Computational Linguistics in the Netherlands Journal 14 (2025) 215-239 Submitted 12/2024; Published 07/2025

Gender Bias and the Role of Context in Human Perception and Machine Translation

Janiça Hackenbuchner JANICA.HACKENBUCHNER@UGENT.BE
Arda Tezcan ARDA.TEZCAN@UGENT.BE
Joke Daems JOKE.DAEMS@UGENT.BE

Language and Translation Technology Team, Ghent University,
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

Abstract

This paper investigates human gender bias and its relation to bias in machine translation (MT), focussing on the role of context in gender interpretation. To this end, we measured human implicit gender bias and conducted an annotation study, followed by a linguistic and computational analysis to compare human gender perceptions among themselves and with a machine translation system. We created a dataset of 60 gender-ambiguous sentences and collected annotations to understand human gender perceptions and specifically which trigger words in context lead to this perception. The study shows that, unlike the MT system tested in this study, humans exhibit highly varied perceptions of gender in ambiguous contexts. A linguistic analysis on annotated trigger words reveals that proper nouns, nouns and adjectives frequently affect human gender perception.

1. Introduction

Research on Neural Machine Translation (NMT) and Large Language Models (LLMs) used for translation agrees that these AI-based language technologies continue to exhibit gender bias, meaning that they discriminate against certain individuals, while favouring others (Friedman and Nissenbaum 1996, Kotek et al. 2023, Vanmassenhove 2024). Fundamentally, bias is an inherently useful characteristic for machine learning systems to generalise on unseen data and predict the statistically most probable output (Mitchell 1980). Systematically useful as this may be, models not only exhibit but amplify biases present in the training data and therefore generate biased outputs (Sun et al. 2019, Mehrabi et al. 2021). Biased MT outputs foster the continued (human and machine) consumption and creation of biased data, further perpetuating social bias.

Among biases in language technologies, gender is a particularly difficult phenomenon for MT systems to handle due to "contrastive linguistic settings that necessitate disambiguation and explicitness in their representation of gender" (Vanmassenhove 2024). When translating from a notional gender language, with limited grammatical gender specification, to a grammatical gender language, where gender is strictly specified grammatically, state-of-the-art MT systems default to stereotypical generalisations learned from gender information embedded in word embeddings (see Section 2). Not unlike MT, humans construct their perception of gender based on grammar in language, whereas in the absence of grammatical cues, they base their gender perception on stereotype information (Gygax et al. 2008). Ambiguous gender scenarios in a (source) text therefore lead to gender bias both when it forces current MT systems to opt for stereotypical (binary) gender norms in the translation process or when it leads to stereotypical human perceptions. These gender implications both for human perceptions and system translations are further analysed in this paper.

Specifically, we investigate human gender perceptions in gender-ambiguous contexts and how this compares to machine translations for the English to German language direction. We focus on how individual factors influence implicit human bias and what we can learn from resulting gender perceptions. Considering contextual cues in the source have been shown to influence how MT systems assign gender in the target language (Kocmi et al. 2020), the underlying idea is to understand what

context in sentences influences human perceptions of gender and whether this can help us better understand gender bias in MT systems. However, not all words in the context are useful in translating a current sentence (Kim et al. 2019) and specific linguistic features in the source make it more difficult for MT to translate without it being clear 'why' this is the case (Don-Yehiya 2022). With a focus on gender, this research addresses these issues by analysing which specific context influences gender perception and translation and 'why' this is the case.

To this end, a manual annotation study was conducted, analysing which (type of) words in context trigger human perceptions of gender. Furthermore, we noted how an MT system (DeepL) translated the same context in terms of gender. Human annotation results were compared to these gendered machine translations. The research questions that guided this study are:

- (RQ1) How do individual factors influence human gender perception?
- (RQ2) To what extent do humans agree on gender perception in the absence of clear (grammatical) cues?
- (RQ3) To what extent does a machine translation system agree with human gender perceptions?
- (RQ4) Based on perceptions from humans that agree most with MT, can we form hypotheses on what the factors are that influence MT bias in gender translations?

The structure of this paper is organised as follows: Section 2 summarises definitions of machine and human gender bias and provides an overview of research conducted on gender bias in MT, gender manifested in word embeddings, and a human-model bias comparison. Section 3 outlines the methodology taken for data collection, machine translation, human bias measure, and manual annotations. Section 4 provides detailed analyses of individual factors that influence human gender perceptions, inter-annotator agreement among annotators as well as for MT-human comparison, and a linguistic analysis of the context that influenced gender perceptions. Section 5 critically summarises the main findings of this article, providing tentative conclusions to the research questions, and points to directions for future work. Data and code are publicly available on GitHub.[1]

2. Related Research

The research presented here focusses both on model and on human gender bias, and a comparison of the two. The term bias in itself can be defined from different angles. From an ethical and societal point of view, systems are said to be biased if they "*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum 1996). In natural language processing (NLP), bias is defined as an unfairness in algorithms, an aggregate effect for demographic groups, and as potentially harming the data (Shah et al. 2020). Models exhibit bias by making "predictions based on gender distribution in the training data, rather than relations between entities in the sentence" (Lior and Stanovsky 2023). In this paper, we will have a closer look at these "relations between entities in the sentence", which (could) lead to bias.

Mehrabi et al. (2021) explain that real-world biases "seep into" the data generation processes. These 'real-world' biases found in training data stem from societal human bias, which is frequently present in the form of 'implicit' bias, referring to "automatic mental associations that can shape our perceptions and decisions without our conscious awareness" (*Project Implicit* n.d.). Humans may therefore have attributes and beliefs that they may be unwilling or unable to identify themselves and that can unconsciously influence their actions and decisions (*Project Implicit n.d.)*.

In natural language, research shows that role names and general stereotypes strongly influence human perceptions of gender (Lardelli and Gromann 2023, Misersky et al. 2014, Gygax et al. 2008).

1. `https://github.com/jhacken/MTGenderContext`

Humans construct their perception of gender based on grammar in language (e.g., *waitress*), however, when lacking such grammatical cues, they base their gender perception on stereotype information (Gygax et al. 2008). Grammatical gender languages such as German (Stahlberg et al. 2007) refer to people in their generic masculine, which is intended to be *generic* but is not typically interpreted as such (Gygax et al. 2008). Where language lacks grammatical gender cues such as pronouns or other gender referents, the reader usually defines the gender as generic masculine or assumes the gender based on previous knowledge or stereotypes. Similarly, MT systems have been shown to primarily translate into masculine (Monti 2020), excluding half the world's population (Vanmassenhove et al. 2018). This paper analyses on which specific contextual information humans, and MT systems, base their gender perceptions or translations in the absence of grammatical cues.

In an attempt to mitigate gender bias in MT systems, challenge sets have been created aimed at balancing the gender of stereotypical professions and adjectives that have been shown to lead to biased translations (Troles and Schmid 2021, Stanovsky et al. 2019, Zhao et al. 2018). Such challenge sets are frequently balanced to contain a female and a male (and in a few cases also non-binary) sentence, as in *"The choreographer finished her work. / The choreographer finished his work. / The choreographer finished their work."* to de-bias an MT system (Saunders and Byrne 2020). In recent years, research on gender bias in MT has extended to include non-binary genders or neutralising sentences (Savoldi et al. 2024, Lardelli and Gromann 2023, Saunders et al. 2021). However, studies on gender in monolingual English data show that, computationally, gender is inherently manifested in a vast range of words, as unveiled by the gender-inflection of their word embeddings (Caliskan et al. 2022, Bolukbasi et al. 2016), not merely in a select number of stereotypical professions and adjectives. Research on word level therefore shows that a large number of word embeddings, which MT systems are trained on, carry an inherent gender inflection. Research on sentence level shows that occupation nouns and adjectives greatly influence the gender inflection in a machine-translated target sentence.

In a recent study on the comparison of human and model evaluations of gender bias, Lior and Stanovsky (2023) showed that, under constrained settings, humans make (sometimes wrong) predictions based on societal and cognitive presupposition and that "model biases reflect human decision-making". Interestingly, they analysed gender bias on both natural and synthetic data (Winogard, WinoBias, and BUG (Zhao et al. 2018, Rudinger et al. 2018, Levy et al. 2021)) and showed that humans exhibit stronger gender bias in naturally-occurring sentences. They conclude that gender bias in coreference resolution is comparable to human biases (Lior and Stanovsky 2023).

