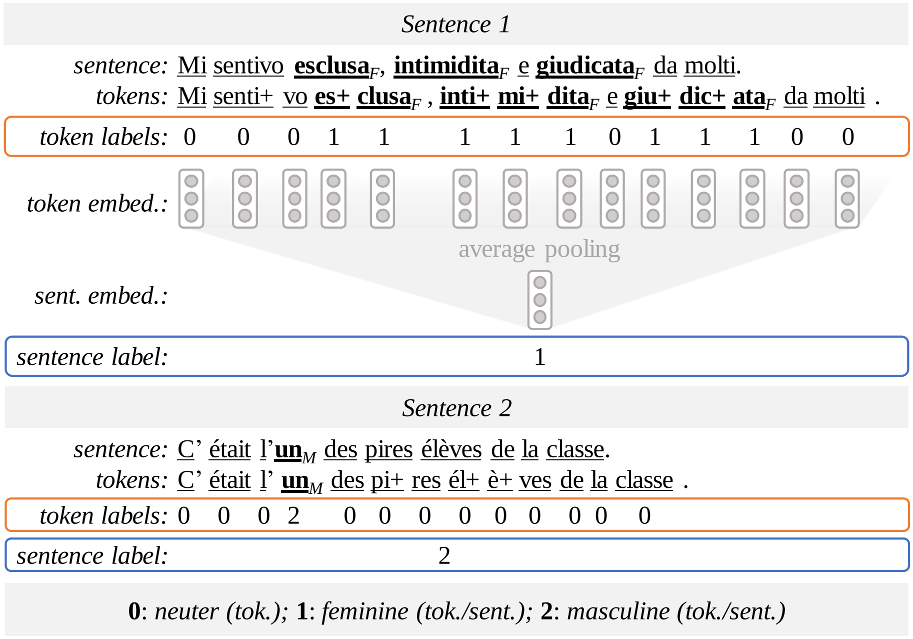

Sentence 1
sentence: Mi sentivo esclusaF, intimiditaF e giudicataF da molti.
tokens: Mi senti+ vo es+ clusaF , inti+ mi+ ditaF e giu+ dic+ ataF da molti .
token labels: 0 0 0 1 1 1 1 1 0 1 1 1 0 0
token embed.:
average pooling
sent. embed.:
sentence label: 1
Sentence 2
sentence: C' était l'unM des pires élèves de la classe.
tokens: C' était l' unM des pi+ res él+ è+ ves de la classe .
token labels: 0 0 0 2 0 0 0 0 0 0 0 0 0
sentence label: 2
0: neuter (tok.); 1: feminine (tok./sent.); 2: masculine (tok./sent.)

Figure 3.6: To assess the difficulty of recovering gender information before and after applying the different model modifications encouraging language-agnostic representations, we trained and tested a classifier to predict the gender conveyed by each token and (after average pooling of the token embeddings) each sentence. Token-level classification considered three classes, *0*: neuter, *1*: feminine, *2*: masculine; sentence-level classification included two classes, *1* and *2*.

3.4 Organization of Data

Generally, communication in languages with a rich inflectional system, such as French or Italian, requires keeping track of agreement features between words, one of which, across many languages, is gender.

3.4.1 Referent Gender

Grammatical gender agreement determines the modification of certain words to express gender congruent with the other words they relate to, which, in our case, were the words designating a *referent* — a person the speaker mentioned or to which the speaker referred. Consequently, the biological (or conceptual) gender of a referent determined the gender of those gender-marked words relating to the referent — i.e., for female referent using feminine words and for male referent using masculine words.

3.4.2 Utterance Categories

In our case, a referent was either the speaker himself or herself or a person not identified as the speaker (nor the addressee(s)/audience). This led us to distinguish between two categories of utterances with different gender phenomena: those with speaker-*related* gender agreement (category 1) and those with speaker-*independent* gender agreement (category 2).

Utterance Category 1: Speaker-Related Gender Agreement

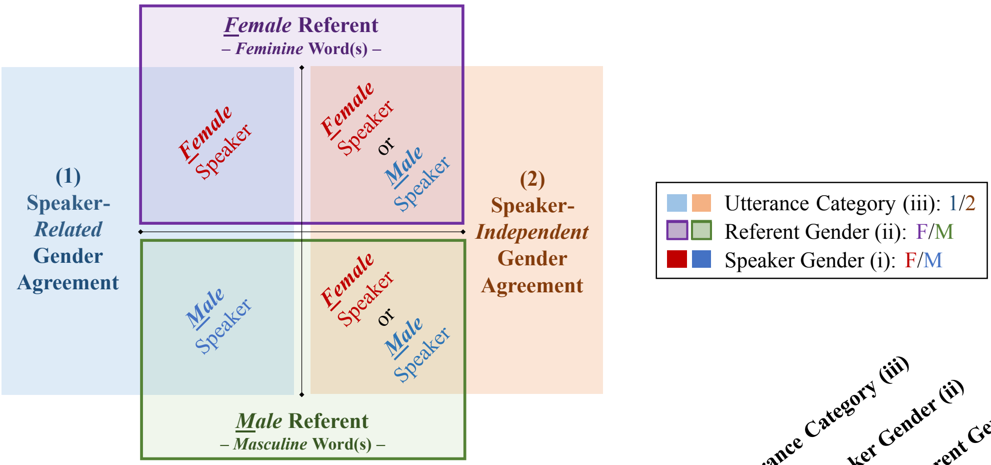
Whenever the speaker is the referent, i.e., the speaker is referring to himself (or herself), there is speaker-related gender agreement among those gender-marked words relating to the speaker. As a piece of speech, besides its linguistic form, an utterance has a context that provides extra-linguistic information, such as speaker-related characteristics (e.g., the speaker's gender), that can affect language comprehension (Hanulíková & Carreiras, 2015). Languages with less pronounced inflection of gender, such as English, can encounter syntactic structures that provide no indication about a speaker's gender. In these cases, extra-linguistic gender information conveyed by a speaker's voice can be imperative for the correct syntactic processing and translation of utterances with gender agreement dependent on the speaker's gender. In contrast, syntactic structures of languages with rich inflectional systems of gender typically encode enough information to unambiguously classify a speaker's gender. In Figure 3.7, example utterances (a) and (b) illustrate the difference between Italian and English sentences; in Italian, multiple words provide unambiguous clues about the speaker's gender, whereas, in English, there is no way of telling the gender of the speaker with certainty just from the sentence itself (without context).

Utterance Category 2: Speaker-Independent Gender Agreement

Whenever a person other than the speaker is the referent, i.e., the speaker is talking about someone else, there is speaker-independent gender agreement among those gender-marked words relating to the referent. In these cases, meaning construction typically does not require the integration of semantic information about the speaker for correct syntactic processing and translation. The gender-inflection of words is therefore often purely based on syntactic agreement with a formally marked subject (here, the referent), making the referent's gender identity explicit in our considered utterances. In these cases, gender violations are always apparent and identifiable by the disagreement or mismatch between the referent gender and gender-inflected words. In contrast, for utterances with speaker-related gender agreement (utterance category 1), gender inflection is incorrect only if the extra-linguistic information about the speaker is considered and integrated into the syntactic build-up of the utterance. In Figure 3.7, example utterances (c)–(f) illustrate that in Italian and in English utterances, there is always at least one word present, signaling the gender of the referent.

3.4.3 Speaker Gender

While the referent gender represents a system for gender distinction in natural language (e.g., grammatical gender), it is not the only gender-related factor that impacts spoken language: Prior research has found differences in language use, including both content and style of language, between men and women, which models can fail to reflect (Savoldi et al., 2021). For instance, Brownlow et al. (2003) have shown that the judicious use of pronouns and articles signals differences in how men and women chose to communicate: Women use more active, responsibility-oriented language including more self-referent pronouns, whereas men tend to use more third person pronouns and more articles as can be seen in passive, more depersonalized language constructions. These gender differences in language use are reflected in spoken as well as in written language (e.g., transcriptions of speech). Since each utterance in our evaluation data originated from either a female or a male speaker, we considered not only differences in referent gender but also speaker gender.


Female Referent
– Feminine Word(s) –
Female Speaker
Female Speaker or Male Speaker
(1) Speaker-Related Gender Agreement
(2) Speaker-Independent Gender Agreement
Male Speaker
Female Speaker or Male Speaker
Male Referent
– Masculine Word(s) –
Utterance Category (iii): 1/2
Referent Gender (ii): F/M
Speaker Gender (i): F/M

	Italian Utterance (with English Translation)	Utterance Category (iii)	Speaker Gender (ii)	Referent Gender (i)
(a)	Mi sentivo **esclusa**$_F$, **intimidita**$_F$ e **giudicata**$_F$ da molti. (“I felt alienated, intimidated, and judged by many.”)	1	F	F
(b)	Ero **uno**$_M$ **studente**$_M$ **ossessivo**$_M$ **compulsivo**$_M$. (“I was an obsessive compulsive student.”)	1	M	M
(c)	Sua madre l'aveva **cresciuta**$_F$ da **sola**$_F$. (“**Her**$_F$ mother raised **her**$_F$ alone.”)	2	F	F
(d)	Era **uno**$_M$ **degli**$_M$ **studenti**$_M$ peggiori della sua classe. (“**He**$_M$ was one of the worst students in class.”)	2	F	M
(e)	Era **dura**$_F$, era forte, era potente. (“**She**$_F$ was tough, **she**$_F$ was strong, **she**$_F$ was powerful. ”)	2	M	F
(f)	E **questo**$_M$ **vice-sceriffo**$_M$ saltò su e travolse l'uomo di colore. (“And this deputy jumped up and **he**$_M$ ran over to this older black man.”)	2	M	M

	French Reference Translations (C-REF: Correct, W-REF: Wrong)	
(a)	C-REF W-REF	Je me sentais **isolée**$_F$, **intimidée**$_F$ et **jugée**$_F$ par beaucoup. Je me sentais **isolé**$_M$, **intimidé**$_M$ et **jugé**$_M$ par beaucoup.
(b)	C-REF W-REF	J’etais **un**$_M$ **étudiant**$_M$ **obsessif**$_M$ **compulsif**$_M$. J’etais **une**$_F$ **étudiante**$_F$ **obsessive**$_F$ **compulsive**$_F$.
(c)	C-REF W-REF	Sa mère l'a **élevée**$_F$ **seule**$_F$. Sa mère l'a **élevé**$_M$ **seul**$_M$.
(d)	C-REF W-REF	C'était l'**un**$_M$ des pires élèves de la classe. C'était l'**une**$_F$ des pires élèves de la classe.
(e)	C-REF W-REF	Elle était difficile, elle était **forte**$_F$, elle était **puissante**$_F$. Elle était difficile, elle était **fort**$_M$, elle était **puissant**$_M$.
(f)	C-REF W-REF	**Cet**$_M$ **adjoint**$_M$ s'est **levé**$_M$ d'un coup et s'est **précipité**$_M$ vers **cet**$_M$ homme **âgé**$_M$ **noir**$_M$. **Cette**$_F$ **adjointe**$_F$ s'est **levée**$_F$ d'un coup et s'est **précipitée**$_F$ vers **cette**$_F$ homme **âgée**$_F$ **noire**$_F$.

Figure 3.7: Studied gender phenomena include (i) the *gender of a referent* mentioned that determined the gender of gender-marked words and (ii) the *speaker’s gender*. The referent could be (1) the speaker or (2) a person other than the speaker; depending on the referent, an utterance had (1) speaker-related or (2) speaker-independent gender agreement (cf. *utterance category* (iii)). Explicit gender clues in depicted example utterances are underlined and gender-marked words in corresponding reference translations are italics.

3.4.4 Summary of the Data's Evaluative Dimensions

Altogether, our evaluation data covered three evaluative dimensions that we considered in our experiments, for each of which an utterance complied with one of two of the following specifications, also illustrated in Figure 3.7 with example utterances and references:

(i) the *gender of the referent*, the person the speaker refers to, that determines the *gender of gender-marked words* relating to the referent:

  - *female* referent with *feminine* word forms, or
  - *male* referent with *masculine* word forms;

(ii) the *gender of the speaker*:

  - *female* speaker, or
  - *male* speaker;

(iii) the *utterance category* reflecting the type of gender agreement as described:

  (1) *speaker-related gender agreement*, where the speaker is the referent, or

  (2) *speaker-independent gender agreement*, where the referent is someone other than the speaker.

It is worth mentioning that the two utterance categories are equivalent to the two groups of MuST-SHE instances between which the authors of the dataset (Bentivogli et al., 2020) distinguish, based on the source of information — audio recording or utterance content — necessary to disambiguate gender (see Section 2.2.2). This original distinction did not hold in our experimental setting because, on the one hand, we did not provide any audio recordings to the models and, on the other hand, we considered different translation directions, for which the utterance content, in French and Italian, always provided gender clues as we focused on the challenge of gender preservation in translation in particular to evaluate gender bias in multilingual MT.

4

Technical Details

This chapter documents the setup of the experiments conducted in this thesis. We describe the data preparation and preprocessing procedures for both the training and evaluation datasets in Section 4.1 and 4.2, provide details about the model training including the hyperparameter configuration in Section 4.3, and describe the inference phase for zero-shot and for pivot-based translation in Section 4.4. Figure B.1 in the appendix provides a visual overview of the experimental setup. The source code used to conduct the experiments is available on GitHub[1].

4.1 Datasets

In our experiments, we used the publicly available corpora MuST-C (Di Gangi et al., 2019) and MuST-SHE (Bentivogli et al., 2020). Specifically, we trained our models on different subsets of MuST-C and evaluated them primarily on a subset of MuST-SHE and, in one instance, on the test set of MuST-C.

4.1.1 Training

For each target language, MuST-C comprises audio recordings from English TED Talks, each of which is aligned at the sentence level with a transcription and a translation. While MuST-C is a multilingual corpus characterized by various speakers, it is worth noting that, like many other corpora, it reflects a gender-imbalanced distribution of only 30% female versus 70% male speakers. The release of version 1.2 of MuST-C includes 14 English-to-X language directions. In our experiments, we included 10 of the 15 available languages from three different language families (Slavic$^{\mathrm{S}}$, Romance$^{\mathrm{R}}$, and Germanic$^{\mathrm{G}}$) in our training data[2] — Czech$^{\mathrm{S}}$ (cs), Dutch$^{\mathrm{G}}$ (nl), English$^{\mathrm{G}}$ (en), French$^{\mathrm{R}}$ (fr), German$^{\mathrm{G}}$ (de), Italian$^{\mathrm{R}}$ (it), Portuguese$^{\mathrm{R}}$ (pt), Romanian$^{\mathrm{R}}$ (ro), Russian$^{\mathrm{S}}$ (ru), and Spanish$^{\mathrm{R}}$ (es) — to which, in short, we refer to as Y. Besides belonging to different language families, the languages in Y vary in their gender-inflectional systems.

To investigate the impact of the bridge language, determined by the language pairs included during training, we formed three training corpora that are subsets of MuST-C with language pairs en↔Y\en, de↔Y\de, and es↔Y\es, where Y\en is the language set Y excluding English. On each of the three corpora, we trained a model and afterward evaluated the three trained

[1] https://github.com/lenacabrera/NMTGMinor

[2] The five remaining MuST-C target languages (in X) were excluded from our experiments since they required different preprocessing procedures and the 10 considered languages (Y) provided a sufficiently large enough database to train our models while keeping the training time at a minimum.

models on our evaluation data. While different model weight initialization can lead to different results, we trained each model only once because of the extensive training times and the number of evaluated models. Since only a portion ($\sim$ 10%) of MuST-C is true-parallel data, i.e., the same sentence is translated in multiple languages, the training corpora with different bridge languages differed in size, as specified in Table 4.1; the en↔Y\en training data comprised more sentences than the other two (which is a true reflection of real-world data availability).

Table 4.1: Overview of the three MuST-C subsets used in the experiments.

Language Pairs	# Sentences per Direction
en ↔ Y\en = {cs, de, es, fr, it, nl, pt, ro, ru}	125,000–267,000
de ↔ Y\de = {cs, en, es, fr, it, nl, pt, ro, ru}	103,000–223,000
es ↔ Y\es = {cs, de, en, fr, it, nl, pt, ro, ru}	102,000–258,000

4.1.2 Evaluation

We evaluated our models on a modified version of MuST-SHE, which is described in detail in Sections 2.2.2 and 3.4. Since MuST-SHE is a subset of MuST-C, we removed overlapping sentences and kept only true-parallel data to obtain a "multiway" version of MuST-SHE unseen to the models. In total, we obtained 278 sentences. Detailed statistics about these sentences are presented in Table 4.2.

Table 4.2: Statistics of the multiway MuST-SHE data used in the experiments. Instances are split into groups according to the referent gender (Feminine/Masculine), gender agreement (Category 1: speaker-related, Category 2: speaker-independent), and speaker gender (Female/Male).

	Feminine		Masculine		**Total**	
	(Female/Male)		(Female/Male)		(Female/Male)	
Category 1	64	(64/0)	56	(0/56)	120	(64/56)
Category 2	72	(58/14)	86	(27/59)	158	(85/73)
Total	136	(122/14)	142	(27/115)	**278**	(149/129)

Using training corpora comprising different language pairs, we built models with different supervised translation directions. Accordingly, the models did not share the same zero-shot directions. For instance, a model trained on the es↔Y\es corpus had seen instances for language pairs including Spanish. Therefore, we did not include translations from and to Spanish in our evaluation to ensure equal zero-shot directions across all models considered in our experiments. Consequently, we evaluated models on fr ↔ it directions, which left us with 556 translations included in our evaluation.