Similarly, Zhu et al. (2024) conducted a study to analyse whether humans (annotators) and language models (LMs) share the same implicit gender bias. Zhu et al. (2024) designed a framework to measure human "contextualised gender bias" (CGB) based on specific questions they designed as part of a framework. Human implicit bias was based on the participants' results to these questions. Their methodology consisted of collecting natural sentences in which they masked gendered terms and then asked the LMs and annotators to choose from provided candidates to fill in the missing gendered term. Focussing on annotators that had a *low* implicit gender bias, they show that LMs and annotators tend to make similar choices, i.e. that both annotators and LMs exhibit low bias in their tasks (Zhu et al. 2024). However, there seems to be a distinction between LMs, where those models that qualitatively perform better, i.e. are more accurate and efficient, tend to exhibit more bias. The research presented here similarly assesses the 'contextualised gender bias' of humans and machines but adopts different measures to assess implicit bias, compares (strong) human bias to MT systems and implements a bottom-up approach to assess which context affects bias.

In previous work (Hackenbuchner et al. 2024), we conducted a comparative study of machine translation and human annotation of role names (words referring to persons) in isolation, out-of-context, and in ambiguous sentence contexts, with the absence of grammatical gender cues. They show that humans frequently (on average, 44% of the time) changed their gender perception of a role name after seeing it in an ambiguous sentence context, in the absence of (grammatical) gender cues. Following up on previous research analysing different angles of gender bias, the interesting gap to fill

is to what extent (which) words in an ambiguous sentence context influence human perceptions of gender, how this is affected by human implicit bias, and how this compares to MT systems. The aim is to analyse whether, in the translation process, MT systems could be affected by the same context that affects human perceptions of gender and whether this information could help us mitigate gender bias by making MT systems more gender-fair and better represent its breadth of users.

3. Task Descriptions

This paper combines and extends the word-level and sentence-level approaches taken, outlined in Section 2, through a bottom-up methodology to focus on how sentence context influences gender. To this end, human perceptions of gender in context are compared to gender present in machine translations. To analyse the impact of context on gender, ambiguous English sentences (i.e., sentences with an absence of grammatical gender cues) were collected and both annotated by humans as well as machine translated into German using DeepL. English is a notional gender language (McConnell-Ginet 2013), where role names generally do not have a gender assigned (e.g., *poet*) apart from kinship relations (*mother, father*) or a few exceptions (*actor, actress*). In comparison, German is a grammatical gender language, where role names generally do have a gender assigned (Stahlberg et al. 2007).

3.1 Data Collection

Existing datasets to test gender bias in machine translation, or generally model behaviour, such as WinoBias and WinoGender (Zhao et al. 2018, Rudinger et al. 2018) can be well-used for controlled experiments. Yet, they consist of a small variety of linguistic constructions (Lior and Stanovsky 2023) and are therefore unrepresentative of real-world, naturally occurring data. To have both a linguistic and topic variety and to resemble real-world scenarios, this research is based on natural data, instead of synthetic or handcrafted data, as methodologically also done in Zhu et al. (2024). The English language has been widely studied with respect to gender bias in coreference resolution with respect to machine learning and psycholinguistics, making it possible for this research to refer to previous studies.

The first step was the creation of a new data set. English data was sampled from monolingual English corpora (StatMT's news-crawl[2], as well as c4 (Raffel et al. 2019) and wiki (Wikimedia Foundation n.d.) as made available on HuggingFace[3]). To sample specific data from these corpora, we defined a list of *seed words* that fulfil certain criteria, as needed for the sampling in question. For the study conducted here, our seed words are 165 role names, words referring to singular persons (e.g., *poet*). 106 of these were chosen based on the gender inflection in their static word embeddings (being inflected as either male or female - 59 and 47 respectively), as collected during previous research (Caliskan et al. 2022, Bolukbasi et al. 2016, Stanovsky et al. 2019). Another 59 words were chosen by prompting ChatGPT to generate gender-neutral words, as shown in Appendix A. From the monolingual English corpora, we automatically sampled sentences containing these seed words, combined with a coreference filter to exclude sentences containing pronouns referring to the seed word while keeping pronouns referring to other entities in the text.

We then manually filtered our sampled dataset to ensure only gender-ambiguous sentences in relation to the seed word were kept. Of the compiled data, a sample of 60 sentences covering a range of topics was selected for this qualitative annotation task. 49 individual role names were represented in the 60 sentences, as some occurred in different sentence contexts, as can be seen by the role name *therapist* in the following two contexts:

- Kensington massage **therapist** jailed for sexually assaulting clients.

2. `https://data.statmt.org/news-crawl/`
3. `https://huggingface.co/`

- There are 52 weeks in a year, my **therapist** continued matter-of-factly, "I know you can't go on a date every single week, but how many do you think you should be going on?"

The sporadic occurrence of role names in varied sentences allowed for a comparison of how different contexts affect gender perception or translation. A more detailed description of how the data was chosen and collected is described in Hackenbuchner et al. (2024).

3.2 Machine Translation

In this case study, DeepL was chosen as the MT system as it is, according to the latest Intento report, the highest-rated machine translation system that "consistently outperform[s] other models"[4], particularly for German, where it is widely used, but nevertheless it continues to exhibit gender bias. German was chosen as the target language as it is one of the top two most-spoken foreign languages in Europe[5], a grammatical gender language with apparent issues of gender bias in machine translation. While working into German via DeepL, this is a case study, where the focus lies on the source language and text.

The sampled data of 60 ambiguous English sentences was machine-translated into German using the DeepL API in November 2024. The gender-ambiguous seed word in each sentence was thus translated into a grammatical gender language. A tally was kept into which (binary) gender the machine translation translated these seed words in German. The two sentences referring to a *therapist*, as outlined above, were translated into German as follows:

- Massage**therapeutin** aus Kensington wegen sexueller Belästigung von Kunden inhaftiert.
- Das Jahr hat 52 Wochen, fuhr mein **Therapeut** sachlich fort: "Ich weiß, dass Sie nicht jede Woche zu einem Date gehen können, aber wie viele sollten es denn sein?"

In the German translations above, DeepL translated the *therapist* referred to in the first sentence as female, and in the second sentence as male. This classification of gendered translations (seed word translated as male or female) was used to calculate a human-MT agreement, as outlined in Section 4.2.1. The aim is to analyse which contextual information in the sentence could have triggered the MT to opt for a certain gender, in the absence of clear gender cues referring to that seed word.

3.3 Human Annotations

MT systems have been shown to exhibit gender bias, leading research to aim for more 'human-like' translations that are less data-rigid or stereotypical. However, humans exhibit gender bias themselves, often very subjectively so. Prior to collecting gender perceptions (in the form of annotations) we conducted implicit bias tests to measure human gender bias subjectivity and to see to what extent this influences their perception of gender (in the annotation task). The aim is to analyse whether human perceptions of gender can help us understand what MT systems (should) take into consideration when translating gender.

3.3.1 Participants

To collect a varied selection of human perceptions, a total of 21 annotators from different genders and countries were recruited. Of the 21 annotators, 10 were male, 9 were female, and 2 were non-binary. All annotators were highly proficient in English and stem from one of the following countries of origin (in alphabetical order): Belgium, Brazil, Bulgaria, France, Germany, India, the Netherlands, Russia, Turkey, and the United Kingdom. The annotator diversity played a key role in balancing gender and demographics, as the tasks were highly subjective and individual.

4. `https://inten.to/machine-translation-report-2024/`
5. ://europa.eu/eurobarometer/surveys/detail/2979

Prior to the annotations, all annotators were duly informed of the content and procedure including their role as annotators and were requested to sign an informed consent form, which had been approved by the ethics committee at the Faculty of Arts and Philosophy, Ghent University, allowing for their annotations to be analysed within the context of this study.

3.3.2 Implicit Association Tests

Before conducting annotations, each annotator was asked to complete two Implicit Association Tests (IATs) from the Project Implicit Social Attitudes from Harvard[6]. The IAT tests measure a person's attributes and beliefs that they may be unwilling or unable to identify themselves. Through automatic associations, the IATs thus bring implicit attitudes to the front, that a person might not be conscious of. A person's automatic associations are evaluated by measuring their response time of how strongly they associate certain concepts with each other. The annotators had to take two IATs, the Gender-Career IAT and Gender-Science IAT, which measure a person's association of women and men with respect to (i) career or family, or with respect to (ii) science or liberal arts, respectively. For each of the two selected IATs, seven results were possible ranging from *strong association for men in science and women in liberal arts* to *strong association for women in science and men in liberal arts* (with the associations being either strong, medium, slight or little). An analysis of the IAT results taken by the annotators and to what extent they influenced their associations is presented in Section 4.1.1.

3.3.3 Annotations

Following the completion of the two IAT tests, the annotators were asked to annotate the sampled dataset of 60 English gender-ambiguous sentences. These in-context annotations were done on the annotation platform *Label Studio*[7]. An example of an annotation task is visualised in Figure 1. The annotation task consisted of two parts:

1. Annotate gender associated with the seed word (gender perception)
2. Annotate other words in the same sentence that trigger this gender perception

In the first task, annotators were asked to annotate a gender they associated with the seed word, the role name referring to a person in the sentence. As an example, in the following sentence "*Kensington massage therapist jailed for sexually assaulting clients.*", annotators had to label the seed word *therapist* with a gender that they associated with that word in this specific sentence context. For each annotation, annotators could choose from three gender options (as shown in Figure 1): 'female', 'male' and 'non-binary', or 'N/A' if they had no gender association or preference for that word in this specific sentence context. There were no *right* or *wrong* annotations.

Annotation task 1 had to be completed by all annotators. Based on annotation task 1, a comparison could be drawn for each annotator, on the one hand, to other annotators, and on the other hand, to the gendered machine translations.

In annotation task 2, annotators were asked to label other words that triggered their gender association in that specific sentence. This second task was an optional, bottom-up approach. Annotators were free to label any and all words that triggered their individual perception of gender in the sentence. Again, there were no *right* or *wrong* annotations.

In the example sentence above, commonly annotated words in context were *sexually assaulting* and *jailed*, which were frequently annotated as a 'male' trigger, or *massage*, which was frequently annotated as a 'female' trigger for the seed word *therapist*. Similarly, to label trigger words, annotators could again choose from 'female', 'male', 'non-binary', or 'N/A' if they had no gender association or

6. `https://implicit.harvard.edu/implicit/selectatest.html`
7. Label Studio `https://labelstud.io/`

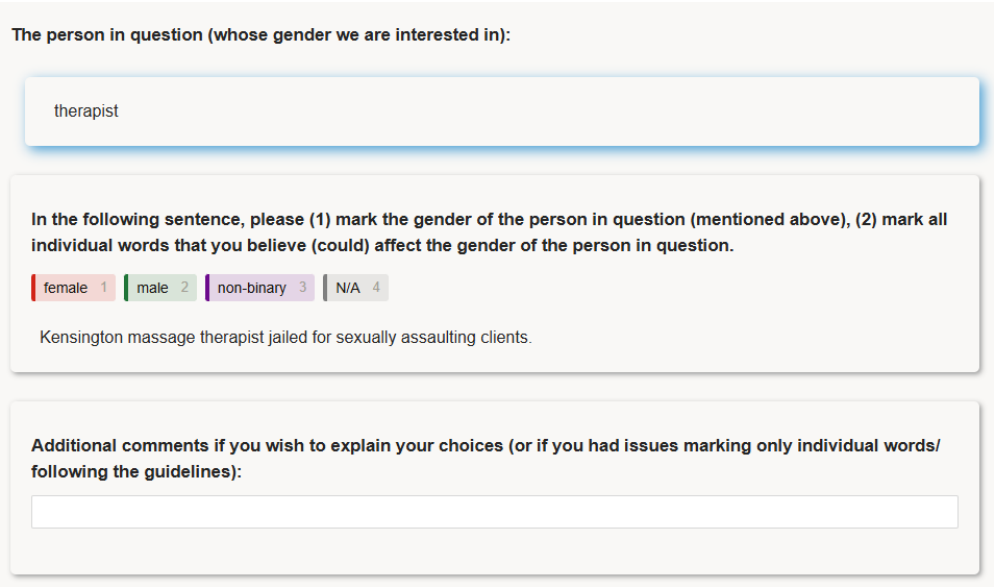

Figure 1: Example visualisation of the annotation task given to each annotator for each sentence.

preference. An analysis of which (type of) words in context were annotated as triggers for human gender perception is presented in Section 4.3.

4. Results

The results section comprises both an analysis of human bias and human perceptions of gender as well as their comparison to gendered machine translations. Based on how similar humans perceived gender and how the MT translated gender, a linguistic analysis of what annotated contextual information triggered human gender perceptions is provided in Section 4.3.

4.1 Human Analysis

This first analysis section provides tentative conclusions to RQ1: How do individual factors influence human gender perception? In Section 4.1.1 we present to which extent the annotation labels of the seed words in context were influenced by the annotators' own gender or by their implicit gender bias, as measured via two IAT tests. In Section 4.1.2 we present an inter-annotator agreement for the gender annotations of the seed word as an answer to RQ2: To what extent do humans agree on gender perception in the absence of clear (grammatical) cues?

4.1.1 Annotation Correlation

To measure the correlations between gender annotations and annotator gender or implicit bias, Cramér's V (Cramér 1999) was calculated. Cramér's V is a measure of association between two values, where 0 refers to 'no association' and 1 refers to 'complete association' between two values. A Cramér's V of below 0.2 refers to a 'weak association, between 0.2 and 0.6 refers to a 'moderate association' and above 0.6 refers to a 'strong association' (IBM Corporation 2024). A specific association interpretation is case-dependent, based on the degrees of freedom that are calculated in the process.

The measure of whether the annotator's own gender influenced their annotations yields a Cramér's V correlation of 0.14. Across all annotators, 'male' was the most frequent label (56.81% on average) annotated for seed words (as shown in Appendix B Figure 9). The primary (yet slight) difference is that both male and female annotators annotated mostly binary (85.36% on average), rarely 'non-binary', and specifically men annotated more 'N/A' than could be expected if there did not exist any relationship between the two categorical variables in question. In turn, non-binary annotators annotated more 'non-binary' and less 'N/A' than expected. The details of this analysis of how annotator genders, on average, correlate with the annotation label of the seed words are shown in Appendix B Figure 10.

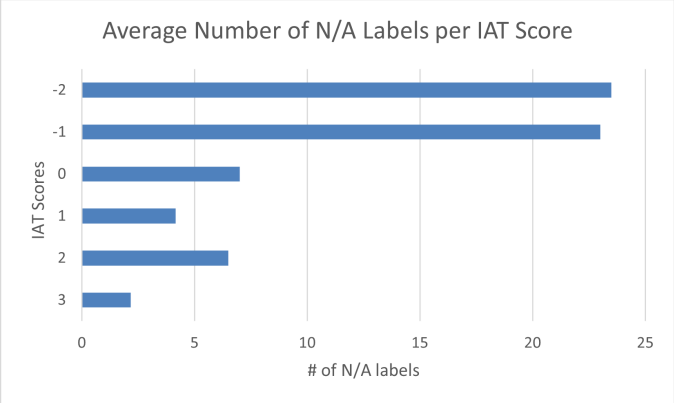

Figure 2: Correlation of Gender-Science IAT results (implicit gender bias) and how many seed words, on average, participants annotated with neutral 'N/A' labels. The IAT scores are converted into numericals (also relevant in the following subsections):
3 = strong association for men in science and women in liberal arts,
2 = moderate association for men in science and women in liberal arts,
1 = slight association for men in science and women in liberal arts,
0 = little or no association for men in science and women in liberal arts,
-1 = slight association for women in science and men in liberal arts,
-2 = moderate association for women in science and men in liberal arts,
-3 = strong association for women in science and men in liberal arts.

The measure of whether the annotators' implicit gender bias (i.e. their IAT results) influenced their annotations of the seed words yields a Cramér's V correlation of 0.18 for the 'Gender-Career' IAT and a Cramér's V of 0.24 for the 'Gender-Science' IAT. The correlation between the annotated labels and the 'Gender-Career' IAT is lower, from which we can see that annotators who had a higher 'Gender-Career' IAT result annotated slightly more 'male' labels than expected, and annotators with a lower IAT result of 0 or 1 annotated more 'N/A' labels than expected, and fewer binary labels.

The higher correlation, above the threshold of 0.2, is between how annotators labelled the seed word and their 'Gender-Science' IAT implicit bias result. Annotators with a low (female) implicit bias result (-1 or -2) annotated less binary than expected but noticeably more 'N/A'. This is visualised in Figure 2, which shows how frequently annotators annotated a seed word with a neutral 'N/A' label with respect to their IAT score. In comparison, the more (male) biased the annotators' IAT results, the more they annotated the seed words as either 'male' or 'female' and the less as 'N/A'. Overall, annotators that had a stronger (male) biased IAT result annotated more binary and less neutral. In comparison, annotators that had a lower (female) biased IAT result, annotated more neutral and less binary.