It is worth mentioning that this subset of MuST-SHE was not ideally gender-balanced in terms of speakers' gender for utterances with speaker-independent gender agreement (category 2); due to the different requirements for the evaluated experimental setting (including only zero-shot directions, removing instances overlapping with the training data, etc.), this constraint was under caution allowed for as further restriction of the evaluation data would have caused it to be reduced to a significantly smaller amount, unsuitable of attesting the significance of made observations.

For our gender classification, we augmented the original MuST-SHE utterances (including the source languages French, Italian, and Spanish)[3] with two sets of labels. For sentence-level classification, we used the information about the gender of the gender-marked words in an utterance provided in MuST-SHE, which are either exclusively feminine (1552 utterances) or masculine (1562). For token-level classification, we augmented each sentence after tokenization with a set of labels the size of the number of tokens in the sentence, assigning each token one of three labels signaling the grammatical gender – neuter (113,934 tokens), feminine (5425), masculine (5431) – of the word the token is part of. We only inspected those results conforming with the gender binary (feminine and masculine) as those are the gender signals we were interested in examining.

4.2 Preprocessing

MuST-C comes with partitioned training, validation, and test sets which we kept unchanged in our experiments, with the additional modifications described above (cf. Section 4.1.1). For each training corpus, including the training and the validation set, we first performed tokenization and truecasing (restoring proper capitalization of words) using the Moses[4] tokenizer and truecaser. Afterwards, we learned BPE using subword-nmt[5] (Sennrich et al., 2016) to create a subword-based vocabulary. We performed 20,000 merge operations on the dataset and only used tokens occurring in the training set with a minimum frequency of 50 times. Our evaluation data was preprocessed in a similar way using the BPE-learned vocabulary.

4.3 Training

For our experimental implementation, we used the sequence-to-sequence toolkit NMTGMinor[6]. Our baseline was a Transformer with 5 encoder and 5 decoder layers with 8 attention heads, an embedding size of 512, and an inner size of 2048. For regularization, we used dropout with a rate of 0.2 and performed label smoothing with a rate of 0.1. Moreover, we used the Adam optimizer (Kingma & Ba, 2014) with the learning rate schedule from Vaswani et al. (2017) with 8,000 warmup steps. The source and target word embeddings were shared. Furthermore, in the decoder, the parameters of the projection from the hidden states to the vocabulary were tied with the transposition of the word lookup table. To specify the output language, we used a target-language-specific begin-token as well as language embeddings concatenated with decoder word embeddings, similar to Pham et al. (2019) and Liu et al. (2021).

As part of our model modifications motivated in Section 2.1.6, we removed a residual connection in the third encoder layer. We trained each model for 64 epochs and averaged the weights of the five best checkpoints ordered by the validation loss. For the auxiliary similarity loss and the adversarial language classifier, we resumed training of the baseline and the model with removed residual connections for 10 additional epochs. By default, we only included the supervised translation directions in the validation set.

When analyzing a model's hidden representations through gender classification performance, we froze the trained encoder-decoder weights and trained the classifier for 100 epochs. The

[3] For the classification, we used the MuST-SHE utterances without removing instances overlapping with MuST-C to enlarge the classifier training and validation data on which we evaluate the classifier, similar to Liu et al. (2021).

[4] https://github.com/moses-smt/mosesdecoder

[5] https://github.com/rsennrich/subword-nmt

[6] https://github.com/nlp-dke/NMTGMinor

classifier used a linear projection from the encoder hidden dimension to the number of classes, followed by a softmax activation. As the classification task was lightweight and convergence was fast, we reduced the warmup steps to 400 while keeping the learning rate schedule unchanged.

4.4 Inference

We compared the translation performance between direct zero-shot translation (implicit bridging with a single model) and pivot-based translation (explicit bridging with cascading two identical models). Since models were trained for multiple translation directions, we provided a beginning-of-sentence (BOS) token to indicate the desired target language. For instance, for French→Italian translation, we indicated the target language by prepending the BOS *IT* to source sentences when performing zero-shot translation. Meanwhile, in the case of pivoting, we first provided the model the BOS token for the pivot language (e.g., *EN* in the case of en↔Y\en training data, or *DE* in the case of de↔Y\de training data) and generated the translation for the pivot language. Afterward, we prepended the BOS token *IT* to the produced translations and fed them back to the model to generate the final output in Italian. Cascading of the model in such manner is illustrated in Figure B.1 (see illustration for "Inference").

5

Results

This chapter presents the results of the experiments performed in this thesis and the discussions thereof. The series of experiments was largely motivated by our central two-part hypothesis that, *on average,* (1) *zero-shot translation generates fewer gender-biased outputs than pivot-based translation, where gender bias is conceived as the systematic and unfair discrimination against a group of individuals of the same sex, here either women or men, in favor of the other gender group,* (2) *while maintaining comparable translation quality.* Looking into both parts separately, we first confirm the reasonable overall translation quality of zero-shot and pivot-based translations and evaluate the gender preservation performances of both approaches afterward.

Primarily, we examine the performance of models bridging via English, i.e., trained on en ↔ Y\en MuST-C language pairs, focusing on the effect of the modifications promoting language-agnostic hidden representations (cf. Section 3.2.3). We conduct a BLEU-based evaluation in Section 5.1 using gender-swapping and perform an additional analysis based on the accuracy, only considering gender-marked target words, in Section 5.2. Secondly, we study the effect of the bridge language in Section 5.3. We compare the performances of models bridging via English to that of models bridging via German or Spanish. Thirdly, in Section 5.4, we measure the extent of gender signals captured in language representations indicated by the models' ability to recover gender information from encoder hidden states using a gender classifier on sentence and on token level.

In the presentation of our results, we use B to refer to the baseline model, R to indicate models with a removed residual connection, AUX_{SIM} represents the auxiliary similarity loss, and ADV_{LAN} the adversarial language classifier. When space is limited, we use ZS and PV to abbreviate the terms zero-shot and pivot-based translation, and $\widehat{en}$ to refer to English-bridging models. We present MuST-SHE results as the average of the scores obtained for the two evaluated zero-shot directions fr ↔ it. Since the discussion of all results obtained in the course of this thesis is beyond the scope, we present only a selection of the results. For the sake of completeness, in their entirety, the results can be found in the appendix B.3.

5.1 BLEU-based Evaluation

In the first part of our evaluation, we examine BLEU scores to evaluate the models' translation performance with a focus on their ability to preserve gender.

5.1.1 Translation Quality

Since sentences always comprise gender-neutral and gender-marked words, a model that only performs well on one of the two is not desirable as it can not be deployed for general application. Therefore, we evaluate the quality of models' translations of the entire utterances, including gender-neutral and gender-marked words, on two datasets: the MuST-C test set and on MuST-SHE (i.e., the *Correct* references, cf. Section 3.3.1). The corresponding results, depicting average BLEU scores for zero-shot directions on either dataset, are presented in Table 5.1.

Table 5.1: Average BLEU scores for zero-shot directions on the MuST-C test set ($\{s \to t | s, t \in \{\text{nl, pt, ro, ru}\} \ \& \ s \neq t\}$) and on MuST-SHE (fr ↔ it). Higher scores are better, the bold score per dataset denotes the best performance of each approach, and underlined are the best of both approaches if one is better.

Model	**MuST-C**		**MuST-SHE**	
$(\widehat{en})$	Zero-Shot	Pivot	Zero-Shot	Pivot
Baseline (B)	3.7	$\underline{\mathbf{17.4}}$	3.8	**25.6**
B + AUX$_{SIM}$	11.1	15.8	15.7	24.0
B + ADV$_{LAN}$	15.9	17.0	25.3	25.0
Residual (R)	14.3	17.2	22.9	24.9
R + AUX$_{SIM}$	15.1	15.6	23.4	23.4
R + ADV$_{LAN}$	**16.0**	17.0	**25.6**	25.4
AVG excl. B	14.5	16.5	22.6	24.5

Starting with our baseline B, we observe far higher BLEU scores for pivoting than for zero-shot translation on both datasets. Adding the auxiliary similarity loss or the adversarial language classifier to the baseline ($B + AUX_{SIM}$, $B + ADV_{LAN}$) improves the zero-shot performance considerably. On MuST-SHE, zero-shot translation with $B + ADV_{LAN}$ even performs slightly better than pivoting using the same model. Removing a residual connection (R/R+) also improves zero-shot translation significantly. On MuST-SHE, we observe an increase of 19.1 in the BLEU score when simply removing a single residual connection from B. Furthermore, combining R with AUX_{SIM} or ADV_{LAN} additionally improves the BLEU score, leading to comparably good zero-shot ($R + ADV_{LAN}$) and pivot-based (B) translation performances in the case of MuST-SHE. Also, it is $R + ADV_{LAN}$ that achieves the highest zero-shot translation scores on both datasets; second-best, not far behind, is $B + ADV_{LAN}$. Hence, ADV_{LAN} achieves the largest performance gains for zero-shot translation.

Generally, as Liu et al. (2021) have showcased before, the results confirm that encouraging language-agnostic representations improves zero-shot translation; this is well summarized by the larger average scores across the five modified models (AVG excl. B) compared to those of B (e.g., $22.6 > 3.8$ for MuST-SHE). Meanwhile, the modifications have a negative effect on pivoting; this is not surprising since more language-independent representations caused by the modifications presumably lead to more errors due to the loss of information along the pivoting translation pipeline. Concerning our hypothesis, we see that zero-shot translation can "[maintain] comparable translation quality" (e.g., $R + ADV_{LAN}$ on MuST-SHE), albeit more often performing not as well as pivoting. Nevertheless, both approaches achieve reasonable translation quality.

5.1.2 Gender Preservation

After examining the translation quality, we continue our evaluation, assessing the models' ability to preserve the gender of gender-marked words: In Table 5.2, we compare the BLEU scores indicative of the similarity of the generated translations of MuST-SHE utterances to the *Correct* references (from Table 5.1) and their gender-reversed counterparts (*Wrong* references), regardless of the gender of gender-marked words appearing in the utterances (i.e., results include both genders). In principle, for a comprehensive model assessment, hereafter, the results per row must be considered in combination (e.g., a high *Correct* score is invalidated to some extent by a large *Wrong* score, indicated by a negative delta of the scores for the two reference sets).

Table 5.2: Average BLEU scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

Model $(\widehat{en})$	*Correct* ↑		*Wrong* ↓		Delta ↑	
	ZS	PV	ZS	PV	ZS	PV
Baseline (B)	3.8	**25.6**	3.8	23.6	0	**2.0**
B + AUX$_{SIM}$	15.7	24.0	14.3	22.3	1.4	1.8
B + ADV$_{LAN}$	25.3	25.0	23.2	23.2	**2.1**	1.8
Residual (R)	22.9	24.9	21.2	23.0	1.7	1.9
R + AUX$_{SIM}$	23.4	23.4	21.4	21.6	2.0	1.9
R + ADV$_{LAN}$	**25.6**	25.4	23.7	23.5	1.9	1.9
AVG excl. B	22.6	24.5	20.8	22.7	1.8	1.9

Our previous analysis has revealed significant zero-shot performance gains of the different model modifications. Comparing the scores for the *Correct* and the *Wrong* reference set shows that the improved performance for the former is always reflected in the latter, which in itself is not unexpected due to the partial correlation between scores for the *Correct* and *Wrong* set (cf. Section 3.3.1). The level of improvement for both reference sets is, however, not the same: We consistently observe positive deltas (=*Correct*−*Wrong*)—i.e., more correct gender translations than gender-reversed (wrong) ones—as well as increased deltas above zero when comparing *B* (has delta of zero) with the other models for zero-shot translation (note, higher delta is better). Accordingly, all model modifications improve zero-shot models' ability of choosing the correct *instead* of the opposite, wrong gender resulting in reduced gender-biased outputs, as the model's translations less frequently ignore and contradict the gender information conveyed by the source sentence. Conversely, we observe the opposite effect (i.e., decreasing deltas) for pivoting with modified models compared to *B*.

Furthermore, it shows that the model with the highest score for the *Correct* set ($R + ADV_{LAN}$, BLEU of 25.6) is not the least biased ($B + ADV_{LAN}$, delta of 2.1, and second-best *Correct* score of 25.3). In contrast, the best pivoting performance accomplishes *B*, also achieving the overall highest BLEU score of 25.6 for the *Correct* set and a delta of 2.0. Despite the noticeable zero-shot performance gains of the model modifications, these results indicate comparable gender preservation performance of both pivoting and zero-shot translation.

Breaking down the results into those for feminine and masculine referents (i.e., feminine and masculine words) in Table 5.3 reveals a notable difference in BLEU scores between both genders, showcasing better masculine than feminine gender preservation as a result of models' biases toward producing masculine outputs more often than ideally expected. Taking a closer look at the BLEU scores for the *Correct* set, it shows that zero-shot translation with $R + ADV_{LAN}$ achieves the highest feminine score (24.5), outperforming pivoting (with B) by 0.7. On masculine words, B produces the best pivoting result (27.1), which is 0.5 scores higher than the best zero-shot translation result achieved by $B + ADV_{LAN}$. It turns out that neither translation approach performs best for both feminine and masculine words.

Table 5.3: Average BLEU scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references, broken down by referent gender and complemented with the scores' difference (Delta ↑ = *Correct*− *Wrong*). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches.

Model $(\widehat{en})$	**Feminine** Referent						**Masculine** Referent					
	Correct ↑		*Wrong* ↓		Delta ↑		*Correct* ↑		*Wrong* ↓		Delta ↑	
	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
Baseline (B)	3.2	**23.8**	3.3	23.4	-0.1	**0.4**	4.3	$\underline{\textbf{27.1}}$	4.1	23.8	0.2	3.3
B + AUX$_{SIM}$	15.6	21.8	14.7	21.4	$\underline{\textbf{0.9}}$	**0.4**	15.7	25.8	13.8	22.8	1.9	3.0
B + ADV$_{LAN}$	23.5	22.8	23.4	22.8	0.1	0.0	**26.6**	26.7	23.0	23.3	$\underline{\textbf{3.6}}$	**3.4**
Residual (R)	21.7	23.0	21.6	22.9	0.1	0.1	23.9	26.4	20.7	23.1	3.2	3.3
R + AUX$_{SIM}$	22.3	21.4	21.5	21.2	0.8	0.2	24.2	24.8	21.1	22.0	3.0	2.8
R + ADV$_{LAN}$	$\underline{\textbf{24.5}}$	23.5	24.2	23.2	0.3	0.3	26.2	26.8	23.2	23.6	3.0	3.2
AVG excl. B	21.5	22.5	21.1	22.3	0.4	0.2	23.3	26.1	20.4	23.0	2.9	3.1

While for *Correct* references, each approach performs better for one gender than the other approach (hence superiority is balanced), the deltas between the *Correct* and the *Wrong* scores indicate better zero-shot performance for both genders. The gap between the approaches' highest (best) deltas is slightly bigger for feminine words ($B + AUX_{SIM}$: $|0.9 - 0.4| = 0.5$) than for masculine words ($B+ADV_{LAN}$: $|3.6-3.4| = 0.2$); hence, on feminine words, zero-shot translation has even fewer gender mix-ups compared to pivoting than on masculine words, where the deltas are very similar. In other words, zero-shot translation has a bigger lead over pivoting regarding preventing gender confusion on feminine words than on masculine words. This is also emphasized by the larger average scores across the five modified models (AVG excl. B) for feminine words (Delta: $0.4 > 0.2$), whereas for masculine words, pivoting achieves, on average, greater results than zero-shot translation ($3.1 > 2.9$). Regarding the model modifications, it stands out that AUX_{SIM} (with B or R) improves zero-shot models' feminine gender preservation most effectively, as indicated by their noticeably higher feminine deltas.

At large, results for both translation approaches share that in the masculine case, deltas are significantly higher than in the feminine case, where deltas are often close to zero. It shows that for feminine words, where the required gender realization is feminine, faulty masculine outputs are essentially almost as frequent as feminine outputs (hence, deltas close to zero); combined with the much higher masculine deltas, this observation suggests that models more frequently produce the masculine gender, and not only when it is required (i.e., mistakenly also for feminine words). Since correct feminine translations are consistently less frequent than their masculine counterparts across all models and approaches, we argue that better feminine gender preservation—as achieved with zero-shot translation—is more desirable in the aim of mitigating gender biases by closing this gender gap.

Summary of Preliminary Findings

The BLEU-based evaluation has offered preliminary insight into the gender preservation behavior of the considered models. For all modifications to our baseline, we observe performance improvements; this strengthens the idea that promoting language-independent representations improves the translation quality and reduces gender biases in zero-shot translation. When comparatively inspecting feminine and masculine performances, we observe a better performance of zero-shot compared to pivot-based translation for feminine words and vice versa for masculine words for the *Correct* references. Zero-shot models prove to be slightly better at preventing gender confusion (i.e., occurs when producing the wrong, gender-reversed translation). Nonetheless, for both approaches we can observe a notable masculine bias discriminating feminine words.