Furthermore, Figure 3 visualises a correlation between annotators' personal gender and their implicit gender bias (as measured by the 'Gender-Science' IAT). It can be seen that female annotators tended to have more (male) biased results (ranging mostly from 1 to 3), non-binary annotators had a slight to moderate (male) gender bias, and male annotators had the most diverse IAT results, with their implicit bias ranging from (female) biased to (male) biased.

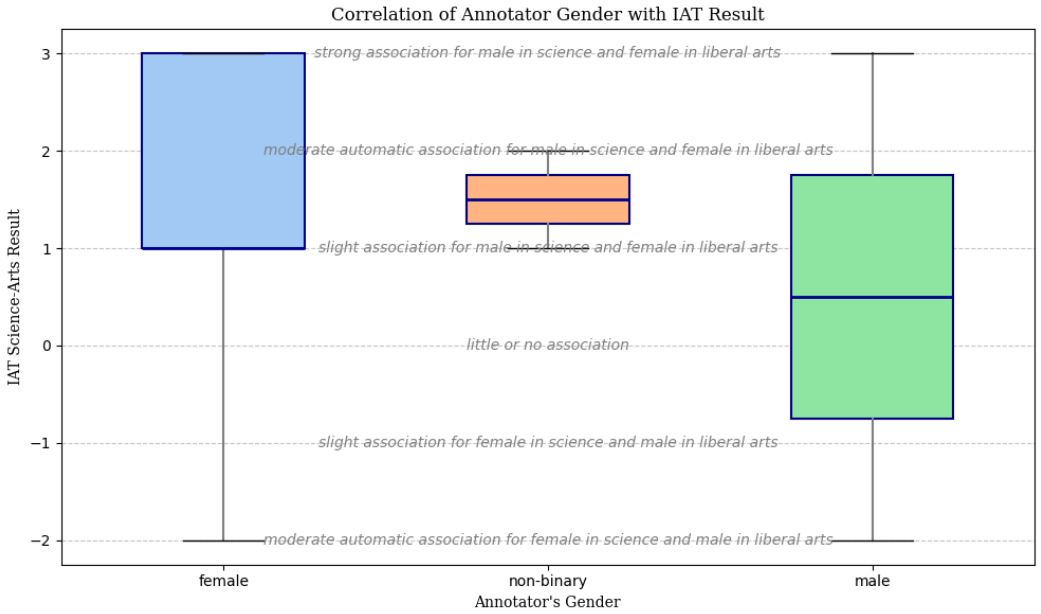

Figure 3: Correlation of the annotator's gender and their IAT 'Gender-Science' result.

4.1.2 Inter-Annotator Agreement

To measure how much humans agree among themselves, the inter-annotator agreement was calculated with Cohen's Kappa (κ) (Cohen 1960), with an average across all 60 sentences and all 21 annotators. Pairwise Cohen's Kappa was chosen as an agreement score as the focus here lies not only on an overall agreement but also on pairwise agreements between the annotators themselves, and between the annotators and the MT system presented in section 4.2.1. The IAA was measured for the gender label of the seed word (Annotation task 1). This IAA measure yields an overall pairwise average of Cohen's $\kappa = 0.383$. Cohen Kappa values reveal a 'moderate agreement' from a measure of 0.4 and a 'substantial agreement' from 0.6. The κ value measured here for gender seed word annotations therefore falls slightly short of showing a moderate agreement among all annotators. To compare, the overall agreement among annotators was also calculated with Fleiss' Kappa (κ) (Fleiss 1971), which resulted in Fleiss' $\kappa = 0.364$, similarly representing a 'fair agreement'.

While Cohen's Kappa average measure of all 21 annotators yields $\kappa = 0.383$, individual IAA between any two of these annotators varies from Cohen's $\kappa = 0.058$ to Cohen's $\kappa = 0.767$. Differences in IAA frequently arose between those annotators who annotated primarily binary or those who often annotated neutral (N/A). Overall, the differences in the gender label chosen for a seed word show noticeable heterogeneity among the annotators. The analysis of individual factors influencing annotations as well as of the inter-annotator agreement highlights the diversity among humans and their gender perceptions.

Example sentence (with the seed word marked in bold) of where the IAA was lowest:

- I'm not a **dancer**.

In this sentence, the personal reference *I* in the sentence influenced the annotator's perception of gender for the word *dancer*. Example sentence of where the IAA was highest:

- The trade is likely to involve a Russian arms **dealer** imprisoned in the US and a North American basketball player, recently detained in Russia.

In this sentence, the context *arms* and *imprisoned* triggered most annotators to perceive the *dealer* as male. A closer analysis of context triggers is outlined in Section 4.3.

4.2 Human-MT Analysis

One of the aims of this study is to analyse human gender perception as a proxy to better understand what context may trigger machine translation systems to produce gender-biased translations. Section 4.2.1 therefore focusses on a comparison between the human annotators and the gendered translations produced by the MT system, as an answer to RQ3: To what extent does a machine translation system agree with human gender perceptions? This section paves the way in providing informed hypotheses for RQ4: Based on perceptions from humans that agree most with MT, can we form hypotheses on what the factors are that influence MT bias in gender translations?

4.2.1 IAA Human-MT Agreements

In comparison to Section 4.1.2, where IAA was averaged between all annotators, this section presents an averaged IAA between each annotator and the MT system. Again, agreement is calculated based on how the seed word was labelled or how it was machine translated as an overall average over the 60 sentences and 21 annotators. Here, however, we face an incongruity, as the annotators could choose between three genders 'male', 'female' or 'non-binary', or a neutral 'N/A', but the machine translation system only provided binary translations, either as 'male' or as 'female'. It is known that commercial MT systems including DeepL do not yet translate into non-binary or inclusive gender, but a neutral 'N/A' translation could have been a possibility. The fact that MT continues to primarily translate into binary genders, with a dominance of the male gender whereas humans perceive gender much more varied and neutral, highlights that MT systems continue to be gender-biased, while humans perceive the world more diversely.

As visualised in Figure 4, the average IAA between annotators and the MT is Cohen's $\kappa = 0.356$. This is a similar average IAA value as for the IAA among annotators themselves. Comparably, the Fleiss Kappa agreement between annotators and the MT is Fleiss's $\kappa = 0.364$. Once again, a wide range of IAA can be seen with the lowest annotator-MT agreement being Cohen's $\kappa = 0.119$ (an example where the annotator labelled mostly 'N/A') and the highest annotator-MT agreement being Cohen's $\kappa = 0.579$ (where the annotator labelled mostly binary, and similar to the MT system). As could already be concluded from previous sections, the annotators vary greatly in terms of gender bias and how they perceive the gender of certain words in context. An example sentence (with the seed word marked in bold) of where the annotator-MT IAA was lowest:

- The wine **salesperson** on the floor might actually have tasted the wines and be able to point out intelligently which ones are fit for tonight's dinner and which ones should be socked away.

In this sentence, *salesperson* was annotated equally as 'male', 'female' or 'N/A', with *wine* and *intelligently* being frequent gender triggers. Example sentence of where the MT-annotator IAA was highest:

- It's a hard fall from grace for someone who worked hard as a lawyer and became a **judge** and is now a convicted felon, Oreskovich said.

Here, most annotators annotated *judge* as 'male' with common triggers being *lawyer* and *convicted felon*, and to a lesser extent *worked hard* and *fall from grace*.

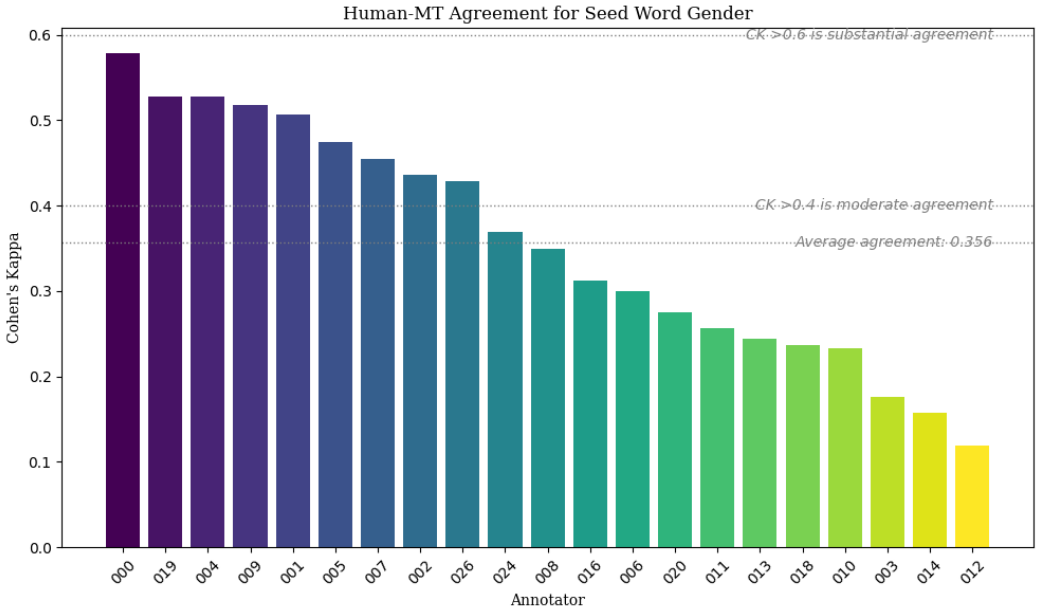

Figure 4: Inter-annotator agreement for Human vs. MT of gender label chosen for the seed/target word (MT gender translation, Human gender annotation). The annotators are visualised on the x-axis and their IAA with the MT (κ) is visualised on the y-axis.

Since a Cohen Kappa value above 0.4 reveals a 'moderate agreement' and above 0.6 a 'substantial agreement', Figure 4 shows that nine annotators had an IAA of 0.4 or higher, meaning that these nine annotators had at least a 'moderate agreement' with MT. It is important to notice the diversity in human gender annotations in this research, as visualised by these examples. However, as one of the aims of the research is to better understand how MT might be triggered towards gender bias, the focus of the analysis in the following sections will lie on these nine annotators who had at least a 'moderate agreement' with MT. Section 4.1.2 showed that the average IAA between all annotators was Cohen's $\kappa = 0.383$ (Fleiss's $\kappa = 0.364$ respectively), however, when looking at only these 'top 9' annotators, an IAA average of Cohen's $\kappa = 0.540$ (Fleiss's $\kappa = 0.541$ respectively) was measured, noticeably above the 'moderate agreement' threshold for both Cohen's and Fleiss's Kappa. Annotators that agreed mostly with MT also agreed most among themselves. The nine annotators encompass a variety of genders, among them one non-binary, four female, and four male annotators.

4.2.2 IATs vs. Human-MT Agreement

As summarised in Section 2, research agrees that MT is gender-biased and should be less so to more accurately reflect humans. To analyse how biased humans actually are, an important aspect of this research was to assess the annotators' implicit gender bias via two gender-focussed IATs. Figure 5 shows a correlation of the 'Gender-Science' IAT result of annotators with their MT-agreement. The results of the nine annotators that agreed mostly with MT is marked as black diamonds in Figure 5. Focussing on the 'top 9' annotators, an interesting pattern emerges: six of the nine annotators (0.67%) had a strong (male) implicit bias result of *3: strong association for men in science and women in liberal arts.* Merely two of the 'top 9' annotators showed no implicit gender bias with a result of *little or no association for men in science and women in liberal arts.* In comparison,

annotators that had the lowest agreement with MT, of Cohen's $\kappa = 0.12$ and $\kappa = 0.16$ as depicted on the right side of Figure 4, have low (female) implicit bias results of -1 and -2.

In other words, annotators with a stronger (male) implicit gender bias have the highest agreement with MT, and annotators with a lower (female) implicit gender bias have the lowest agreement with MT. Particularly on the left side of Figure 5, it is clearly depicted that those annotators that had the strongest implicit gender bias results (3) also had the highest agreements with MT (0.4 and above). That the more gender-biased humans have the most similar results to MT once again illustrates that MT systems produce gender-biased (and stereotypical) translations and that humans can be (very) gender-biased themselves.

4.3 Context Analysis

This research focusses on the contextual information (which is combined with stereotyping) that may sway an MT system to translate into one gender in favour of the other. With the aim to better understand MT systems, we present an analysis of what words in context triggered human perception of gender in ambiguous sentences. Following Section 4.2, this section aims to provide informed hypotheses for RQ4: Based on perceptions from humans that agree most with MT, can we form hypotheses on what the factors are that influence MT bias in gender translations?

4.3.1 Agreement on Trigger Words

Similarly to the IAA calculated for the annotations of the seed word, an agreement between annotators was calculated for the span annotations of trigger words. For this scenario, the gamma agreement, a chance-adjusted measure of the agreement between annotators, was chosen (Mathet 2015). Gamma compares the annotated spans between annotators as an alignment of units both on position

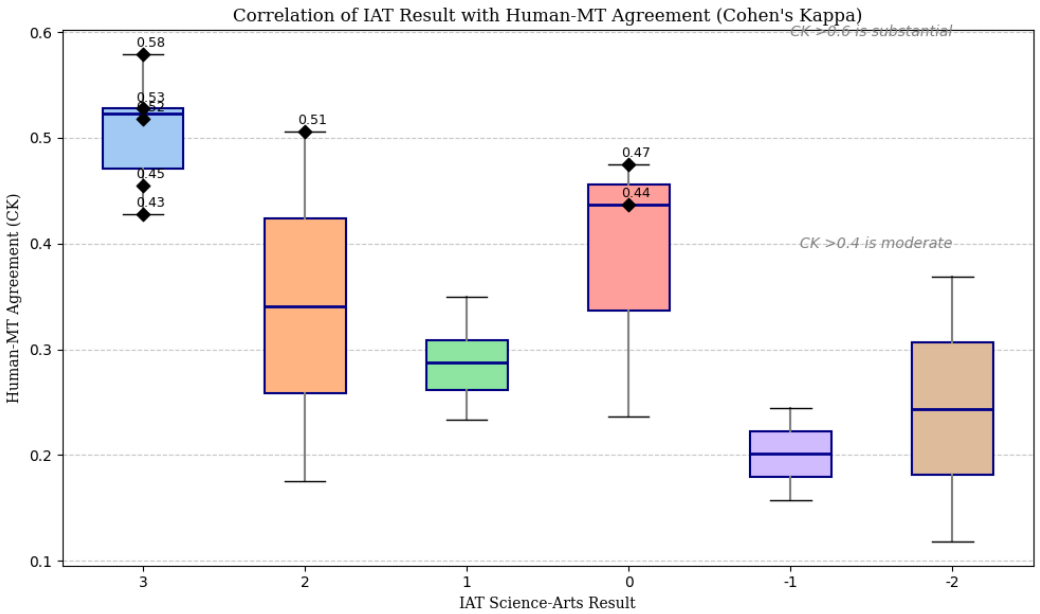

Figure 5: Correlation of IAT Science-Arts test result with human-MT agreement (Cohen's Kappa). The nine annotators that had the highest agreement with MT are highlighted as black diamonds in the plot.

(segment boundaries of the annotations) as well as on annotation category (the label of the annotated segment) (Mathet 2015). Gamma agreement is calculated from the best span alignment, which is computed based on dissimilarity and continuum between annotators. A continuum stores the set of annotations produced by several annotators for each annotated file. An example of a continuum is visualised in Appendix C Figure 11. A dissimilarity is a function that shows to what degree a unitary alignment between annotators "should be considered as different, taking into account such features as their positions, their annotations, or a combination of the two" (Mathet 2015). Gamma agreement measures are bounded by [-inf, 1], where the higher the value (closer to 1), the higher the agreement.