Altogether, the results of this preliminary analysis fail to reject our hypothesis, but rather support it. We find that zero-shot translation can maintain comparable translation quality but more often performs worse than pivoting. Furthermore, the results illustrate marginal superiority of zero-shot translation over pivoting regarding gender preservation, which is, however, not to be considered significant. Since BLEU is only an indirect and coarse indicator of better or worse gender preservation performance, we find it has limitations regarding gender bias evaluation which we demonstrate in the Appendix B.1.

5.2 Accuracy-based Evaluation

While the evaluation of BLEU scores has already shed light on the models' ability to preserve gender, it is insightful to also perform a more discriminative analysis, pointing only to the actual words through which gender is expressed. Compared to BLEU, the accuracy is a more transparent metric for our evaluation since better or worse performance measured is reliably attributed to better or worse translation of gender-inflected words only. Complementary to the BLEU-based evaluation, we examine the differences between zero-shot and pivot-based translations' accuracies for gender-marked words and compare the accuracies for the feminine and masculine gender.

For the *Correct* MuST-SHE references, Table 5.4 presents the accuracies ($\gamma_{correct}$, cf. Section 3.3.1) of the translated gender-marked words, regardless of the referent gender (All), as well as broken down by referent gender (Feminine, Masculine); additionally, it includes the magnitude of the gender gap ($\delta_{referent}$, cf. Section 3.3.1), which, when large is interpreted as bias toward better preservation of one gender at the expense of the other. Higher accuracies are a sign of better gender preservation performances, whereas smaller gender gaps between accuracies for the two genders indicate more fairly balanced preservation of both genders and thus fewer gender biases. We highlight the best results per metric (i.e., per two columns) for better overview, but it is worth noting that, similar to the BLEU-based evaluation, the metrics have to be considered in combination.

First of all, it shows that our tested models are able to preserve the gender of gender-inflected words adequately in approximately 40% of the cases (see All). Consistent with our BLEU-based evaluation is the masculine bias that is found throughout all models' performances, indicated by higher masculine than feminine results. However, compared to the BLEU scores, the difference between the accuracy scores for feminine and masculine words is much more noticeable, revealing a stronger masculine bias than derivable from the BLEU-based evaluation.

Comparing the performance of the baseline B to the average of all other models (AVG excl. B), it is confirmed that zero-shot translation significantly benefits from the different modifications promoting language-agnostic hidden representations (e.g., gender gap shrinks from 0.69 to 0.44), while pivoting performs better using B (gender gap of 0.47 is unchanged but with higher accuracies for B). The results also confirm the previous finding that $B + ADV_{LAN}$ and $R + ADV_{LAN}$

Table 5.4: Average accuracy scores for *Correct* MuST-SHE references ($\gamma_{correct}$, higher ↑ is better), broken down by referent gender and complemented with the referent gender gap ($\delta_{referent}$, lower ↓ is better). Bold scores denote the best results per approach and underlined are the best of both approaches.

Model ($\widehat{en}$)	All ↑		Feminine ↑		Masculine ↑		Gender Gap ↓	
	ZS	PV	ZS	PV	ZS	PV	ZS	PV
Baseline (B)	6.4	**42.5**	3.0	29.1	9.7	55.4	0.69	0.47
B + AUX$_{SIM}$	30.8	40.0	23.4	27.0	37.8	52.5	**0.38**	0.49
B + ADV$_{LAN}$	**43.0**	41.9	27.5	26.6	**57.8**	**56.5**	0.52	0.53
Residual (R)	38.8	42.4	25.9	29.8	51.1	54.5	0.49	0.45
R + AUX$_{SIM}$	41.1	39.2	30.0	28.0	51.7	49.9	0.42	**0.44**
R + ADV$_{LAN}$	**43.0**	**42.5**	**32.1**	**30.0**	53.4	54.5	0.40	0.45
AVG excl. B	39.3	41.2	27.8	28.3	50.4	53.6	0.44	0.47

achieve the highest accuracies for zero-shot translation (All: 43.0), with $B{+}ADV_{LAN}$ also achieving the best masculine result (57.8) and $R + ADV_{LAN}$ achieving the best feminine result (32.1), all of which are better than the best pivoting results. Regarding the gender gap, it shows that $R + ADV_{LAN}$ achieves closer accuracies for both genders, i.e., a smaller gender gap (0.40) than $B + ADV_{LAN}$ (0.50). This emphasizes that with already significantly higher masculine than feminine results, the feminine results and their improvement specifically are a decisive factor when aiming to reduce the gender gap and with it models' gender biases.

Zero-shot translation using $R{+}ADV_{LAN}$ outperforms pivoting on feminine words (by 2.1 for $R{+}ADV_{LAN}$, and by 3.0 compared to pivoting with B), but pivoting achieves higher masculine scores (higher by 1.0 for $R + ADV_{LAN}$ and by 2.0 using B). In comparison, the larger lead on feminine words makes up for the smaller lag on masculine words, making the gender gap for zero-shot translation using $R{+}ADV_{LAN}$ slightly smaller than for pivoting (smaller by 0.05 using $R + ADV_{LAN}$ and smaller by 0.07 compared to pivoting with B).

Consulting the complementary results for the *Wrong* references, presented in Table 5.5, supports the observation that the betterment of feminine gender preservation through modified zero-shot models more than compensates the slightly worse masculine performance: Comparing the deltas between the accuracies for the *Correct* and the *Wrong* set of $R + ADV_{LAN}$ for both genders shows a larger (better) delta for zero-shot translation than for pivoting ($0.5 > -4.9$), while pivoting achieves a masculine delta larger by a smaller magnitude ($43.9 > 43.0$). In contrast to the BLEU-based evaluation, these outcomes show that for feminine words, where the required gender realization is feminine, pivoting produces faulty masculine outputs more often than feminine outputs; hence, the negative feminine delta for pivot-based translation (not only for $R + ADV_{LAN}$ but for all models). In comparison, zero-shot models achieve better but still relatively poor feminine gender preservation results, as also revealed by the BLEU scores. In the masculine case, the accuracies for the *Wrong* set are much lower; combined with the much higher masculine *Correct* scores (cf. Table 5.4), the masculine deltas are positive and well above zero with the highest deltas of 49.2 and 49.0 for zero-shot and pivot-based translation respectively (both $B + ADV_{LAN}$).

Altogether, the results show once more the big difference between both genders, and reveal that models more frequently generate the masculine gender, and not only when it is required, leading to better masculine than feminine gender preservation. Overall, it seems that zero-shot translation yields slightly more equitable gender preservation results than pivoting, with

Table 5.5: Average accuracy scores for *Wrong* (γ_{wrong}, lower ↓ is better) MuST-SHE references, broken down by referent gender and complemented with the difference Delta ↑ = *Correct*− *Wrong* (cf. *Correct* scores in Table 5.4). For the Deltas, bold scores denote the best results per approach, underlined are the best of both approaches.

Model ($\widehat{en}$)	**Feminine** Referent				**Masculine** Referent			
	Wrong ↓		Delta ↑		*Wrong* ↓		Delta ↑	
	ZS	PV	ZS	PV	ZS	PV	ZS	PV
Baseline (B)	5.5	36.9	-2.5	-7.8	2.2	9.1	7.5	46.3
B + AUX$_{SIM}$	20.2	34.2	**3.2**	-7.2	7.1	8.3	30.7	44.2
B + ADV$_{LAN}$	34.3	39.2	-6.8	-12.6	8.6	7.5	**49.2**	**49.0**
Residual (R)	30.8	35.9	-4.9	-6.1	9.1	10.0	42.0	44.5
R + AUX$_{SIM}$	27.6	34.6	2.4	-6.6	8.9	9.9	42.8	40.0
R + ADV$_{LAN}$	31.6	34.9	0.5	**-4.9**	10.4	10.6	43.0	43.9
AVG excl. B	28.9	35.8	-1.1	-7.5	8.8	9.3	41.5	44.3

R+ADV$_{LAN}$ accomplishing the overall best result (highest feminine accuracy and second lowest gender gap); pivoting, on the other hand, achieves better masculine gender preservation which, at the already significantly lower feminine gender preservation performance contributes to larger gender gaps that indicate more gender inequality.

5.2.1 Distinguishing Between Female & Male Speakers

Our evaluation, so far, has revealed a masculine bias throughout all models' performances: We observe unfair translation behavior manifesting itself in a consistent misrepresentation of feminine word forms in the models' outputs. This bias detriments the translation performance for sentences mentioning women and benefits the translation of sentences mentioning men; accordingly, we observe unfair treatment of one referent gender in favor of the other. Continuing our evaluation, we want to analyze whether our identified gender-biased translations are, in fact, a result of the underrepresentation of *communicative repertoires of women*; therefore, we break down the results for speakers of both genders and examine them for differences between both speaker groups, men and women. Furthermore, we refine our evaluation by analyzing the effect of the low gender-inflectional English bridge language on gender preservation for utterances with speaker-*related* and speaker-*independent* gender agreement.

Utterance Category 1: Speaker-Related Gender Agreement

First, we look into those utterances with speaker-related gender agreement, where correct gender inflection can be evaluated based on the gender of the speaker (i.e., "Je me sentais *isolée$_F$*" uttered by a woman requires the adjective inflection to be congruent with the gender of the speaker). We hypothesized that when bridging via English, the models lose knowledge about the gender of the referent (= speaker) necessary for the correct gender-inflectional build-up of the translation, resulting in gender ambiguity impairing gender preservation; by examining the results for utterances with speaker-related gender agreement presented in Table 5.6, we want to analyze whether zero-shot and pivot-based translation are indeed affected by this. Given the nature of utterances with speaker-related gender agreement, additionally accounting for the impact of speakers' gender on models' gender biases is straightforward: results for feminine referents/words represent results for female speakers only, i.e., women, and those for masculine referents/words only represent results for male speakers, i.e., men.

Table 5.6: Average accuracy scores for *Correct* ($\gamma_{correct}$, higher ↑ is better) MuST-SHE references with speaker-related gender agreement, broken down by referent gender (agrees with speaker gender), and complemented with the gender gap ($\delta_{referent}$, lower ↓ is better). Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 1: Speaker-*Related* Gender Agreement —						
Model ($\widehat{en}$)	**Feminine** Referent ↑ (*Female* Speaker)		**Masculine** Referent ↑ (*Male* Speaker)		Gender Gap ↓	
	ZS	PV	ZS	PV	ZS	PV
Baseline (B)	4.0	15.3	8.2	53.2	**0.51**	0.71
B + AUX$_{SIM}$	16.9	16.3	36.6	49.9	0.54	0.67
B + ADV$_{LAN}$	17.3	16.5	**56.6**	**55.5**	0.69	0.70
Residual (R)	17.6	18.3	48.8	48.2	0.64	0.62
R + AUX$_{SIM}$	17.8	**19.2**	47.7	45.0	0.63	**0.57**
R + ADV$_{LAN}$	**20.2**	**19.2**	48.2	48.7	0.58	0.61
AVG excl. B	18.0	17.9	47.6	49.5	0.62	0.64

First of all, it shows that the masculine bias is even more pronounced in the translation results for utterances of this category than for both utterance categories combined (cf. Table 5.4): Comparing the results for both speaker groups shows more than twice as high scores for male speakers than for female speakers; because of this, the gender gaps are generally larger than in Table 5.4. This observation confirms the notion that models struggle to preserve gender for utterances with speaker-related gender agreement *when bridging via English*, more so than for utterances with speaker-independent gender agreement (cf. Table 5.6). Furthermore, it shows that the poorer gender preservation affects the feminine gender more than the masculine gender when considering the only slightly worse performance for masculine words (AVG excl. B lower by approx. 3–4 scores) while observing a significant drop in scores for feminine words (AVG excl. B lower by approx. 10–11 scores) compared to the results for all utterances (cf. Table 5.4), making the previously found discrimination against women more prominent.

Both zero-shot and pivot-based translation are affected by the low gender-inflectional system of the English bridge language and the associated limitations to preserving and recovering gender information. On average (AVG excl. B), zero-shot models achieve equal to better feminine gender preservation performance compared to pivoting ($18.0 > 17.9$) in contrast to slightly worse feminine performance for all utterances combined ($27.8 < 28.3$, cf. Table 5.4); moreover, the worse masculine performance of zero-shot translation compared to pivoting is not as pronounced ($|47.6 - 49.5 = 1.9|$) as before, for all utterances combined ($|50.4 - 53.6| = 3.2$). While the differences are small, this indicates that zero-shot translation suffers slightly less from the limitations of the English bridge language than pivoting in the case of speaker-related gender agreement, as we hypothesized; however, the effect of the limitations of the English bridge language on both approaches' performances is much more similar than expected.

Regarding the model modifications, we generally observe similar outcomes as before: the best scores by a large margin for female and male speakers (i.e., for feminine and masculine referents) yield $R + ADV_{LAN}$ and $B + ADV_{LAN}$ respectively; the lowest gender gap of the two again achieves $R + ADV_{LAN}$ (0.58 for ZS and 0.61 for PV). Using $R + ADV_{LAN}$, zero-shot translation outperforms pivoting for feminine words (20.2 > 19.2), whereas it is the opposite for masculine words (48.2 < 48.7); the larger difference in feminine (±1.0) than in masculine results (±0.5) between both approaches is again the reason for zero-shot translation achieving a smaller, better gender gap (0.58.0 < 0.61) and, thus, fewer gender-biased outputs.

After all, it remains an open question whether the observed gender differences are due to models' inability to properly model women's use of language ("women talking") or whether the discrimination trivially narrows down to worse translations of feminine words ("talking about women"), regardless of who is using them. Accounting for speakers' gender for utterances in which feminine and masculine words are used by speakers of both genders (as is possible for utterances with speaker-independent gender agreement) promises to shed more light on the cause of the repeatedly visible masculine bias.

Utterance Category 2: Speaker-Independent Gender Agreement

In utterances with speaker-independent gender agreement, gender information is always present and explicit, even in English (e.g., "[...] *she* was strong [...]", "[...] *elle* était *forte*$_F$ [...]"); hence, using the low gender-inflectional English bridge language does not impair gender preservation for that matter. Due to the independence of gender agreement from the speaker, these utterances allow us to look into the gender preservation in translations of feminine and masculine words used by *both* speaker groups, i.e., both speaker groups "talking about men *and* about women"; the corresponding results are presented in Table 5.7. The trend of better masculine than feminine

Table 5.7: Average accuracy scores for *Correct* ($\gamma_{correct}$, higher ↑ is better) MuST-SHE references with speaker-independent gender agreement. Results are broken down by referent gender (Feminine, Masculine) and speaker gender (Female, Male), and are complemented with the referent and the speaker gender gap ($\delta_{referent}$, $\delta^{F}_{speaker}$ and $\delta^{M}_{speaker}$, lower ↓ is better). Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 2: Speaker-*Independent* Gender Agreement —

		Referent						Speaker Gender Gap ↓			
Model ($\widehat{en}$)	Speaker	**Feminine** ↑		**Masculine** ↑		Gender Gap ↓		**Fem.** Ref.		**Masc.** Ref.	
		ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
Baseline	Female	2.1	**37.8**	6.6	57.4	0.68	**0.34**	**0.00**	0.35	0.47	0.02
	Male	2.1	**58.8**	12.4	56.4	0.83	0.04				
+ AUX$_{SIM}$	Female	26.7	34.3	38.3	55.1	**0.30**	0.38	0.36	0.27	**0.01**	0.02
	Male	41.5	47.0	38.7	53.8	**0.07**	0.13				
+ ADV$_{LAN}$	Female	33.1	32.6	57.8	59.6	0.43	0.45	0.37	0.35	0.02	0.06
	Male	**52.7**	50.2	**58.8**	56.0	0.10	0.10				
Residual	Female	32.6	36.5	53.6	61.7	0.39	0.41	0.13	0.35	0.03	0.07
	Male	37.3	56.1	52.0	**57.2**	0.28	**0.02**				
+ AUX$_{SIM}$	Female	40.0	34.9	**58.9**	53.5	0.32	0.35	0.11	**0.14**	0.12	**0.01**
	Male	44.8	40.6	52.1	52.8	0.14	0.23				
+ ADV$_{LAN}$	Female	**41.0**	37.4	58.2	**61.9**	**0.30**	0.40	0.20	0.26	0.04	0.09
	Male	51.1	50.5	56.1	56.6	0.09	0.11				
AVG excl. B	Female	34.7	35.1	53.4	58.4	0.30	0.40	0.19	0.29	0.11	0.05
	Male	45.5	48.9	51.5	55.3	0.07	0.02				

performance is again reflected, although less prominent than for utterances with speaker-related gender agreement (cf. Table 5.6), which emphasizes the notion of better (less challenging) gender preservation for speaker-independent gender agreement *when bridging via English.*

For feminine words ("talking about women"), each model yields higher scores for male ("men talking") than for female speakers ("women talking"), regardless of the approach[1]; feminine words by female speakers are, on average, lower by more than ten scores than those of male speakers (compare AVG excl. B: 34.7 vs. 45.5 for ZS and 35.1 vs. 48.9 for PV). For masculine words ("talking about men"), we observe a mixed picture: For zero-shot translation, half of the models (B, $B + AUX_{SIM}$, $B + ADV_{LAN}$) achieve lower scores for female speakers than for male speakers, and the others (R, $R+AUX_{SIM}$, $R+ADV_{LAN}$) do vice versa; meanwhile, for pivoting, all models achieve higher scores for female than for male speakers.