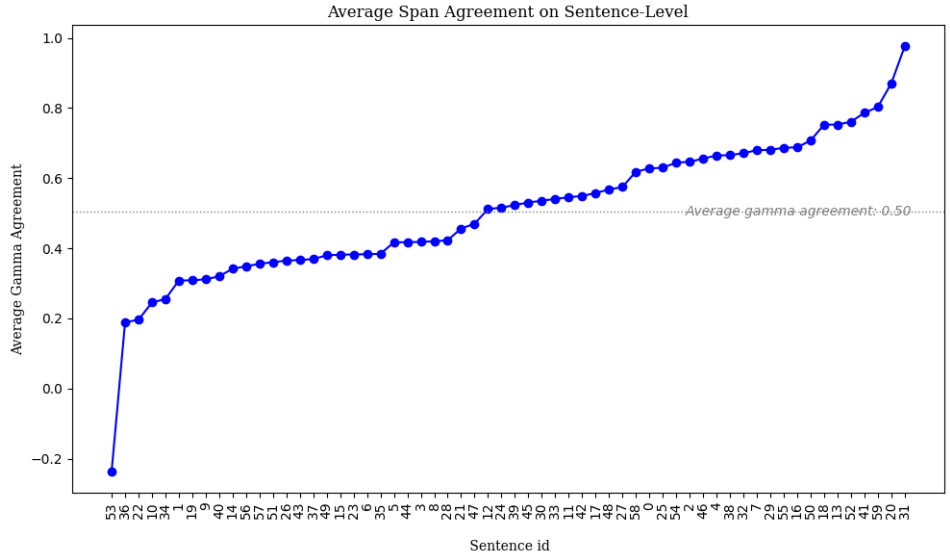

Figure 6: Average agreement (gamma) for trigger words on a sentence-level for the annotators that agreed mostly with MT.

Focussing again on the nine annotators that agreed most with MT, using the gamma python module[8], we compute a gamma agreement averaged over all 60 sentences of $\gamma = 0.504$. The trigger word agreements are visualised per sentence in Figure 6. This metric shows a broad range of agreements and disagreements for different sentence contexts (based on how neutral or stereotypical the sentence context might be). A factor to take into account is that gamma agreement can only be calculated between annotators that have at least one annotation. Considering our methodology was bottom-up and annotators were free to choose whether they annotated trigger words or not, not all nine annotators have at least one annotation per sentence. Therefore, the gamma agreements per sentence presented in Figure 6 have been calculated between those annotators that *did* annotate at least one trigger word.

Examples of where the nine annotators highly *disagreed* on context words as gender triggers for the seed word (in bold) are:

- Another **user** commented, "You guys are the best social media account in the country."

In the sentence above, annotators annotated trigger words such as *You guys*, *the best in the country* or *social media* either as individual words or partial phrases using mixed male or female

8. `https://pygamma-agreement.readthedocs.io/en/latest/index.html`

	agreement (γ)
gamma	.504
gamma α=1, β=2	.488
gamma α=2, β=1	.538

Table 1: Annotator gamma agreement using the general gamma, and gamma with varied positional (α) and categorial (β) weights.

gender labels. In contrast, examples of where the nine annotators highly *agreed* on context words as gender triggers (in italics) for the seed word (in bold) are:

- There will be a roughly 90-minute inspection after the *race* and the **winner** will not be deemed official until the process is completed.

In this sentence, all annotators labelled the word *race* as a male trigger.

Using the py-gamma documentation, it is possible to calculate slight variations of gamma or change the weights given to the position (segment annotated) or category (label chosen for annotation). In this study, we are firstly interested in whether a word has been annotated as a trigger or not (position), and secondly in how this trigger has been labelled (category). Table 1 presents various computations of gamma agreement for this set of data. First, the 'general' gamma with equal weights was calculated, which results in an agreement of $\gamma = 0.504$. The last two gamma computations focus on adapting the weights of positional and categorial mismatches. For the second calculation, a higher weight was assigned to categorial mismatches (β=2) and a lower weight was assigned to positional mismatches (α=1). This yields a lower gamma agreement of $\gamma = 0.488$. In comparison, the third calculation was computed by assigning a lower weight to categorial mismatches (β=1) and a higher weight to positional mismatches (α=2). This yields a higher gamma agreement of $\gamma = 0.538$. These two values show that annotators agreed more on *which* words they annotated as triggers (position) rather than *how* they annotated these words (category).

Furthermore, a limitation to gamma is that with an increase in the number of annotators, computation increases rapidly[9]. For a total of 21 annotators, computation increased unrealistically. A parallelised version (parallelisation script compiled and available here: `https://github.com/TomMoeras/parallel-pygamma`) of gamma was run multiple times to compute an agreement between all 21 annotators, but after running for 72 hours even on the VSC supercomputer (`https://www.vscentrum.be/home`) and then automatically stopping, we ended this intent.

4.3.2 Linguistic Analysis

Following an agreement measure of whether annotators agreed on the words they annotated in context as being triggers for their gender perception of that sentence, the next interesting aspect to look at is which types of words these annotators actually annotated. The aim is to understand both what gender is triggered in a certain context but also what specific context actually triggered that gender perception (or maybe also translation). Two computationally linguistic analyses were conducted to provide a deeper understanding of what these trigger words are and how they stand in relation to the seed word. First, a parts of speech (POS) analysis was conducted to analyse what these trigger words are linguistically. Secondly, to understand in what relation these triggers stand to the person in question, the dependency between annotated trigger words and the seed word was measured. Both analyses presented here are based on the nine annotators that agreed most with MT.

9. The pygamma documentation (`https://pygamma-agreement.readthedocs.io/en/latest/performance.html`) outlines: "We're aware that for a high amount of annotators, the computation takes a lot of time and cannot be viable for realistic input. The theoretical complexity cannot be reduced."

4.3.3 POS Analysis

The POS labels were computed using spaCy's[10] POS labeling, which is based on the Universal Dependencies annotation scheme[11]. Depicted in Figure 7 is the POS analysis of the words annotated as gender triggers by the 'top 9' annotators. As an example, of all proper nouns (PROPN) present in the 60 sentences, the annotators annotated an average of 24.80% of these. From Figure 7 we can see that, by percentage frequency, the most annotated words were proper nouns (PROPN) with a frequency of 24.80%, followed by adjectives (ADJ) with a frequency of 20.39% and nouns (NOUN) with a frequency of 19.87%. Annotated with a frequency percentage of 14.12% are verbs (VERB), followed by pronouns (PRON) with a frequency of 11.00% and to a lesser extent by adverbs (ADV) with a frequency of 8.82%.

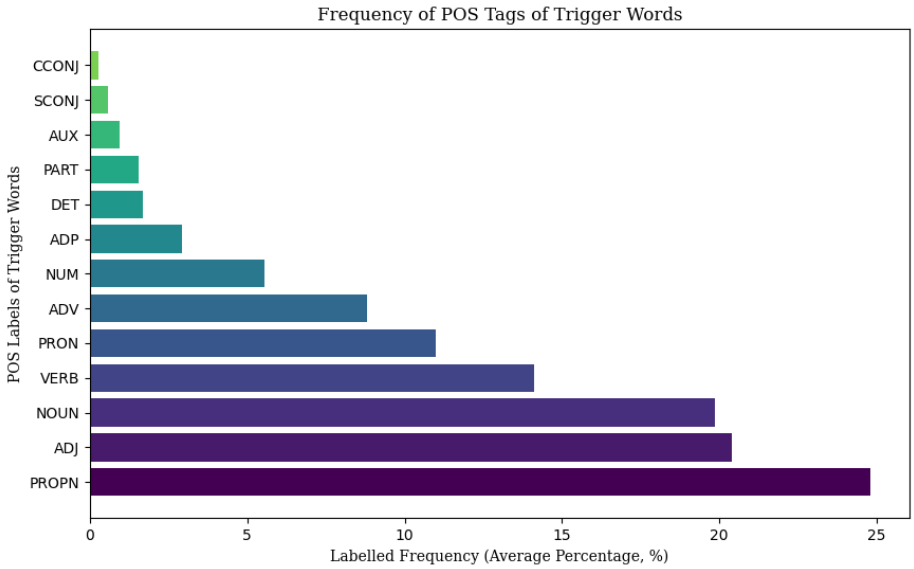

Figure 7: Average percentage of parts of speech of annotated triggers words for nine annotators that agreed most with MT. The y-axis represents the POS labels of these trigger words and the x-axis represents the frequency percentage of these words being annotated.

To clarify, these annotated trigger words do *not* include the seed word, and the context in the sentences was gender-ambiguous in relation to the seed word. For example, annotated pronouns were either not gendered in relation to the seed word (they may have been personal pronouns, such as *I* or *you*) or they were gendered pronouns referring to another person in the sentence, based on which a gender for the seed word may have been assumed.

Below are example sentences for each of the above-described categories, where triggers are marked in italics and seed words in bold:

- Proper noun as trigger: In other lyrics, the vocal *Donald Trump* **supporter** says: "Inflation's up like the minimum wage. So it's all the same. Not a damn thing changed."
- Noun as trigger: And it is far from a sure political **winner** in the upcoming *election*.
- Adjective as trigger: One day, she visited a friend who worked as an *assistant* production **coordinator** on a set, and she was intrigued by the location department.

10. `https://spacy.io/`
11. `https://universaldependencies.org/`

- Verbs as triggers: Like me, Imogen gets her "dream job" and thinks her life is finally starting - but her confidence and happiness is constantly *threatened* and *undermined* by a toxic **colleague**.
- Pronoun as trigger: After a **friend** suggested *she* try it, Ann said, "Sure!"
- Adverb as trigger: Kensington massage **therapist** jailed for *sexually* assaulting clients.

In machine translation, it has been shown that adjectives directly referring to a person can affect the gender of that person in question, also tested to counter bias and termed 'fighting bias with bias'. For example, a male-biased word, e.g., *lawyer*, is translated as female if a female-inflected adjective is added in front (e.g., *the pretty lawyer*, (Stanovsky et al. 2019)). The POS analysis presented in Section 4.3.3 shows a diverse representation of annotated words that affect human perception of gender.

4.3.4 Dependency Distances Analysis

Previous studies on adjectives highlight that words affecting the translation of a person's gender are close to the target word in terms of dependency structure. We want to have a better idea if there is a pattern/relationship between the dependency distance of a given word and how likely it could be a trigger word. Therefore, the second aspect linguistically analysed is the dependency relation between annotated trigger words and the seed word. This section analyses how these different words stand in relation to the seed word (to the person in question). The dependency distances were computed using spaCy's dependency tree calculation.

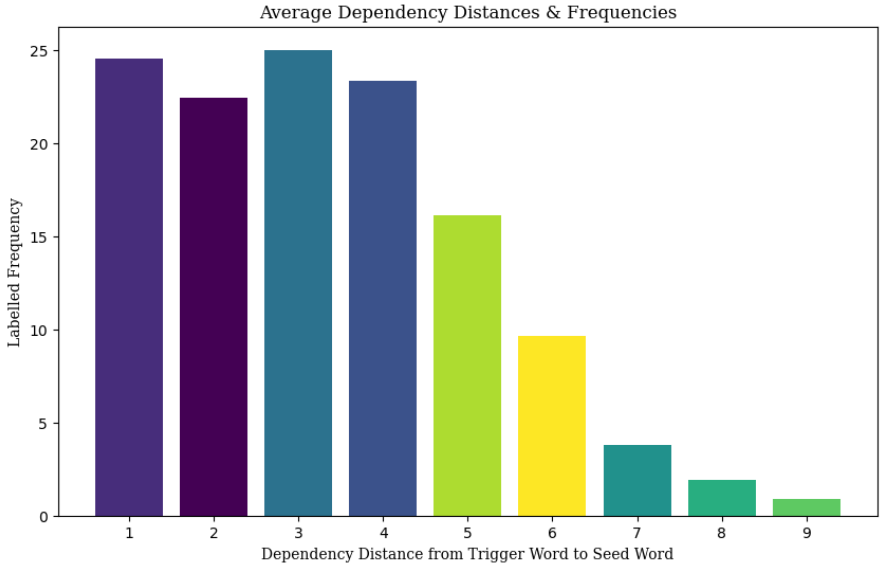

Figure 8: Average frequency of dependency distances from annotated trigger words to the seed word for nine annotators that agreed most with MT. Presented on the x-axis is the dependency distance measured from the trigger word to the seed word and the y-axis shows the frequency of this dependency distance.

Figure 8 visualises the average dependency distances, the number of edges in the dependency tree of a given sentence that separate two words, measured between an annotated trigger word and the seed word for the nine annotators that agreed most with MT. This figure shows that the

dependency distance of trigger words to the seed word largely ranged between 1 and 5, but even includes a few triggers measured at a distance of up to 9. Interestingly, trigger words were not just at a dependency distance of 1 away from the seed word, but equally frequently at a distance of 2, 3 and 4 away from the seed word. The dependency distance of 3, by a slight margin, even encompasses the most annotated trigger words. This analysis shows that contextual information even at larger distances can be trigger words but that words at a distance of up to 5 cover the majority of triggers marked for how humans perceive the gender of a seed word.

Below are example sentences for the highest four dependency distances, and for the dependency distance of 9. Annotated triggers are marked in italics, the seed words are marked in bold:

- Dependency distance 1: For someone who is normally a *business* **writer**, that is very valuable.
- Dependency distance 2: The German boss is reportedly convinced that the *teenager* can become a hugely important **player** in the near future.
- Dependency distance 3: The irony of this metaphor being solemnly valorised by the **boss** of a powerful *tech* corporation seems to be lost on the industry.
- Dependency distance 4: It appears the **filmmaker** is going to be making a much more elegiac and sobering *gangster* movie this time around.
- Dependency distance 9: The **socialite** who got a countess to write a 50-page manual on such things as how full a box of *tissues* had to be before it was thrown away - half - never noticed the fresh child factory production line of underage girls coming and going.

5. Discussion & Conclusion

The main contributions of this research are (1) a small sample of ambiguous natural data to analyse the effect of gender (bias), (2) an extended human gender analysis to include both non-binary and 'neutral' options, (3) a bottom-up analysis of what context triggers gender perception, and (4) a comparison of human gender perceptions and MT gender bias in (ambiguous) contexts, with a focus on translation. We conclude by forming hypotheses about sources of gender bias in MT source context, which will be further tested in future studies. These hypotheses have been based on an analysis of human gender perceptions based on implicit tests of bias and inter-annotator agreement and tested in context.

In Section 4, we explored how individual factors influence human gender perception (RQ1), we calculated to what extent humans agree on gender perception in the absence of clear (grammatical) cues (RQ2), as well as to what extent machine translations agree with human gender perceptions (RQ3). Based on the analyses presented on how and what triggers humans that agree with MT most to perceive gender in ambiguous sentences, can we form hypotheses about what context leads to gender bias in MT? (RQ4)

From our study, we can confirm that without contradicting (grammatical) gender cues present in the source text, MT primarily translates into generic masculine, as shown in previous research (Monti 2020, Vanmassenhove et al. 2018). However, and importantly, this occurs with the exceptions where the generic masculine is trumped by information in the source context. Table 2 in Appendix D shows a list of sentences, where seed words with a male-inflected or neutral word embedding were translated into a female gender in German. This shows that MT does take context (selectively) into account when translating gender. The same sentences were similarly annotated as 'female' by the annotators. Based on the analysis of human gender perceptions and a comparison of the nine annotators that had the highest agreement with MT, we arrive at the following hypotheses to be explored in future work (see Figure D for examples):

- Hypothesis 1: MT takes proper nouns and names into account when translating gender (e.g., "All too Well"; "Global Women Network"; "Ann").

- Hypothesis 2: MT is influenced by pronouns referring to other people in the sentence (e.g., stepping onto the court against a "she"; "her" friend).

- Hypothesis 3: MT translates ambiguous relations between partners (friendships or relationships) as male and female (e.g., "billionaire" and "companion"; "Ann" and "friend").

Summarising the work presented here, from a human analysis, we show that (i) human gender perceptions are only slightly influenced by their own gender, (ii) human gender perceptions are moderately influenced by their implicit gender bias, as primarily shown by the Gender-Science IAT results, (iii) across a diversity of annotators, human gender perception of seed words in ambiguous sentences leads to no noticeable inter-annotator agreement, and (iv) inter-annotator agreement is higher across the more biased annotators. These results show that to a certain extent human gender perception is influenced by stereotype information, as expressed by Gygax et al. (2008), specifically true for the top nine annotators, but that this varies greatly among all 21 annotators where agreement on gender perception is low. This shows that influences of stereotypes and resulting gender perceptions cannot be generalised across humans.

From a human-MT comparison, we show that (i) IAA values between humans and gendered machine translations show very varied results, and that (ii) human-MT inter-annotator agreements are correlated with human implicit gender bias: annotators with the strongest (male) implicit gender bias have the highest agreement with MT, and annotators with the lowest (female) implicit gender bias have the lowest agreement with MT. This once again underlines that MT predominantly outputs gender-biased and stereotypical translations, as shown in previous research. Where Lior and Stanovsky (2023) showed that "model biases reflect human decision-making", we show that this holds for a certain group of biased annotators (top nine), whereas, however, the MT behaviour does not reflect human perceptions across all 21 annotators, where agreement is low. Furthermore, where Zhu et al. (2024) show that annotators with *low* implicit gender bias tend to make similar choices as LMs, we, on the contrary, show that annotators with *high* (male) implicit gender bias tend to make similar choices as the MT system tested in this study. The obvious gender bias and annotation diversity among all annotators highlights the need for a more diverse and individual (or customisable) MT that can be adapted to the user's need.