Regarding the gender differences, we consider the *referent gender gap* (between results for feminine and masculine referents/words, $\delta_{referent}$) and the *speaker gender gaps* (between results for female and male speakers, $\delta^{F}_{speaker}$ and $\delta^{M}_{speaker}$). Comparing the referent gender gaps for the two speaker groups shows noticeably higher (worse) gaps for female speakers than for male speakers (e.g., AVG excl. B: 0.30 > 0.07 for ZS and 0.40 > 0.02 for PV). Comparing the speaker gender gaps for the two referent genders shows, on average, higher (worse) gaps for feminine words than for masculine words (e.g., AVG excl. B: 0.19 > 0.11 for ZS and 0.29 > 0.05 for PV). Consequently, for female speakers ("women talking"), the preservation results for the two referent genders are further apart than for male speakers, with feminine words ("talking about women") generally being less well preserved than masculine words.

Furthermore, the average results reveal slightly smaller differences between the referent gender gaps ($|0.30-0.07| = 0.23$ for ZS $< |0.40-0.02| = 0.38$ for PV) and the speaker gender gaps ($|0.19-0.11| = 0.08$ for ZS $< |0.29 - 0.05| = 0.24$ for PV) of zero-shot models compared to pivoting; this indicates slightly more balanced performances for the genders considered in each case; for instance, the referent gender gaps for the two speaker genders are of more similar magnitude for zero-shot translation—in other words, the extent to which masculine gender preservation is better than feminine gender preservation is more similar between speaker groups—signaling fairer gender treatment of zero-shot models than is the case for pivoting.

A further comparison between approaches can be made for zero-shot translation using $R + ADV_{LAN}$ (the model that, so far, has arguably achieved the overall best zero-shot results, including consideration of gender equality) and pivoting using B (the model that achieves the best pivoting scores for feminine referents/words for both speaker groups): We observe a smaller referent gender gap for *female* speakers (0.30 < 0.34) and a smaller speaker gender gap for *feminine* words (0.20 < 0.35) of zero-shot translation, whereas pivoting achieves a smaller referent gender gap for *male* speakers (0.09 > 0.04) and a smaller speaker gender gap for *masculine* words (0.04 > 0.02); here, the biggest advantage of one approach over the other has zero-shot translation regarding the difference in gender preservation between female and male speakers for feminine words (smaller by 0.15), which proves to be the biggest source of gender inequality (i.e., the difference between men and women when "talking about women") in our evaluation; the other advantages of one approach over the other are arguably negligible.

Tendentially, pivoting can achieve smaller gender gaps between the two referent genders used by *male* speakers and the two speaker groups using *masculine* words. On the other hand, zero-shot translation achieves smaller gender gaps between the two referent genders used by *female* speakers and a (much) smaller gap between the two speaker groups using *feminine* words. Since referent gender gaps for female speakers and speaker gender gaps for feminine words are almost always significantly higher and, thus, worse than their gender-reversed counterparts, it is reasonable to conclude that reducing the former serves the mitigation of gender bias more

[1] Zero-shot translation with B is an exception that is negligible as B performs poorly.

than reducing the latter; this tendency is demonstrated by zero-shot translation more than by pivoting.

Notably, gender preservation for *feminine* words used by *female* speakers ("women talking about women") achieves by far the lowest result (the highest scores is 41.0) among much higher and otherwise more or less similar best scores for the other referent-speaker-gender combinations (58.8 for "men talking about women", 58.8 for "men talking about men", and 61.9 for "women talking about men"). Accordingly, the considerably higher performance for feminine words used by men compared to women is a primary factor contributing to models' *masculine bias*, because the results suggest that the models are *worse* at modeling the language use of *women* than that of men *when they use feminine words* (i.e., when they talk about women). In contrast, for the use of masculine words, we observe better performance for female than for male speakers, which shows that models are *not* systematically discriminating female speakers. However, these findings, combined with the fact that women use feminine words more often than men (inherently whenever they talk about themselves, cf. speaker-related gender agreement), show that our tested models undeniably put women at a disadvantage for many different application scenarios of translation technology.

Summary of Preliminary Findings

Overall, the results of this analysis add to our previous findings that besides better translation performance for masculine words, the models' masculine bias manifests itself in better performance for male speakers—but only in the case of feminine words. Accordingly, we observe gender-biased behavior of all models, *discriminating feminine words used by female speakers* (i.e., "women talking about women") through worse gender preservation performance; we observe this type of gender bias for both approaches, however, zero-shot translation produces slightly fewer gender-biased results than pivoting. It also shows that models are *not* systematically discriminating female speakers, since gender preservation for *masculine words* used by *female speakers* works tendentially even *better* than for male speakers.

Generally, pivoting achieves smaller gender gaps between the two referent genders used by *male* speakers and the two speaker groups using *masculine* words; zero-shot translation achieves smaller gender gaps between the two referent genders used by *female* speakers and a (much) smaller gap between the two speaker groups using *feminine* words. Since referent gender gaps for female speakers and speaker gender gaps for feminine words are almost always significantly higher and, thus, worse than their gender-reversed counterparts, we conclude that reducing the former serves the mitigation of gender bias more than reducing the latter; this desired tendency is demonstrated by zero-shot translation more than by pivoting.

The results confirm better gender preservation performances for utterances with speaker-related than for those with speaker-independent gender agreement when bridging via English. Zero-shot translation is slightly less affected by the loss of gender information in the English language for utterances with speaker-related gender agreement; however, the effect of the low gender-inflectional bridge language on both explicit and implicitly-learned bridging is much more similar than hypothesized.

5.3 Effect of The Bridge Language

In light of our second research question, we want to investigate how the bridge language influences the masculine bias of zero-shot and pivot-based translation. We hypothesized using a bridge language with a rich gender-inflected system, such as grammatical gender languages, enables better preservation of gender information necessary to disambiguate gender, unlike languages that lack a comparable gender-inflectional system, such as English, which is typically the most common language in training data and therefore the most obvious choice for a bridge language. As we have previously presented the results for models using English as an explicit or implicit bridge, we now compare them to the performances of models trained using the grammatical gender bridge languages German and Spanish.

Considering the performance aspects evaluated so far, we focus on the two different utterance categories with speaker-related and speaker-independent gender agreement, as we expect to see notable differences between the three bridge languages, English, German, and Spanish. To facilitate a better presentation of the results, we use abbreviations such as $\overset{\frown}{en}$ for English-bridging models, $ZS_{\overset{\frown}{es}}$ to indicate zero-shot translation with models using Spanish as the bridge language, and $PV_{\overset{\frown}{de}}$ for pivoting using German as the bridge/pivot language.

Utterance Category 1: Speaker-Related Gender Agreement

The results for $\overset{\frown}{de}$ and $\overset{\frown}{es}$ models for utterances with speaker-related gender agreement are presented in Table 5.8, complemented with the difference in performance compared to $\overset{\frown}{en}$ models in parentheses. First of all, we notice that B achieves higher scores for $ZS_{\overset{\frown}{de}}$ and $ZS_{\overset{\frown}{es}}$ than for

Table 5.8: Average accuracy scores for *Correct* ($\gamma_{correct}$, higher ↑ is better) MuST-SHE references with speaker-related gender agreement (category 1) when bridging via German (de) or Spanish (es); the change in accuracy compared to bridging via English (cf. Table 5.6) is enclosed in parentheses. Results are broken down by referent gender (agrees with speaker gender). Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 1: Speaker-*Related* Gender Agreement —									
Training Data X ↔	Model	**Feminine** Referent ↑ (*Female* Speaker)				**Masculine** Referent ↑ (*Male* Speaker)			
		Zero-Shot		Pivot		Zero-Shot		Pivot	
de	Baseline (B)	8.8	(+4.8)	17.8	(−2.5)	25.9	(+17.7)	**53.2**	(±0)
	B + AUX$_{SIM}$	13.7	(−3.2)	**19.9**	(+3.6)	28.5	(−8.1)	47.0	(−2.9)
	B + ADV$_{LAN}$	14.7	(−2.6)	14.8	(−1.7)	44.1	(−12.5)	44.1	(−11.4)
	Residual (R)	18.7	(+1.1)	15.2	(−3.1)	48.9	(+0.1)	50.1	(+1.9)
	R + AUX$_{SIM}$	**19.5**	(−1.7)	15.3	(−3.9)	44.8	(−2.9)	46.4	(−2.9)
	R + ADV$_{LAN}$	15.1	(−5.1)	18.4	(−0.8)	**51.1**	(+2.9)	49.8	(+1.1)
	AVG excl. B	16.3	(−2.3)	16.7	(−1.2)	43.5	(−4.1)	47.5	(−2.8)
es	Baseline (B)	33.0	(+29.0)	41.5	(+26.2)	41.8	(+33.6)	52.1	(−1.1)
	B + AUX$_{SIM}$	39.2	(+22.3)	41.9	(+25.6)	46.1	(+9.5)	44.2	(−5.7)
	B + ADV$_{LAN}$	**41.0**	(+23.7)	**45.6**	(+29.1)	46.4	(−10.2)	46.8	(−8.7)
	Residual (R)	32.4	(+14.8)	34.3	(+16.0)	**53.1**	(+4.3)	**53.6**	(+5.4)
	R + AUX$_{SIM}$	33.3	(+15.5)	32.4	(+13.2)	51.4	(+3.7)	41.3	(−3.7)
	R + ADV$_{LAN}$	23.8	(+3.6)	29.4	(+10.2)	50.6	(+2.4)	45.7	(+3.0)
	AVG excl. B	33.9	(+16.0)	36.7	(+18.8)	49.5	(+1.9)	46.3	(−1.9)

$ZS_{\overset{\frown}{en}}$, and modifying B can in both cases further improve the results. Regarding the different model modifications, there is no clear pattern visible as to one model performing best for the three bridge languages for the feminine or the masculine gender.

The best zero-shot performances for feminine words using the bridge languages German, Spanish, and English are 19.5 ($R + AUX_{SIM}$), 41.0 ($B + ADV_{LAN}$), and 20.2 ($R + ADV_{LAN}$), respectively. Accordingly, $ZS_{\widehat{es}}$ produces more than twice as many correct feminine translations than $ZS_{\widehat{de}}$ or $ZS_{\widehat{en}}$. The best performances for masculine words for $ZS_{\widehat{de}}$ and $ZS_{\widehat{es}}$ are 51.1 ($R + ADV_{LAN}$) and 53.1 (R), while for $ZS_{\widehat{en}}$ it is 56.6 ($B + ADV_{LAN}$). $ZS_{\widehat{es}}$ is again more accurate than $ZS_{\widehat{de}}$ when it comes to masculine translation, but by a much smaller margin; however, both are outperformed by $ZS_{\widehat{en}}$.

While the masculine results of $ZS_{\widehat{en}}$ defeat those of the other two bridge languages, it is worth noting that, at the same time, $ZS_{\widehat{en}}$ produces larger gender gaps, reflected by negative values in parenthesis in Table 5.9 (especially for $\widehat{es}$ models), due to lower feminine and higher masculine results (cf. Table 5.6). Most likely, the cause is $ZS_{\widehat{en}}$ models not reliably recognizing the correct gender form due to the low gender-inflected bridge language and instead reflecting learned probabilities of gender occurrence whose distribution is skewed toward the masculine gender.[2]

Table 5.9: Referent gender gaps ($\delta_{referent}$, lower ↓) between feminine and masculine accuracies for *Correct* MuST-SHE references (category 1, cf. Section 5.8). Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 1: Speaker-*Related* Gender Agreement —					
Training Data X ↔	Model	Gender Gap ↓			
		Zero-Shot		Pivot	
de	Baseline (B)	0.66	(+0.15)	0.67	(+0.04)
	B + AUX$_{SIM}$	$\underline{\mathbf{0.52}}$	(−0.02)	**0.58**	(−0.09)
	B + ADV$_{LAN}$	0.67	(−0.02)	0.66	(−0.04)
	Residual (R)	0.62	(−0.02)	0.70	(+0.08)
	R + AUX$_{SIM}$	0.56	(−0.07)	0.67	(−0.10)
	R + ADV$_{LAN}$	0.70	(+0.12)	0.63	(−0.02)
	AVG excl. B	0.61	(−0.04)	0.65	(−0.03)
es	Baseline (B)	0.21	(−0.30)	0.20	(−0.51)
	B + AUX$_{SIM}$	0.15	(−0.39)	0.05	(−0.62)
	B + ADV$_{LAN}$	**0.12**	(−0.57)	$\underline{\mathbf{0.03}}$	(−0.67)
	Residual (R)	0.39	(−0.25)	0.36	(−0.26)
	R + AUX$_{SIM}$	0.35	(−0.28)	0.22	(−0.35)
	R + ADV$_{LAN}$	0.53	(−0.05)	0.36	(−0.25)
	AVG excl. B	0.31	(−0.31)	0.20	(−0.43)

In contrast, for the models using the two grammatical gender bridge languages there is less evidence of simply reflecting a learned gender distribution. Turning to the deltas between the accuracies for the *Correct* and the *Wrong* set, presented in Table 5.10, it shows that both $ZS_{\widehat{de}}$ and $ZS_{\widehat{es}}$ models achieved noticeably higher (better) feminine deltas compared to $ZS_{\widehat{en}}$ indicated by positive values in parentheses. This is evidence of the $ZS_{\widehat{en}}$ models much more often choosing the wrong (masculine) over the correct gender (feminine) compared to $ZS_{\widehat{de}}$ and $ZS_{\widehat{es}}$ models for feminine words. Furthermore, almost all $ZS_{\widehat{es}}$ models also produce higher masculine deltas

[2] In MuST-C training corpus approximately 30% female versus 70% male speakers.

Table 5.10: Differences (Delta, higher ↑ is better) between average accuracy scores for *Correct* and *Wrong* MuST-SHE references (category 1, cf. *Correct* scores in Table 5.8) when bridging via German (de) or Spanish (es); the change in difference compared to bridging via English (cf. Table 5.6) is enclosed in parentheses. Results are broken down by referent gender. Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 1: Speaker-Related Gender Agreement —									
Training Data X ↔	Model	Delta ↑							
		Feminine Referent ↑ (*Female* Speaker)				**Masculine** Referent ↑ (*Male* Speaker)			
		Zero-Shot		Pivot		Zero-Shot		Pivot	
de	Baseline (B)	-12.1	(−8.3)	-16.1	(+17.9)	22.4	(+16.5)	**44.2**	(+1.9)
	B + AUX$_{SIM}$	-10.1	(−3.2)	$\underline{\textbf{-9.5}}$	(+18.2)	23.0	(−4.2)	38.4	(−2.8)
	B + ADV$_{LAN}$	-15.3	(+8.0)	-17.7	(+14.7)	38.6	(−9.3)	37.4	(−10.4)
	Residual (R)	**-9.6**	(+9.2)	-20.1	(+12.3)	43.3	(−4.6)	41.8	(−6.0)
	R + AUX$_{SIM}$	-12.0	(+3.5)	-18.8	(+5.0)	36.3	(+1.4)	38.4	(+4.8)
	R + ADV$_{LAN}$	-13.6	(+2.5)	-13.2	(+10.6)	$\underline{\textbf{44.9}}$	(+6.3)	41.7	(+7.1)
	AVG excl. B	-12.1	(+4.0)	-15.9	(+12.2)	37.2	(−2.1)	39.5	(−1.5)
es	Baseline (B)	19.2	(+23.0)	21.0	(+55.0)	35.9	(+30.0)	43.5	(+1.2)
	B + AUX$_{SIM}$	**27.4**	(+34.3)	25.9	(+53.6)	37.2	(+10.0)	34.4	(−6.8)
	B + ADV$_{LAN}$	27.3	(+50.5)	$\underline{\textbf{29.6}}$	(+62.0)	36.6	(−11.3)	37.5	(−10.3)
	Residual (R)	8.9	(+27.7)	4.3	(+31.4)	**44.2**	(+4.0)	$\underline{\textbf{45.3}}$	(+9.1)
	R + AUX$_{SIM}$	11.3	(+26.8)	7.1	(+30.3)	42.6	(+5.0)	32.1	(−1.9)
	R + ADV$_{LAN}$	-5.8	(+10.3)	3.2	(+27.0)	39.3	(+0.7)	35.5	(+0.9)
	AVG excl. B	13.8	(+29.9)	14.2	(+40.9)	40.0	(+1.7)	37.0	(−1.8)

than $ZS_{\overset{\frown}{en}}$ models, which shows that the better gender preservation of Spanish indeed benefits both genders.

This is also reflected in the gender gaps depicted in Table 5.9: Compared to the others, $\overset{\frown}{es}$ models reduce the gender gap, and thus bias, by producing significantly better feminine and (only) slightly better masculine translations, which brings models closer toward closing the gender gap. In contrast, $ZS_{\overset{\frown}{de}}$ models, similar to $ZS_{\overset{\frown}{en}}$ models, rather than relying on the gender clues from the source, seem to frequently produce the wrong gender form according to a bias that detriments feminine and benefits masculine translation, resulting in negative feminine deltas and larger gender gaps. Arguably, the Spanish language has an even richer gender-inflectional system than German (e.g., Spanish adjectives, unlike German adjectives, are inflected for gender), which can likely better preserve gender signals conveyed by the source languages, which, on another note, share the same language family as Spanish; for speaker-related gender agreement in particular, this could explain $\overset{\frown}{es}$ models' better ability to preserve gender, while $\overset{\frown}{de}$ models are more on par with $\overset{\frown}{en}$ models.

For both $\overset{\frown}{de}$ and $\overset{\frown}{es}$ models, we observe higher scores for pivoting than for zero-shot translation for either referent gender in Table 5.8. Comparing the gender gaps of both approaches in Table 5.9, we see that pivoting consistently achieves smaller gender gaps than zero-shot translation for $\overset{\frown}{es}$ models, some of which are very close to zero, indicating equally good preservation results for both genders. On the other hand, zero-shot translation can achieve smaller gaps for

$\widehat{de}$ models, of similar magnitude as those of $\widehat{en}$ models. These outcomes suggest that with an increased level of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairly balanced gender preservation. Furthermore, the similarity of the source and bridge language could also play a decisive role in gender preservation: When gender inflection of languages is similar, gender preservation does not require strong generalizability of gender representation that underlies zero-shot but not pivot-based translation; hence, explicit bridging (pivoting) is even advantageous for gender preservation into languages that have a similar gender-inflectional system to the source language but disadvantageous for dissimilar languages.

Regarding the different model modifications, a notable difference to previous results is that the removed residual connection does not improve the zero-shot performance for feminine words in the case of $\widehat{es}$ models most times; it can even decrease the accuracy (e.g., 32.4 and 23.8 using R and $R + ADV_{LAN}$) compared to the baseline (33.0). For masculine words, this is not the case. Nevertheless, the gender gaps also reflect significantly worse performances for models with a removed residual connection (e.g., 0.39 and 0.53 using R and $R + ADV_{LAN}$) compared to the baseline (0.21). Accordingly, lifting the positional correspondence to input tokens to some extent in $\widehat{es}$ models' language representations does not seem to improve but rather degrade zero-shot (and pivoting) performances for feminine words, whereas for masculine words it has the opposite effect—both attributes to larger gender gaps. This observation strengthens our idea that in the Spanish case, where the bridge language is very similar to the source languages, lifting the positional correspondence of input tokens to hidden representations impairs the model's ability to preserve the correct gender rather than improving it, causing it to rely on biases instead that favor masculine over feminine outputs.

Utterance Category 2: Speaker-Independent Gender Agreement

For sentences with speaker-independent gender agreement, the results for $\widehat{de}$ and $\widehat{es}$ models are presented in Table 5.11, again complemented with the difference in performance compared to $\widehat{en}$ models in parentheses. Similar to $\widehat{en}$ models, we observe better $\widehat{de}$ and $\widehat{es}$ performances for the utterances with speaker-independent gender agreement than for those where gender agreement is speaker-related, most likely due to more explicit and unambiguous gender clues that can be conveyed in a variety of different languages and are thus easier for models to preserve in the process of (explicit or implicit) bridging.

While the masculine scores are slightly better, feminine scores are significantly higher compared to the results for utterances with speaker-related gender agreement (utterance category 1), especially for $\widehat{de}$ models (cf. Table 5.8 vs. Table 5.11); this is also reflected in the gender gaps in Table 5.12. Compared to utterance category 1, the gender gaps for $\widehat{de}$ and $\widehat{es}$ models for utterance category 2 are of much more similar magnitude, even if $\widehat{es}$ models still perform better.

The performance difference to $\widehat{en}$ models indicated in parentheses also highlights the fact that the grammatical gender bridge languages achieve fewer gender-biased results, since the gender gaps are consistently smaller (hence, the negative values in parentheses). In large part, this can again be attributed to higher feminine preservation scores for the two bridge languages compared to English, as shown in Table 5.11. At 43.1, the best feminine result of $\widehat{de}$ models is no longer that far behind the best feminine result of $\widehat{es}$ models with 50.3 and is better than 42.8, the highest feminine result of $\widehat{en}$ models. In particular, $\widehat{de}$ models move the feminine scores much closer to the masculine ones than in the previous utterance scenario, where the gender gaps are considerably larger (e.g., for $R + ADV_{LAN}$, compare 0.19 in Table 5.12 vs. 0.70 in Table 5.9).

In contrast to speaker-related gender agreement, comparing zero-shot to pivot-based translation for speaker-independent gender agreement shows that the former can perform equally well as the latter regarding feminine gender preservation for $\widehat{es}$ models, which is also reflected in the

Table 5.11: Average accuracy scores for *Correct* ($\gamma_{correct}$, higher ↑ is better) MuST-SHE references with speaker-independent gender agreement (category 2) when bridging via German (de) or Spanish (es); the change in accuracy compared to bridging via English is enclosed in parentheses. Results are broken down by referent gender. Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 2: Speaker-*Independent* Gender Agreement —									
Training Data X ↔	Model	**Feminine** Referent ↑ (*Female/Male* Speaker)				**Masculine** Referent ↑ (*Female/Male* Speaker)			
		Zero-Shot		Pivot		Zero-Shot		Pivot	
de	Baseline (B)	26.9	(+24.8)	41.3	(+0.3)	39.5	(+28.9)	52.5	(−4.2)
	B + AUX$_{SIM}$	32.1	(+2.7)	40.3	(+3.7)	39.2	(+28.6)	46.7	(−7.2)
	B + ADV$_{LAN}$	37.2	(+0.5)	**43.1**	(+7.3)	49.9	(−8.6)	54.0	(−3.1)
	Residual (R)	41.0	(+7.5)	40.6	(+0.5)	49.9	(−2.6)	**54.7**	(−3.9)
	R + AUX$_{SIM}$	**42.4**	(+1.5)	40.8	(+4.8)	49.0	(−5.2)	49.7	(−3.3)
	R + ADV$_{LAN}$	40.4	(−2.4)	43.6	(+3.8)	**50.1**	(−6.6)	55.6	(−2.7)
	AVG excl. B	38.6	(+2.0)	41.7	(+4.0)	47.6	(+1.1)	52.1	(−4.0)
es	Baseline (B)	43.1	(+41.0)	**48.8**	(+7.2)	50.4	(+39.8)	57.7	(+1.0)
	B + AUX$_{SIM}$	46.2	(+16.8)	46.5	(+9.9)	52.7	(+14.1)	51.6	(−2.6)
	B + ADV$_{LAN}$	**50.3**	(+13.6)	47.9	(+12.1)	55.1	(−3.4)	54.1	(−3.0)
	Residual (R)	49.4	(+15.9)	44.9	(+4.8)	54.5	(+2.0)	**59.2**	(+0.6)
	R + AUX$_{SIM}$	47.1	(+6.2)	42.4	(+6.4)	53.5	(−0.7)	50.5	(−2.5)
	R + ADV$_{LAN}$	49.6	(+6.8)	45.3	(+5.5)	**57.7**	(+1.0)	55.0	(−3.3)
	AVG excl. B	48.5	(+11.9)	45.4	(+7.7)	54.7	(+2.6)	54.1	(−2.2)

gender gaps: $ZS_{\overset{\frown}{es}}$ with $B + ADV_{LAN}$ or with R achieves a gender gap of 0.09, which is close to zero, and similar to the best gender gap for pivoting (0.10). For $\overset{\frown}{de}$, zero-shot and pivot-based translation also achieve similar gender gaps of 0.13 and 0.14, respectively.

It stands out that the smallest (best) $PV_{\overset{\frown}{es}}$ gender gaps for utterance category 1 achieved by $B + ADV_{LAN}$ and $B + AUX_{SIM}$, which are close to zero (0.03 and 0.05), are noticeably smaller than their equivalents for utterance category 2 (0.11 and 0.10, respectively), despite them conveying fewer gender ambiguity. In the case of $B + ADV_{LAN}$, $ZS_{\overset{\frown}{es}}$ surpasses $PV_{\overset{\frown}{es}}$ regarding fewer gender biases ($0.09 < 0,11$); this is the opposite of what we expected to see. These results show that zero-shot translation is at least on par with pivoting for both grammatical gender bridge languages, if not slightly better, regarding bias in gender preservation for utterances with speaker-independent gender agreement and we conclude that zero-shot models are not necessarily exposed to the risk of losing explicit gender clues through generalized, language-independent representations.

Lastly, comparing the average gender gaps achieved by the modified models (AVG excl. B) for the two utterance categories (utterance category 1: $ZS_{\overset{\frown}{en}}/ZS_{\overset{\frown}{de}}/ZS_{\overset{\frown}{es}} = 0.62/0.61/0.31$ versus $PV_{\overset{\frown}{en}}/PV_{\overset{\frown}{de}}/PV_{\overset{\frown}{es}} = 0.64/0.65/0.20$, utterance category 2: $ZS_{\overset{\frown}{en}}/ZS_{\overset{\frown}{de}}/ZS_{\overset{\frown}{es}} = 0.38/0.19/0.11$ versus $PV_{\overset{\frown}{en}}/PV_{\overset{\frown}{de}}/\ PV_{\overset{\frown}{es}} = 0.32/0.20/0.16$) reveals slightly more robustness of zero-shot results for speaker-related gender agreement and of pivoting results for speaker-independent gender agreement. The variance between zero-shot results ($|0.31 - 0.62| = 0.31$, $|0.11 - 0.38| = 0.27$) is smaller than for pivoting results ($|0.11 - 0.38| = 0.27$, $|0.32 - 0.16| = 0.16$), which suggests

Table 5.12: Referent gender gaps ($\delta_{referent}$, lower ↓) between feminine and masculine accuracies for *Correct* MuST-SHE references (category 2, cf. Section 5.11). Bold scores denote the best results per approach and underlined are the best of both approaches.

— Utterance Category 2: Speaker-*Independent* Gender Agreement —					
Training Data X ↔	Model	Referent Gender Gap ↓			
		Zero-Shot		Pivot	
de	Baseline (B)	0.32	(−0.48)	0.21	(−0.06)
	B + AUX$_{SIM}$	0.18	(−0.06)	**0.14**	(−0.18)
	B + ADV$_{LAN}$	0.25	(−0.12)	0.20	(−0.17)
	Residual (R)	0.18	(−0.18)	0.26	(−0.06)
	R + AUX$_{SIM}$	**0.13**	(−0.12)	0.18	(−0.14)
	R + ADV$_{LAN}$	0.19	(−0.06)	0.22	(−0.10)
	AVG excl. B	0.19	(−0.11)	0.20	(−0.13)
es	Baseline (B)	0.14	(−0.66)	0.15	(−0.12)
	B + AUX$_{SIM}$	0.12	(−0.12)	**0.10**	(−0.17)
	B + ADV$_{LAN}$	**0.09**	(−0.28)	0.11	(−0.26)
	Residual (R)	**0.09**	(−0.27)	0.24	(−0.08)
	R + AUX$_{SIM}$	0.12	(−0.13)	0.16	(−0.16)
	R + ADV$_{LAN}$	0.14	(−0.11)	0.18	(−0.14)
	AVG excl. B	0.11	(−0.18)	0.16	(−0.16)

that gender preservation of pivoting is more influenced by the choice of the bridge language than zero-shot translation, which is more robust in this regard.

Summary of Findings

Evaluating the effect of the bridge language on gender translation reveals poorer performance for utterances with speaker-related gender agreement than for speaker-independent gender agreement regardless of the bridge language, which is as expected. Using Spanish as the bridge language yields by far the best gender preservation performances; feminine words, in particular, are modeled better for both zero-shot and pivot-based translation using Spanish. In contrast, bridging via German yields only marginally different outcomes compared to bridging via English for speaker-related gender agreement. For speaker-independent gender agreement, bridging via German outperforms bridging via English by a small margin.

With an increased level of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairly balanced gender preservation for speaker-related gender agreement. For speaker-independent gender agreement, both zero-shot and pivot-based translation are on par regarding balanced gender preservation for feminine and masculine words. Combining the findings for both utterance categories, the zero-shot results are slightly less variable, suggesting that zero-shot translation is less affected and thus less dependent on the choice of bridge language. Nevertheless, the differences in performance between the two approaches are not as big as initially expected.

5.4 Probing Models' Hidden Representations For Gender

So far, in our evaluation, we have explored the effect of different model modifications and bridge languages on the translation of gender-specific expressions. However, it remains an open question whether the observed improvements are, in fact, a result of models' enhanced ability to disambiguate gender, in particular, due to better preservation and rendering of gender signals in language representations. Because of this, we round off our evaluation with a final set of experiments to analyze and potentially explain the previously observed outcomes to some extent. In this final set of experiments, we freeze each model, preventing its weights from being modified, and train a classifier on the output of the model's encoder. We perform a sentence-level and a token-level classification on the source sentence representations produced by the encoder to assess the difficulty of recovering gender-specific information before and after applying the different model modifications. We do so for models using English or Spanish as the bridge language, with the results depicted in Table 5.13. Consistent with the previous experiments, the source languages include French and Italian, and, this time, also Spanish.

Table 5.13: Average accuracy scores for gender classifiers trained to recover gender signals (indicative of referent gender) from encoder outputs, either meanpooled per sentence or processed on the token-level. The results include the accuracy for both genders combined (Fem. & Masc.) and those for each gender (e.g., Fem. or Masc.). For each bridge language, bold scores denote the best results.

Training Data X ↔	Model	**Sentence**			**Token**		
		Fem. & Masc.	Fem.	Masc.	Fem. & Masc.	Fem.	Masc.
en	Baseline (B)	62.3	62.5	62.1	20.9	20.0	21.9
	B + AUX$_{SIM}$	63.1	64.1	62.1	13.6	12.9	14.4
	B + ADV$_{LAN}$	61.5	39.8	**83.9**	19.5	0.05	**36.1**
	Residual (R)	63.9	56.3	71.8	**29.8**	**34.3**	24.8
	R + AUX$_{SIM}$	61.1	54.7	67.7	25.5	27.9	22.8
	R + ADV$_{LAN}$	**67.5**	**74.2**	60.5	11.0	0.04	19.0
	AVG. excl. B	63.4	57.8	69.2	19.9	15.0	23.4
es	Baseline (B)	63.9	74.2	53.2	22.6	19.3	26.2
	B + AUX$_{SIM}$	62.3	75.0	49.2	27.3	24.4	30.5
	B + ADV$_{LAN}$	63.5	**86.7**	39.5	16.4	0.06	28.6
	Residual (R)	**64.3**	66.4	**62.1**	20.4	15.0	26.4
	R + AUX$_{SIM}$	57.9	55.5	60.5	25.4	**30.5**	19.7
	R + ADV$_{LAN}$	63.9	79.7	47.6	**32.7**	24.3	**42.1**
	AVG. excl. B	62.4	72.7	51.8	24.4	18.9	29.5

5.4.1 Sentence-level Gender Classification

For both bridge languages, we notice that the classification accuracies are higher than the translation accuracies observed in our previous experiments. This is somewhat expected since the sentence-level classifier can rely on multiple gender clues in cases where there is more than one gender-marked word in a sentence. Differences between both bridge languages in the magnitude of accuracy are very marginal. Most prominent is the outcome of better feminine accuracy for bridging via Spanish than English, and vice versa for the masculine case. It is noticeable that the highest accuracies for one gender sometimes entail significantly lower accuracies for the other gender (e.g., for English, $B + ADV_{LAN}$ yields feminine accuracy of 39.8 and masculine accuracy of 83.9). This showcases a trade-off between the optimization for either gender. Comparing the results for the baselines ($B/B+$) to those for models with removed residual connections ($R/R+$), there is no significant difference noticeable. This finding confirms that lifting positional correspondence to the input tokens in the latter cases does not necessarily worsen sentence-level classification despite less alignment with word-specific gender information.

Perhaps most interestingly, the results contradict our prior notion of a consistent masculine bias, as we can observe higher feminine than masculine accuracies in multiple cases, especially for Spanish. Despite correct gender prediction on the sentence level, erroneous feminine translations may result from models choosing different (potentially synonymous) word translations than those appearing in the reference or, for sentences with multiple gender-marked words, generating the correct translation for one but not the other word(s); in those cases, the classification accuracy can be higher than the actual gender preservation accuracy discussed in our accuracy-based evaluation. Against this background, sentence-level classification demonstrates its limits as an instrument used to explain why some models perform better than others.

5.4.2 Token-level Gender Classification

Instead of two classes, the token-level classifier distinguishes three (feminine, masculine, neuter), of which we closely examine only the results of the two classes conforming with binary feminine/masculine dichotomy. We observe several differences compared to the results for sentence-level classification. The accuracies are noticeably lower than previously and, considering only the best for both genders, masculine accuracies are higher than feminine accuracies for both bridge languages. Compared to the sentence-level accuracies, the widely poorer token-level classification results indicate the difficulty of recovering (correct) gender information from the token embeddings. Combining the embeddings of tokens that are part of the same word (*word* embeddings) and feeding those to the classifier might lead to different results.

Interestingly, the token-level classifier seems to perform better on representations of models with a removed residual connection ($R/R+$) than on those of the other models ($B/B+$); however, we also notice a large amount of variability in the results, making it difficult to reliably attribute better performance to the removed residual connection. Since removing residual connections lifts positional correspondence of the representations to the input tokens, it is surprising that the classifier predicts the token gender more accurately in some of these cases. It is worth mentioning that the genders of gender-marked words in an utterance are consistent, namely exclusively feminine or masculine. This property of our evaluated utterances might indirectly influence the outcomes due to gender signals of all gender-marked words amplifying each other; performing classification on sentences with mixed gender forms could confirm this or prove otherwise.

Looking at the results of both classifications for both bridge languages, it is striking that the baseline accuracies (B) are not the lowest among all accuracies. This finding contradicts the consistent improvement of the model modifications observed in all our previous experiments (for zero-shot translation). Generally, we find that the outcomes of this final set of experiments are a first step toward analyzing and explaining previous findings. Nevertheless, more fine-grained analyses, including inspection of different sentences and word translations could provide even more insights and explanations, as model transparency and explainability are cornerstones in the efforts toward gender bias mitigation in translation technology.

6

Conclusion

This chapter concludes this thesis, in which we explored gender bias in MNMT in the context of gender preservation for zero-shot translation directions. We compared the gender preservation performances of pivot-based and zero-shot translation while studying the role of the bridge language and the effect of encouraging language-agnostic hidden representations on models' ability to preserve the feminine and masculine gender equally in their outputs. In Section 6.1, we address our research questions related to these studied aspects and, in Section 6.2, we discuss limitations and directions to extend this work in the future.

6.1 Answers to Research Questions

Based on the results obtained from our experiments, we answer the research questions stated in Section 1.3 as follows.

Research Question 1: How do zero-shot and pivot-based translation compare regarding gender-biased outputs in a zero-shot translation setting?

Regarding this question, we tested and confirmed the hypothesis that, *on average, zero-shot translation generates fewer gender-biased outputs than pivot-based translation, where gender bias is conceived as the systematic and unfair discrimination against a group of individuals of the same sex, here either women or men, in favor of the other gender group, while maintaining comparable translation quality*, when the bridge language is English.

Similar to existing works, our evaluation has revealed a masculine bias throughout all evaluated models' performances, meaning models frequently produced the wrong gender form according to a bias that detriments feminine and benefits masculine gender preservation. Using our baseline model, pivoting outperformed zero-shot translation by a large margin, producing fewer gender-biased outputs. When modifying the baseline, namely encouraging language-agnostic representations, zero-shot translation improved noticeably, performing slightly better than pivoting in terms of more fairly balanced preservation of both genders. In particular, pivoting struggling with the feminine gender preservation, more so than zero-shot translation, contributed to more biased outputs since, for feminine-intended words, pivoting produced the (wrong) masculine gender more often than the (correct) feminine gender.

Considering the gender of the speaker using the gender-inflected words, we found gender preservation of feminine words used by female speakers to be significantly worse than for male speakers. In contrast, for the use of masculine words, we observed slightly better performances for female than for male speakers, which shows that models did not systematically discriminate female speakers; however, their gender preservation behavior did, in fact, *discriminate female speaker using feminine words* ("women talking about women", i.e., about themselves or about other women).

Regarding this revealing performance difference, our evaluation showed that zero-shot translation generally achieves a better balance between gender preservation results for female and male speakers using feminine words than pivoting. While the differences were smaller than we had hypothesized, this outcome confirms that zero-shot translation generates fewer gender-biased outputs than pivot-based translation (when bridging via English), also because the unfavorable "translation treatment" of women talking about women, albeit still existent, is not as pronounced and, thus, not as damaging as for pivoting. According to our analyses, the difference in treating this specific case scenario contributed most notably to better zero-shot than pivot-based translation regarding gender bias.

Research Question 2: Does the bridge language affect the gender biases perpetuated by zero-shot and pivot-based translations?

Our experiments, in which we evaluated models' performances using three different bridge languages—English, German, and Spanish—each complying with a distinctive gender-inflectional system, revealed their impact on models' gender-biased translation behavior. The differences became apparent when comparing the translation for sentences with speaker-related and speaker-independent gender agreement: In English, knowledge about the gender of the referent necessary for the correct gender-inflectional build-up of the translation of utterances with speaker-related gender agreement was non-existent in our utterances; hence, there was no certainty of telling the correct gender, whereas German or Spanish have gender-inflectional systems that support better gender disambiguation required for speaker-related gender agreement—the latter even more so than the former.

These gender-related language differences were reflected in the gender preservation performances of models using the three different bridge languages. Translations for utterances with speaker-related gender agreement were generally less accurate regarding the correct gender preservation than those for utterances with speaker-independent gender agreement; when gender ambiguity was higher (more often for speaker-related gender agreement), feminine gender preservation suffered due to more frequent masculine outputs, most likely as a result of models simply reflecting learned probabilities of gender occurrence in the training data. The models bridging via Spanish achieved the most accurate and least biased gender translations for both types of utterances. Bridging via German performed better than bridging via English for the translation of utterances with speaker-independent gender agreement, while performing equally well for speaker-related gender agreement.

Most notably, we find that with an increased level of gender inflection in the bridge language, pivoting surpasses zero-shot translation regarding fairly balanced gender preservation for speaker-related gender agreement. For speaker-independent gender agreement, we find that both zero-shot and pivot-based translation are at least on par regarding balanced gender preservation for feminine and masculine words. Altogether, zero-shot translation was less dependent on the bridge language, making it more robust to changing training data or live data.

Research Question 3: Do translation quality improvements of zero-shot models reduce their gender biases?

All three evaluated modifications to the baseline encouraging language-agnostic hidden representations, namely

1. removing a residual connection in the Transformer encoder to lessen positional correspondence to the input tokens,
2. promoting similar source and target language representations through an auxiliary loss, and
3. joint adversarial training penalizing successful recovery of source language signals in the representations,

improved the gender preservation performances and reduced gender biases of zero-shot models noticeably. Despite the persisting masculine bias, modifying the baseline contributed toward closing the gender gap regarding more similar gender preservation results for feminine and masculine words. Hence, the modifications reduced the detrimental treatment of feminine words, especially when used by female speakers—which, inherently, is more common because it applies whenever women speak about themselves (where men would use masculine words instead) or about other women—since correct translation accuracies were more balanced between both speaker groups. Consequently, modifications improving zero-shot translation by generating more language-independent hidden representations, making them more generalizable, also mitigated zero-shot models' gender biases.

None of the different modifications consistently outperformed the others and the level of improvement for each modification was different depending on the considered translation scenario. However, the adversarial language classifier most notably improved the translation accuracy of feminine and masculine words, whereas the auxiliary similarity loss most often had the most positive effect on balanced gender preservation results.

6.2 Limitations & Future Work

Besides our findings, this work also features some limitations that can be explored in future work. First, the data used in our experimental evaluation limited the scenarios to be examined. Future work can examine translation of sentences with mixed gender (i.e., sentences include both feminine and masculine words) and directions in- cluding languages from different language families and with different gender systems to further study language differences. Second, developing a large gender-annotated corpus suitable for MT training could most likely be used to improve models' gender preservation performance. A well-performing gender classifier could be used to annotate the MuST-C dataset with token- or word-level gender labels. Third, we believe that the metrics currently used to evaluate models' gender biases are not ideal. about the phenomenon of gender bias in translation requires appropriate and established metrics; the lack thereof currently leaves room for improvement in evaluative procedures.

While there is a lot of potential for further research on this topic, it is crucial to acknowledge that, ultimately, translation technology is bound by the principles of language, which subtly reproduces societal asymmetries and embeds signs of sexism. Examples of evident linguistic gender inequality include the consensual norm according to which the prototypical human is male and the more subtle convention by which expressions referring to females are grammatically more complex in many languages. Consequently, combating gender biases in translation technology requires aware- ness of language use, as it is one of the most powerful means through which sexism and gender discrimination is perpetrated and reproduced.

Bibliography

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *5th International Conference on Learning Representations*.

Alhafni, B., Habash, N., & Bouamor, H. (2020). Gender-aware reinflection using linguistically enhanced neural models. *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing*, 139–150.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., & Macherey, W. (2019). The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Basta, C., Ruiz Costa-jussà, M. R., & Rodríguez Fonollosa, J. A. (2020). Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. *Proceedings of the 4th Widening Natural Language Processing Workshop*, 99–102.

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. *arXiv preprint arXiv:2006.05754*.

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

Brownlow, S., Rosamond, J. A., & Parker, J. A. (2003). Gender-linked linguistic behavior in television interviews. *Sex Roles*, *49*(3), 121–132.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014b). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1724–1734.

Cho, W. I., Kim, J. W., Kim, S. M., & Kim, N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, 173–181.

Costa-jussà, M. R., & de Jorge, A. (2020). Fine-tuning neural machine translation on gender-balanced datasets. *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing*, 26–34.

Costa-jussà, M. R., Escolano, C., Basta, C., Ferrando, J., Batlle, R., & Kharitonova, K. (2020). Gender bias in multilingual neural machine translation: The architecture matters. *arXiv preprint arXiv:2012.13176*.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., & Turchi, M. (2019). MuST-C: A multilingual speech translation corpus. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers)*, *1*, 2012–2017.

Escudé Font, J., & Costa-Jussa, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Friedman, B., & Nissenbaum, H. (2017). Bias in computer systems. In *Computer Ethics* (pp. 215–232). Routledge.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, *17*(1), 2096–2030.

Hanulíková, A., & Carreiras, M. (2015). Electrophysiology of subject-verb agreement mediated by speakers' gender. *Frontiers in Psychology*, *6*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709.

Kiefer, J., & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 462–466.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133.

Liu, D., & Niehues, J. (2021). Maastricht University's large-scale multilingual machine translation system for WMT 2021. *Proceedings of the 6th Conference on Machine Translation*, 425–430.

Liu, D., Niehues, J., Cross, J., Guzmán, F., & Li, X. (2021). Improving zero-shot translation by disentangling positional information. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)*, *1*, 1259–1273.

Moryossef, A., Aharoni, R., & Goldberg, Y. (2019). Filling gender & number gaps in neural machine translation with black-box context injection. *arXiv preprint arXiv:1903.03467*.

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.

Pham, N.-Q., Niehues, J., Ha, T.-L., & Waibel, A. (2019–August 2). Improving zero-shot translation with language-independent constraints. *Proceedings of the 4th Conference on Machine Translation*, *1*, 13–23.

Post, M. (2018). A call for clarity in reporting BLEU scores. *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, 186–191.

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, *32*(10), 6363–6381.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Saunders, D., & Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

Saunders, D., Sallis, R., & Byrne, B. (2020). Neural machine translation doesn't translate gender coreference right unless you make it. *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing*, 35–43.

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.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, *1*, 1715–1725.

Shatz, I. (2017). Native language influence during second language acquisition: A large-scale learner corpus analysis. *Pacific Second Language Research Forum 2016*, 175–188.

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, *27*.

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3003–3008.

Vargha, D. (2021). *Hungarian is a gender neutral language, it has no gendered pronouns, so google translate automatically chooses the gender for you. here is how everyday sexism is consistently encoded in 2021. fuck you, google.* [Tweet by @DoraVargha, 2021, March 20]. Twitter. Retrieved 01/10/2023, from https://twitter.com/doravargha/status/1373211762108076034

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in machine translation. *Engineering*, *18*, 143–153.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zoph, B., & Knight, K. (2016). Multi-source neural translation. *arXiv preprint arXiv:1601.00710*.

Appendix A

Background

A.1 Basic Principles of Deep Learning

Deep learning refers to the broader family of ML algorithms based on ANNs leveraging representation learning. ANNs are a biologically-inspired programming paradigm. They emerged as an attempt to exploit the architecture of the human brain to enable computers to learn from observational data to perform tasks that conventional algorithms had little success with. Specifically, they are designed based on a core concept of natural intelligence, namely learning. In principle, learning involves the ability to generalize from past experience to deal with new situations. As such, learning can be narrowed down to the task of approximating a function that defines the relationship between a set of inputs and an output. In view of this, an ANN is essentially a function approximator, comprising a set of parameters which transform an input to an output.

Given large numbers of input and output pairs, the network can gradually adjust its parameters based on measured deviations of its generated outputs from desired targets. This process of updating the parameters is referred to as learning or training. During training, an ANN processes an input into an output by passing it through subsequent layers, transforming it along the way. Typically, an ANN consists of an input layer and an output layer, and one or multiple so-called hidden layers in between. A layer consists of small individual units called neurons and different layers may perform different transformations on their inputs. Stacking multiple layers enables an ANN to perform efficient hierarchical feature learning, during which the hidden layers learn useful representations of the data.

A key characteristic of ANNs is that they perform “end-to-end” learning: all parameters are trained jointly and thus optimized simultaneously instead of sequentially. As a result, a single algorithm learns to model the entire pipeline from initial inputs to desired outputs at once. After a model is trained, it is tested on new instances to validate its ability to generalize, namely being able to digest new data instances and make accurate predictions. The process of applying a trained ANN to infer outputs for unseen instances is referred to as inference. While deep architectures are a decade-old concept, it was not until recently that they delivered unprecedented levels of performance. Reasons for the recent success of deep learning are increased accessibility of training data, advancements in dedicated hardware customized for training, as well as open-source toolkits enabling joint efforts for progress from the entire community. Nowadays, NMT is widely adopted in the industry and is deployed in production systems by Google, Facebook, Microsoft, Amazon, and many more.

Appendix B

Experiments & Results

B.1 The "Blind Spot" in BLEU-based Gender Bias Evaluation Causing Misleading Results

BLEU accounts differently for the neighboring (often gender-neutral) words surrounding the gender-marked words in a translation, depending on whether the latter are translated correctly. As a result, our BLEU-based assessment using gender-swapping is flawed in some specific scenarios. The central idea behind BLEU is to measure "translation closeness" by counting matches of n-grams, here, in a candidate and two reference translations, the *Correct* and the gender-swapped *Wrong* reference. The number of gender-marked words in relation to gender-neutral words in a sentence is small. Consequently, the majority of words participating in n-grams are gender-neutral words, which are identical in *Correct* and *Wrong* references, as they only differ in gender-marked words. If we imagine a candidate translation that matches one gender-marked word (w_C) contained in the *Correct* reference and another one (w_W) contained in the *Wrong* reference, BLEU-based evaluation can fail to reflect the equal number of word matches for both references.

The problem is that a single matched word not only contributes to a higher BLEU score as a unigram but potentially also through broader n-gram matches in which its neighboring words (if translated correctly) participate. Because of this, different word orders in otherwise identical sentences can produce different BLEU scores, which is counterintuitive and misleading. A simple scenario in which the unwanted effect emerges provides our example: If longer sequences of matching gender-neutral words – which are the same in both references – adjoin w_C than w_W, then the BLEU score for the *Correct* reference is higher than the score for the *Wrong* reference, despite having the same number of matching gender-marked words (one for each reference) and otherwise identical gender-neutral words. Consequently, BLEU can signal different tendencies than, for example, the accuracy, and is thereby misleading.

B.2 Experiment Overview

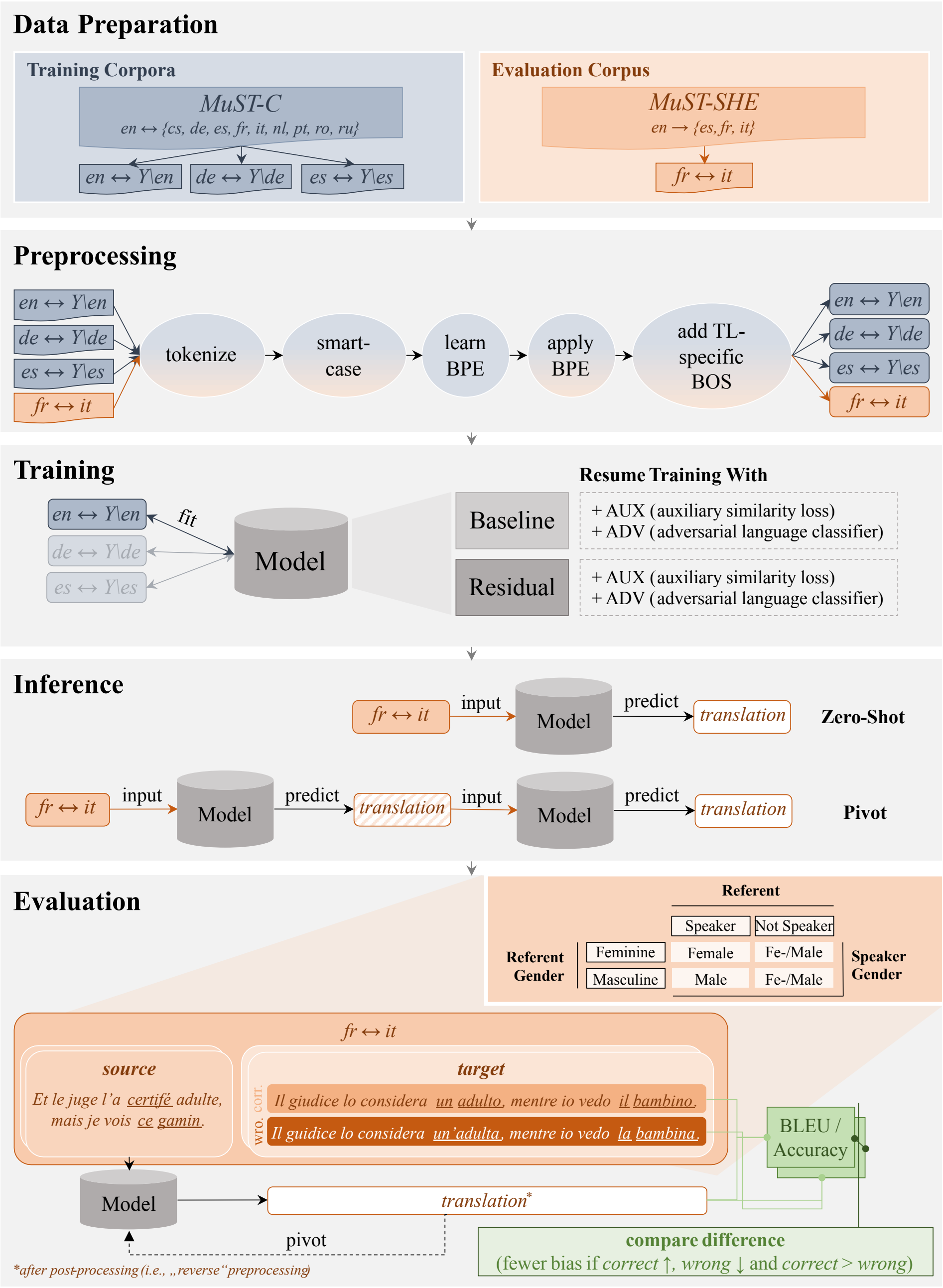


Figure B.1: Overview of the experimental setup and the procedures followed to investigate gender bias in multilingual MT, specifically zero-shot and pivot-based translation, with and without the proposed model modifications and using different bridge languages.

B.3 Results

Table B.1: Average BLEU scores for zero-shot directions on the MuST-C test set and on MuST-SHE. Zero-shot directions are {nl, pt, ro, ru} ↔ {nl, pt, ro, ru} for the MuST-C test set and fr ↔ it for MuST-SHE.

Training Data X ↔	Model	**MuST-C**		**MuST-SHE**	
		Zero-Shot	Pivot	Zero-Shot	Pivot
en	Baseline (B)	3.7	17.4	3.8	25.6
	B + AUX$_{SIM}$	11.1	15.7	15.8	24.0
	B + ADV$_{LAN}$	15.9	17.0	25.3	25.0
	Residual (R)	14.3	17.2	22.9	24.9
	R + AUX$_{SIM}$	15.1	15.6	23.4	23.4
	R + ADV$_{LAN}$	16.0	17.0	25.6	25.4
de	Baseline (B)	6.8	14.6	13.7	21.6
	B + AUX$_{SIM}$	8.6	13.5	15.8	20.4
	B + ADV$_{LAN}$	10.7	14.0	18.6	20.2
	Residual (R)	10.9	14.3	19.8	20.8
	R + AUX$_{SIM}$	12.4	13.3	20.0	19.8
	R + ADV$_{LAN}$	11.1	14.3	19.9	21.3
es	Baseline (B)	13.2	16.5	20.0	26.1
	B + AUX$_{SIM}$	14.1	14.6	23.1	24.4
	B + ADV$_{LAN}$	15.5	16.1	25.2	25.6
	Residual (R)	15.1	16.3	24.5	25.9
	R + AUX$_{SIM}$	14.7	14.4	24.3	23.8
	R + ADV$_{LAN}$	15.3	15.8	24.7	24.4

Table B.2: Utterance Category 1 & 2: Average BLEU scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 1 & 2 (BLEU) –																			
Training Data X ↔	Model	**All**						**Feminine**						**Masculine**					
		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
		ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline (B)	3.8	25.6	3.8	23.6	0	2.0	3.2	23.8	3.3	23.4	-0.1	0.4	4.3	27.1	4.1	23.8	0.2	3.3
	B + AUX$_{SIM}$	15.7	24.0	14.3	22.3	1.4	1.8	15.6	21.8	14.7	21.4	0.9	0.4	15.7	25.8	13.8	22.8	1.9	3.0
	B + ADV$_{LAN}$	25.3	25.0	23.2	23.2	2.1	1.8	23.5	22.8	23.4	22.8	0.1	0.0	26.6	26.7	23.0	23.3	3.6	3.4
	Residual (R)	22.9	24.9	21.2	23.0	1.7	1.9	21.7	23.0	21.6	22.9	0.1	0.1	23.9	26.4	20.7	23.1	3.2	3.3
	R + AUX$_{SIM}$	23.4	23.4	21.4	21.6	2.0	1.9	22.3	21.4	21.5	21.2	0.8	0.2	24.2	24.8	21.1	22.0	3.0	2.8
	R + ADV$_{LAN}$	25.6	25.4	23.7	23.5	1.9	1.9	24.5	23.5	24.2	23.2	0.3	0.3	26.2	26.8	23.2	23.6	3.0	3.2
de	Baseline (B)	13.7	21.6	12.5	19.6	1.2	2.0	13.0	20.2	12.3	19.4	0.7	0.8	14.3	22.8	12.6	19.8	1.7	3.0
	B + AUX$_{SIM}$	15.8	20.4	14.6	18.4	1.2	2.0	15.0	18.4	14.4	17.6	0.6	0.8	16.4	21.8	14.7	19.0	1.7	2.8
	B + ADV$_{LAN}$	18.6	20.2	17.0	18.2	1.6	2.0	18.0	19.1	17.4	18.2	0.6	0.9	19.1	20.4	16.7	17.6	2.4	2.8
	Residual (R)	19.8	20.8	17.8	18.8	2.0	2.0	18.5	19.8	17.4	19.2	1.1	0.6	20.9	21.7	18.0	18.6	2.9	3.1
	R + AUX$_{SIM}$	20.0	19.8	18.0	18.0	2.0	1.8	19.4	18.1	18.3	17.6	1.1	0.5	20.3	21.2	17.6	18.4	2.7	2.8
	R + ADV$_{LAN}$	19.9	21.3	17.8	19.0	2.1	2.3	19.2	20.6	18.2	19.4	1.0	1.2	20.5	21.8	17.5	18.6	3.0	3.2
es	Baseline (B)	20.0	26.1	17.6	23.0	2.4	3.1	19.5	25.2	17.2	22.5	2.3	2.7	20.4	26.8	17.8	23.3	2.6	3.5
	B + AUX$_{SIM}$	23.1	24.4	20.4	21.6	2.7	2.8	22.8	23.8	20.2	21.0	2.6	2.8	23.3	25.0	20.4	22.2	2.9	2.8
	B + ADV$_{LAN}$	25.2	25.6	22.2	22.6	3.0	3.0	24.5	25.4	21.5	22.8	3.0	2.6	25.8	25.6	22.7	22.4	3.1	3.2
	Residual (R)	24.5	25.9	21.8	23.2	2.7	2.7	23.8	25.0	21.4	23.6	2.4	1.4	25.0	26.6	21.8	23.0	3.2	3.6
	R + AUX$_{SIM}$	24.3	23.8	21.6	21.6	2.7	2.2	23.3	22.8	21.1	21.5	2.2	1.3	25.2	24.4	22.1	21.6	3.1	2.8
	R + ADV$_{LAN}$	24.7	24.4	22.2	22.2	2.5	2.2	23.8	24.0	22.0	22.6	1.8	1.4	25.4	24.8	22.4	21.8	3.0	3.0

Table B.3: Utterance Category 1 & 2: Average accuracy scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 1 & 2 (Accuracy) –																			
Training Data X ↔	Model	**All**						**Feminine**						**Masculine**					
		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
		ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline (B)	6.4	42.5	3.8	22.7	2.6	19.8	3.0	29.1	5.5	36.9	-2.5	-7.8	9.7	55.4	2.2	9.1	7.5	46.3
	B + AUX$_{SIM}$	30.8	40.0	13.5	21.0	17.3	19.0	23.4	27.0	20.2	34.2	3.2	-7.2	37.8	52.5	7.1	8.3	30.7	44.2
	B + ADV$_{LAN}$	43.0	41.9	21.2	23.0	21.8	18.9	27.5	26.6	34.3	39.2	-6.8	-12.6	57.8	56.5	8.6	7.5	49.2	49.0
	Residual (R)	38.8	42.4	19.7	22.7	19.1	19.7	25.9	29.8	30.8	35.9	-4.9	-6.1	51.1	54.5	9.1	10.0	42.0	44.5
	R + AUX$_{SIM}$	41.1	39.2	18.0	22.0	23.1	17.2	30.0	28.0	27.6	34.6	2.4	-6.6	51.7	49.9	8.9	9.9	42.8	40.0
	R + ADV$_{LAN}$	43.0	42.5	20.8	22.5	22.0	20.0	32.1	30.0	31.6	34.9	0.5	-4.9	53.4	54.5	10.4	10.6	43.0	43.9
de	Baseline (B)	26.4	41.7	11.1	16.5	15.3	25.2	18.3	30.2	16.6	26.0	1.7	4.2	34.2	52.8	5.9	7.4	28.3	45.4
	B + AUX$_{SIM}$	29.3	38.9	14.0	15.9	15.3	23.0	23.3	30.6	20.4	24.5	2.9	6.1	35.1	46.8	7.8	7.6	27.3	39.2
	B + ADV$_{LAN}$	37.3	40.1	15.7	16.3	21.6	23.8	26.5	29.7	24.4	26.4	2.1	3.3	47.7	50.1	7.5	6.6	40.2	43.5
	Residual (R)	40.2	41.0	15.5	17.1	24.7	23.9	30.4	28.6	23.2	26.7	7.2	1.9	49.5	52.9	8.2	7.9	41.3	45.0
	R + AUX$_{SIM}$	39.6	38.8	17.4	17.2	22.2	21.6	31.6	28.7	25.6	27.2	6.0	1.5	47.4	48.4	9.5	7.7	37.9	40.7
	R + ADV$_{LAN}$	39.7	42.7	15.6	15.4	24.1	27.3	28.4	31.6	23.7	24.0	4.7	7.6	50.5	53.3	7.8	7.1	42.7	46.2
es	Baseline (B)	42.8	50.5	10.4	13.9	32.4	36.6	38.3	45.3	13.2	18.8	25.1	26.5	47.1	55.5	7.6	9.2	39.5	46.3
	B + AUX$_{SIM}$	46.6	46.6	10.9	12.3	35.7	34.3	42.8	44.3	14.0	15.9	28.8	28.4	50.2	48.7	7.9	8.8	42.3	39.9
	B + ADV$_{LAN}$	48.8	49.1	11.0	13.0	37.8	36.1	45.9	46.8	13.0	17.0	32.9	29.8	51.7	51.2	9.0	9.1	42.7	42.1
	Residual (R)	47.8	48.6	14.0	16.9	33.8	31.7	41.3	39.9	19.7	25.3	21.6	14.6	54.0	57.0	8.5	8.9	45.5	48.1
	R + AUX$_{SIM}$	46.7	42.4	13.4	15.6	33.3	26.8	40.5	37.6	18.8	22.3	21.7	15.3	52.7	46.9	8.2	9.3	44.5	37.6
	R + ADV$_{LAN}$	46.3	44.7	15.4	16.0	30.9	28.7	37.4	37.8	21.8	22.3	15.6	15.5	54.9	51.4	9.3	9.9	45.6	41.5

Table B.4: Utterance Category 1: Average accuracy scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references with speaker-related gender agreement, complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 1 (Accuracy) –																			
Training Data X ↔	Model	**All**						**Feminine**						**Masculine**					
		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
		ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline (B)	6.0	32.8	5.3	31.5	0.7	1.3	4.0	15.3	7.8	49.3	-3.8	-34.0	8.2	53.2	2.3	10.9	5.9	42.3
	B + AUX$_{SIM}$	26.0	31.8	17.2	27.7	8.8	4.1	16.9	16.3	23.8	44.0	-6.9	-27.7	36.6	49.9	9.4	8.7	27.2	41.2
	B + ADV$_{LAN}$	35.5	34.5	25.8	29.9	9.7	4.6	17.3	16.5	40.5	48.9	-23.2	-32.4	56.6	55.5	8.7	7.7	47.9	47.8
	Residual (R)	32.0	32.1	23.6	30.0	8.4	2.1	17.6	18.3	36.4	45.4	-18.8	-27.1	48.8	48.2	8.6	12.0	40.2	36.2
	R + AUX$_{SIM}$	31.6	31.1	22.5	27.9	9.1	3.2	17.8	19.2	33.3	42.4	-15.5	-23.2	47.7	45.0	10.1	11.0	37.6	34.0
	R + ADV$_{LAN}$	33.1	32.8	24.0	29.7	9.1	3.1	20.2	19.2	36.3	43.0	-16.1	-23.8	48.2	48.7	9.6	14.1	38.6	34.6
de	Baseline (B)	16.7	34.2	12.8	22.4	3.9	11.8	8.8	17.8	20.9	33.9	-12.1	-16.1	25.9	53.2	3.5	9.0	22.4	44.2
	B + AUX$_{SIM}$	20.5	32.4	15.3	19.8	5.2	12.6	13.7	19.9	23.8	29.4	-10.1	-9.5	28.5	47.0	5.5	8.6	23.0	38.4
	B + ADV$_{LAN}$	28.3	28.3	18.7	20.6	9.6	7.7	14.7	14.8	30.0	32.5	-15.3	-17.7	44.1	44.1	5.5	6.7	38.6	37.4
	Residual (R)	32.7	31.3	17.8	22.8	14.9	8.5	18.7	15.2	28.3	35.3	-9.6	-20.1	48.9	50.1	5.6	8.3	43.3	41.8
	R + AUX$_{SIM}$	31.2	29.6	20.8	22.1	10.4	7.5	19.5	15.3	31.5	34.1	-12.0	-18.8	44.8	46.4	8.5	8.0	36.3	38.4
	R + ADV$_{LAN}$	31.7	32.9	18.3	20.8	13.4	12.1	15.1	18.4	28.7	31.6	-13.6	-13.2	51.1	49.8	6.2	8.1	44.9	41.7
es	Baseline (B)	37.1	46.4	10.1	15.0	27.0	31.4	33.0	41.5	13.8	20.5	19.2	21.0	41.8	52.1	5.9	8.6	35.9	43.5
	B + AUX$_{SIM}$	42.4	43.0	10.5	13.1	31.9	29.9	39.2	41.9	11.8	16.0	27.4	25.9	46.1	44.2	8.9	9.8	37.2	34.4
	B + ADV$_{LAN}$	43.5	46.2	11.9	12.9	31.6	33.3	41.0	45.6	13.7	16.0	27.3	29.6	46.4	46.8	9.8	9.3	36.6	37.5
	Residual (R)	42.0	43.2	16.8	20.0	25.2	23.2	32.4	34.3	23.5	30.0	8.9	4.3	53.1	53.6	8.9	8.3	44.2	45.3
	R + AUX$_{SIM}$	41.7	36.5	15.9	17.9	25.8	18.6	33.3	32.4	22.0	25.3	11.3	7.1	51.4	41.3	8.8	9.2	42.6	32.1
	R + ADV$_{LAN}$	36.2	37.0	21.1	18.8	15.1	18.2	23.8	29.4	29.6	26.2	-5.8	3.2	50.6	45.7	11.3	10.2	39.3	35.5

Table B.5: Utterance Category 2: Average accuracy scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references with speaker-independent gender agreement, complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 2 (Accuracy) –																			
Training Data X ↔	Model	**All**						**Feminine**						**Masculine**					
		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
		ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline (B)	6.8	49.9	2.7	16.0	4.1	33.9	2.1	41.6	3.5	25.8	-1.4	15.8	10.6	56.7	2.0	7.9	8.6	48.8
	B + AUX$_{SIM}$	34.4	46.3	10.7	15.9	23.7	30.4	29.4	36.6	17.0	25.4	12.4	11.2	38.6	54.2	5.6	8.0	33.0	46.2
	B + ADV$_{LAN}$	48.6	47.5	17.7	17.8	30.9	29.7	36.7	35.8	28.8	30.5	7.9	5.3	58.5	57.1	8.6	7.3	49.9	49.8
	Residual (R)	43.9	50.3	16.7	17.1	27.2	33.2	33.5	40.1	25.7	27.3	7.8	12.8	52.5	58.6	9.3	8.7	43.2	49.9
	R + AUX$_{SIM}$	48.2	45.3	14.6	17.5	33.6	27.8	40.9	36.0	22.4	27.5	18.5	8.5	54.2	53.0	8.2	9.3	46.0	43.7
	R + ADV$_{LAN}$	50.4	49.9	18.3	17.0	32.1	32.9	42.8	39.8	27.4	27.5	15.4	12.3	56.7	58.3	10.8	8.4	45.9	49.9
de	Baseline (B)	33.8	47.4	9.8	12.0	24.0	35.4	26.9	41.3	12.7	18.9	14.2	22.4	39.5	52.5	7.5	6.4	32.0	46.1
	B + AUX$_{SIM}$	36.0	43.8	13.0	12.9	23.0	30.9	32.1	40.3	17.4	20.1	14.7	20.2	39.2	46.7	9.3	7.0	29.9	39.7
	B + ADV$_{LAN}$	44.2	49.1	13.5	13.1	30.7	36.0	37.2	43.1	19.3	21.0	17.9	22.1	49.9	54.0	8.8	6.5	41.1	47.5
	Residual (R)	45.9	48.3	13.8	12.8	32.1	35.5	41.0	40.6	18.7	19.0	22.3	21.6	49.9	54.7	9.8	7.8	40.1	46.9
	R + AUX$_{SIM}$	46.1	45.7	14.8	13.6	31.3	32.1	42.4	40.8	20.3	21.0	22.1	19.8	49.0	49.7	10.2	7.4	38.8	42.3
	R + ADV$_{LAN}$	45.7	50.2	13.5	11.3	32.2	38.9	40.4	43.6	19.2	17.1	21.2	26.5	50.1	55.6	8.8	6.5	41.3	49.1
es	Baseline (B)	47.1	53.6	10.5	13.1	36.6	40.5	43.1	48.8	12.8	17.3	30.3	31.5	50.4	57.7	8.7	9.6	41.7	48.1
	B + AUX$_{SIM}$	49.8	49.3	11.2	11.7	38.6	37.6	46.2	46.5	16.0	15.8	30.2	30.7	52.7	51.6	7.2	8.3	45.5	43.3
	B + ADV$_{LAN}$	52.9	51.3	10.3	13.0	42.6	38.3	50.3	47.9	12.4	17.8	37.9	30.1	55.1	54.1	8.5	9.0	46.6	45.1
	Residual (R)	52.2	52.7	11.8	14.6	40.4	38.1	49.4	44.9	16.2	20.9	33.2	24.0	54.5	59.2	8.2	9.4	46.3	49.8
	R + AUX$_{SIM}$	50.6	46.8	11.5	13.9	39.1	32.9	47.1	42.4	16.0	19.6	31.1	22.8	53.5	50.5	7.7	9.3	45.8	41.2
	R + ADV$_{LAN}$	54.0	50.6	11.1	13.8	42.9	36.8	49.6	45.3	14.9	18.8	34.7	26.5	57.7	55.0	8.0	9.7	49.7	45.3

Table B.6: Utterance Category 1 & 2 Average accuracy scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine) and by speaker gender (Female or Male). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 1 & 2 (Accuracy) –																				
Training Data X ↔	Model	Speaker	All						Feminine						Masculine					
			Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
			ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline	F	3.7	31.7	4.8	33.0	-1.1	-1.3	3.1	26.0	5.7	38.4	-2.6	-12.4	6.6	57.4	0.9	8.7	5.7	48.7
		M	9.5	55.3	2.6	10.6	6.9	44.7	2.1	58.5	3.6	23.2	-1.5	35.3	10.4	54.9	2.4	9.2	8.0	45.7
	+ AUX$_{SIM}$	F	24.6	30.3	17.6	30.5	7.0	-0.2	21.5	24.8	20.3	35.0	1.2	-10.2	38.3	55.1	5.4	10.1	32.9	45.0
		M	38.1	51.4	8.7	9.9	29.4	41.5	41.5	47.0	19.4	27.4	22.1	19.6	37.7	51.9	7.4	7.9	30.3	44.0
	+ ADV$_{LAN}$	F	30.8	30.6	30.3	34.7	0.5	-4.1	24.8	24.1	34.8	40.3	-10.0	-16.2	57.8	59.6	10.0	9.5	47.8	50.1
		M	57.2	55.2	10.5	9.2	46.7	46.0	52.7	50.2	30.2	28.9	22.5	21.3	57.8	55.8	8.3	7.0	49.5	48.8
	Residual	F	30.0	33.3	26.8	31.4	3.2	1.9	24.7	27.0	30.8	36.6	-6.1	-9.6	53.6	61.7	8.6	7.6	45.0	54.1
		M	49.1	53.2	11.3	12.5	37.8	40.7	37.3	56.1	30.0	28.9	7.3	27.2	50.5	52.8	9.2	10.6	41.3	42.2
	+ AUX$_{SIM}$	F	33.9	31.5	24.0	30.8	9.9	0.7	28.4	26.7	27.7	35.1	0.7	-8.4	58.9	53.5	7.0	11.3	51.9	42.2
		M	49.5	48.2	11.1	11.6	38.4	36.6	44.8	40.6	26.2	29.4	18.6	11.2	50.0	49.0	9.4	9.6	40.6	39.4
	+ ADV$_{LAN}$	F	35.2	34.0	28.0	30.4	7.2	3.6	30.1	27.8	31.5	35.5	-1.4	-7.7	58.2	61.9	12.1	7.6	46.1	54.3
		M	52.2	52.6	12.2	13.1	40.0	39.5	51.1	50.5	32.4	28.9	18.7	21.6	52.3	52.8	9.9	11.3	42.4	41.5
de	Baseline	F	21.3	34.1	15.4	22.1	5.9	12.0	16.7	28.8	16.6	26.3	0.1	2.5	42.1	58.2	9.9	3.1	32.2	55.1
		M	32.5	50.6	6.1	10.0	26.4	40.6	34.0	43.1	15.8	23.7	18.2	19.4	32.3	51.5	5.0	8.4	27.3	43.1
	+ AUX$_{SIM}$	F	25.8	33.1	19.0	21.5	6.8	11.6	22.0	29.3	20.4	24.5	1.6	4.8	43.1	50.5	12.7	7.8	30.4	42.7
		M	33.4	45.7	8.1	9.3	25.3	36.4	35.9	43.6	20.9	24.5	15.0	19.1	33.1	46.0	6.7	7.6	26.4	38.4
	+ ADV$_{LAN}$	F	29.8	34.3	22.3	22.4	7.5	11.9	24.6	28.2	24.9	26.6	-0.3	1.6	53.2	62.1	10.8	3.6	42.4	58.5
		M	46.2	46.9	8.0	9.2	38.2	37.7	44.7	43.8	19.3	25.2	25.4	18.6	46.4	47.3	6.7	7.3	39.7	40.0
	Residual	F	33.3	33.4	21.0	23.2	12.3	10.2	28.9	27.4	23.1	26.7	5.8	0.7	53.4	60.3	11.6	7.3	41.8	53.0
		M	48.2	50.0	9.1	10.0	39.1	40.0	45.0	39.7	24.1	27.1	20.9	12.6	48.6	51.1	7.4	8.1	41.2	43.0
	+ AUX$_{SIM}$	F	33.4	32.8	23.2	24.0	10.2	8.8	29.7	27.3	25.7	27.7	4.0	-0.4	49.8	57.5	11.8	7.3	38.0	50.2
		M	47.0	45.9	10.6	9.3	36.4	36.6	48.5	42.2	24.5	23.3	24.0	18.9	46.8	46.3	9.0	7.7	37.8	38.6
	+ ADV$_{LAN}$	F	32.2	35.7	21.2	20.9	11.0	14.8	26.9	30.4	23.6	24.3	3.3	6.1	55.9	59.7	10.7	5.5	45.2	54.2
		M	48.5	50.9	8.9	9.0	39.6	41.9	42.4	43.0	25.1	21.3	17.3	21.7	49.2	51.9	7.1	7.5	42.1	44.4
es	Baseline	F	39.8	47.0	12.5	17.4	27.3	29.6	37.0	44.1	13.1	19.0	23.9	25.1	52.5	60.1	9.5	10.0	43.0	50.1
		M	46.3	54.7	7.9	9.8	38.4	44.9	51.0	57.3	14.3	16.8	36.7	40.5	45.8	54.4	7.2	9.0	38.6	45.4
	+ AUX$_{SIM}$	F	44.4	45.4	12.1	14.2	32.3	31.2	41.3	43.1	13.2	15.8	28.1	27.3	58.5	56.0	7.2	7.2	51.3	48.8
		M	49.1	47.9	9.4	10.0	39.7	37.9	57.7	55.9	21.6	16.8	36.1	39.1	48.2	47.0	8.0	9.2	40.2	37.8
	+ ADV$_{LAN}$	F	46.0	47.7	11.8	15.7	34.2	32.0	44.1	45.4	12.4	17.0	31.7	28.4	54.5	57.9	9.1	9.5	45.4	48.4
		M	52.2	50.7	10.0	9.8	42.2	40.9	62.6	59.9	18.6	16.3	44.0	43.6	51.0	49.7	9.0	9.0	42.0	40.7
	Residual	F	42.4	42.6	18.1	22.9	24.3	19.7	39.8	38.2	19.6	25.9	20.2	12.3	54.2	62.7	11.6	9.2	42.6	53.5
		M	54.0	55.7	9.0	9.9	45.0	45.8	55.4	55.6	20.2	18.7	35.2	36.9	53.9	55.7	7.8	8.9	46.1	46.8
	+ AUX$_{SIM}$	F	41.0	39.7	17.0	19.9	24.0	19.8	38.6	36.7	18.7	22.6	19.9	14.1	51.4	53.4	9.5	7.7	41.9	45.7
		M	53.5	45.5	9.1	10.6	44.4	34.9	58.3	46.3	20.2	19.2	38.1	27.1	53.0	45.4	7.8	9.6	45.2	35.8
	+ ADV$_{LAN}$	F	38.9	39.0	19.9	21.0	19.0	18.0	34.5	35.7	22.4	22.7	12.1	13.0	58.9	54.2	8.2	13.2	50.7	41.0
		M	55.0	51.5	10.2	10.1	44.8	41.4	64.1	57.8	16.4	18.7	47.7	39.1	54.0	50.8	9.5	9.1	44.5	41.7

Table B.7: Utterance Category 2: Average accuracy scores for *Correct* (higher ↑ is better) and *Wrong* (lower ↓ is better) MuST-SHE references with speaker-independent gender agreement, complemented with the difference (Delta ↑ = *Correct*− *Wrong*) of the scores for the two reference sets. Results for both gender word forms (All) are additionally broken down by referent gender (Feminine, Masculine) and by speaker gender (Female or Male). For the *Correct* scores and the Deltas, bold scores denote the best results per approach and underlined are the best of both approaches if one is better.

– Utterance Category 2 (Accuracy) –																		
Training Data X ↔	Model	Speaker	**All**						**Feminine**						**Masculine**			
			Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑		Correct ↑		Wrong ↓		Delta ↑	
			ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV	ZS	PV
en	Baseline	F	3.5	44.0	2.6	20.8	0.9	23.2	2.1	37.8	3.4	26.4	-1.3	11.4	6.6	57.4	0.9	8.7	5.7	48.7
		M	10.6	56.8	2.7	10.4	7.9	46.4	2.1	58.5	3.6	23.2	-1.5	35.3	12.4	56.4	2.5	7.6	9.9	48.8
	+ AUX$_{SIM}$	F	30.3	40.9	12.9	20.2	17.4	20.7	26.7	34.3	16.4	25.0	10.3	9.3	38.3	55.1	5.4	10.1	32.9	45.0
		M	39.2	52.6	8.1	10.7	31.1	41.9	41.5	47.0	19.4	27.4	22.1	19.6	38.7	53.8	5.6	7.1	33.1	46.7
	+ ADV$_{LAN}$	F	40.9	41.2	22.6	24.1	18.3	17.1	33.1	32.6	28.5	30.9	4.6	1.7	57.8	59.6	10.0	9.5	47.8	50.1
		M	57.7	55.0	11.9	10.4	45.8	44.6	52.7	50.2	30.2	28.9	22.5	21.3	58.8	56.0	7.9	6.3	50.9	49.7
	Residual	F	39.3	44.5	19.6	20.8	19.7	23.7	32.6	36.5	24.7	27.0	7.9	9.5	53.6	61.7	8.6	7.6	45.0	54.1
		M	49.4	57.0	13.3	12.8	36.1	44.2	37.3	56.1	30.0	28.9	7.3	27.2	52.0	57.2	9.7	9.2	42.3	48.0
	+ AUX$_{SIM}$	F	46.0	40.8	16.9	22.1	29.1	18.7	40.0	34.9	21.6	27.1	18.4	7.8	58.9	53.5	7.0	11.3	51.9	42.2
		M	50.8	50.6	11.9	12.1	38.9	38.5	44.8	40.6	26.2	29.4	18.6	11.2	52.1	52.8	8.7	8.3	43.3	44.5
	+ ADV$_{LAN}$	F	46.4	45.2	21.8	21.0	24.6	24.2	41.0	37.4	26.3	27.2	14.7	10.2	58.2	61.9	12.1	7.6	46.1	54.3
		M	55.2	55.5	14.2	12.4	41.0	43.1	51.1	50.5	32.4	28.9	18.7	21.6	56.1	56.6	10.2	8.7	45.9	47.9
de	Baseline	F	30.7	46.4	11.3	13.1	19.4	33.3	25.3	40.9	12.0	17.8	13.3	23.1	42.1	58.2	9.9	3.1	32.2	55.1
		M	37.5	48.6	8.0	10.7	29.5	37.9	34.0	43.1	15.8	23.7	18.2	19.4	38.2	49.9	6.3	7.9	31.9	42.0
	+ AUX$_{SIM}$	F	35.0	43.1	15.3	15.5	19.7	27.6	31.2	39.6	16.6	19.1	14.6	20.5	43.1	50.5	12.7	7.8	30.4	42.7
		M	37.2	44.7	10.2	9.9	27.0	34.8	35.9	43.6	20.9	24.5	15.0	19.1	37.5	45.0	7.8	6.7	29.7	38.3
	+ ADV$_{LAN}$	F	41.1	49.0	16.6	14.8	24.5	34.2	35.5	42.9	19.3	20.0	16.2	22.9	53.2	62.1	10.8	3.6	42.4	58.5
		M	47.8	49.1	9.9	11.0	37.9	38.1	44.7	43.8	19.3	25.2	25.4	18.6	48.5	50.3	7.8	7.9	40.7	42.4
	Residual	F	44.3	47.0	15.6	14.0	28.7	33.0	40.0	40.8	17.4	17.2	22.6	23.6	53.4	60.3	11.6	7.3	41.8	53.0
		M	47.7	49.9	11.7	11.4	36.0	38.5	45.0	39.7	24.1	27.1	20.9	12.6	48.3	52.1	9.0	8.0	39.3	44.1
	+ AUX$_{SIM}$	F	43.8	45.9	17.0	16.3	26.8	29.6	41.1	40.5	19.4	20.5	21.7	20.0	49.8	57.5	11.8	7.3	41.8	53.0
		M	48.7	45.5	12.1	10.3	36.6	35.2	48.5	42.2	24.5	23.3	24.0	18.9	48.7	46.2	9.4	7.5	39.3	38.7
	+ ADV$_{LAN}$	F	45.0	48.8	15.6	12.7	29.4	36.1	40.0	43.7	17.9	16.1	22.1	27.6	55.9	59.7	10.7	5.5	45.2	54.2
		M	46.5	51.8	11.0	9.6	35.5	42.2	42.4	43.0	25.1	21.3	17.3	21.7	47.4	53.8	7.9	7.0	39.5	46.8
es	Baseline	F	44.9	51.1	11.5	15.0	33.4	36.1	41.3	46.9	12.4	17.4	28.9	29.5	52.5	60.1	9.5	10.0	43.0	50.1
		M	49.8	56.7	9.4	10.8	40.4	45.9	51.0	57.3	14.3	16.8	36.7	40.5	49.5	56.5	8.3	9.4	41.2	47.1
	+ AUX$_{SIM}$	F	48.3	48.1	12.4	12.9	35.9	35.2	43.6	44.4	14.8	15.6	28.8	28.8	58.5	56.0	7.2	7.2	51.3	48.8
		M	51.5	50.8	9.8	10.2	41.7	40.6	57.7	55.9	21.6	16.8	36.1	39.1	50.1	49.7	7.2	8.7	42.9	41.0
	+ ADV$_{LAN}$	F	49.7	49.3	10.4	15.4	39.3	33.9	47.5	45.2	11.0	18.2	36.5	27.0	54.5	57.9	9.1	9.5	45.4	48.4
		M	56.7	53.7	10.2	10.1	46.5	43.6	62.6	59.9	18.6	16.3	44.0	43.6	55.3	52.3	8.3	8.7	47.0	43.6
	Residual	F	50.0	48.9	14.1	17.6	35.9	31.3	48.0	42.5	15.3	21.4	32.7	21.1	54.2	62.7	11.6	9.2	42.6	53.5
		M	54.7	57.2	9.1	11.1	45.6	46.1	55.4	55.6	20.2	18.7	35.2	36.9	54.6	57.6	6.7	9.4	47.9	48.2
	+ AUX$_{SIM}$	F	46.7	45.3	13.3	15.9	33.4	29.4	44.5	41.5	15.1	19.7	29.4	21.8	51.4	53.4	9.5	7.7	41.9	45.7
		M	55.2	48.6	9.3	11.7	45.9	36.9	58.3	46.3	20.2	19.2	31.8	27.1	54.5	49.1	6.9	10.0	47.6	39.1
	+ ADV$_{LAN}$	F	50.3	46.2	12.5	17.0	37.8	29.2	46.3	42.5	14.5	18.8	31.8	23.7	58.9	54.2	8.2	13.2	50.7	41.0
		M	58.4	55.9	9.4	10.0	49.0	45.9	64.1	57.8	16.4	18.7	47.7	39.1	57.2	55.4	7.9	8.1	49.3	47.3