From a linguistic perspective, we show that (i) annotated contextual information in ambiguous sentences reveals that humans are primarily triggered by proper nouns, nouns and adjectives in their perception of gender, and that (ii) humans are triggered by words in context that are at a dependency distance of up to 5, and even more, away from the seed word in question. These results from a human analysis show that specific type of contextual cues in the source (Kocmi et al. 2020) influence gender perceptions and 'why' this is the case (Don-Yehiya 2022). The results further show which type of words influence the gender perception, and which don't (Kim et al. 2019), and which relations between entities in the sentence (Lior and Stanovsky 2023) lead to human gender perceptions.

Overall, we highlight that (ambiguous) context has a demonstrable impact on human gender perceptions and how this is influenced by personal factors. We show that in the absence of (grammatical) gender cues in language, human perceptions of gender are extremely varied among themselves and are greatly influenced by context, as initially outlined in Hackenbuchner et al. (2024). Unsurprisingly, the MT system in question continues to primarily translate into the generic masculine, however, with the important exception of highly stereotypical scenarios or certain cues in the sentence context.

6. Limitations

This study includes a number of limitations. Among these is the fact that this study was focussed on DeepL as the sole MT system and in a single language direction. In the future, we aim to look at open-source MT models where we can extract more information from the model itself and look at different grammatical gender languages. Furthermore, while annotators were free to annotate using a 'male', 'female', 'neutral' or 'non-binary' label, the MT system merely translated the sentence into either 'male' or 'female'. Even though the MT system could technically translate into a 'neutral' form using language at its disposal (but not yet into a gender-inclusive form), all translations were of a binary gender. While this is an unfair comparison of the gender between human annotations and machine translations, it highlights the continued gender bias present in MT systems, compared to the much more diverse perception of humans. Furthermore, this study was merely conducted on a sample of 60 gender-ambiguous sentences. We are aware that this is, especially in current NLP research, a small dataset. The focus on this case study lay on a qualitative analysis and on evaluating correlations of annotators' implicit gender bias and their agreement with MT. For this qualitative analysis, a small data sample was chosen. In future research, we plan to release a big dataset of gender-ambiguous sentences, as compiled for this study.

7. Acknowledgements

This study is part of a strategic basic PhD research (1SH5V24N) fully funded by The Research Foundation – Flanders (FWO) for the time span of four years, from 01.11.2023 until 31.10.2027, and hosted within the Language and Translation Technology Team (LT3) at Ghent University. This research, including the information letter, study guidelines and informed consent form, has been ethically approved by the ethics committee at the Faculty of Arts and Philosophy at Ghent University. The authors would like to thank all annotators for their voluntary and patient annotations, without whom this study could not have been done. The computational resources (Stevin Supercomputer Infrastructure) and services partially used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. Thank you to Jasper Degraeuwe and Thomas Moerman for analytical and computational support.

References

Bolukbasi, T., K. W. Chang, J. Zou, V. Saligrama, and A. Kalai (2016), Man is to computer programmer as woman is to homemaker? debiasing word embeddings, *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Caliskan, A., P. Parth Ajay, T. Charlesworth, Wolfe R., and Banaji M. (2022), Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics., *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 156–170.

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20** (1), pp. 37–46, Sage Publications Sage CA: Thousand Oaks, CA.

Cramér, H. (1999), *Mathematical methods of statistics*, Vol. 26, Princeton university press.

Don-Yehiya, L.; Abend O., S.; Choshen (2022), Prequel: Quality estimation of machine translation outputs in advance, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing)*, Association for Computational Linguistics, p. 11170–11183.

Fleiss, Joseph L (1971), Measuring nominal scale agreement among many raters., *Psychological bulletin* **76** (5), pp. 378, American Psychological Association.

Friedman, B. and H. Nissenbaum (1996), Bias in computer systems, **14** (3), pp. 330–347.

Gygax, P., U. Gabriel, O. Sarrasin, J. Oakhill, and A. Garnham (2008), Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men, *LANGUAGE AND COGNITIVE PROCESSES*, Vol. 23:3, pp. 464–485.

Hackenbuchner, J., A. Tezcan, A. Maladry, and J. Daems (2024), You shall know a word's gender by the company it keeps, *Proceedings of the 2nd Workshop on Gender-Inclusive Translation Technologies*.

IBM Corporation, . (2024), Cramér's v.

Kim, Y., D. T. Tran, and H. Ney (2019), When and why is document-level context useful in neural machine translation?, *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT)*, Association for Computational Linguistics, p. 24–34.

Kocmi, T., T. Limisiewicz, and G. Stanovsky (2020), Gender coreference and bias evaluation at wmt 2020, *Proceedings of the 5th Conference on Machine Translation (WMT)*, Association for Computational Linguistics, p. 357–364.

Kotek, H., R. Dockum, and D. Sun (2023), Gender bias and stereotypes in large language models, *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24.

Lardelli, M. and D. Gromann (2023), Gender-fair (machine) translation, *New Trends in Translation Technology (NeTTT)*, Rhodes Island, Greece, p. 166–177.

Levy, S., K. Lazar, and G. Stanovsky (2021), Collecting a large-schale gender bias dataset for coreference resolution and machine translation, *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Domenican Republic, p. 2470–2480.

Lior, G. and G. Stanovsky (2023), Comparing humans and models on a similar scale: Towards cognitive gender bias evaluation in coreference resolution, *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, p. 755–762.

Mathet, A.; Métivier J. P., Y.; Widlöcher (2015), The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment, *Computational Linguistics* **41**, pp. 437–479.

McConnell-Ginet, S. (2013), Gender and its relation to sex: The myth of 'natural' gender., *In G. G. Corbett (Ed.), The expression of gender. DE GRUYTER.* p. 3–38.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021), A survey on bias and fairness in machine learning, *ACM Computing Surveys (CSUR)* **54**, pp. 1–35.

Misersky, J., P. Gygax, P. Canal, U. Gabriel, A. Garnham, F. Braun, T. Chiarini, K. Englund, A. Hanulikova, A. Öttl, J. Valdrova, L. Von Stockhausen, and S. Sczesny (2014), Norms on the gender perception of role nouns in czech, english, french, german, italian, norwegian, and slovak, *Behav Res* **46**, pp. 841–871.

Mitchell, T. M. (1980), The need for biases in learning generalizations, Rutgers University New Brunswick, NJ 08904.

Monti, J. (2020), *Gender issues in machine translation: An unsolved problem?*, The Routledge Handbook of Translation, Feminism and Gender.

Project Implicit (n.d.).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, Wei Li, and P. J. Liu (2019), Exploring the limits of transfer learning with a unified text-to-text transformer, *arXiv e-prints*.

Rudinger, R., J. Naradowsky, B. Leonard, and B. Van Durme (2018), Gender bias in coreference resolution, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2 (Short Papers), Association for Computational Linguistics, pp. 8–14.

Saunders, D. and B. Byrne (2020), Reducing gender bias in neural machine translation as a domain adaptation problem, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 7724–7736.

Saunders, D., R. Sallis, and B. Byrne (2021), Neural machine translation doesn't translate gender coreference right unless you make it, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Association for Computational Linguistics, p. 35–43.

Savoldi, B., A. Piergentili, D. Fucci, M. Negri, and L. Bentivogli (2024), A prompt response to the demand for automatic gender-neutral translation, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2: Short Papers, Association for Computational Linguistics, p. 256–267.

Shah, D., H. A. Schwartz, and Dirk Hovy (2020), Predictive biases in natural language processing models: A conceptual framework and overview., *Association for Computational Linguistics*, p. 5248–5284.

Stahlberg, D., F. Braun, L. Irmen, and S. Sczesny (2007), Representation of sexes in the language, *Social Communication, Frontiers of Social Psychology*, Psychology Press, New York, NY, p. 163–187.

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

Sun, T., A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, Kai-Wei Chang, and William Y. W. (2019), Mitigating Gender Bias in Natural Language Processing: Literature Review, *Proceedings of ACL*, Association for Computational Linguistics, Florence, IT, pp. 1630–1640. https://www.aclweb.org/anthology/P19-1159.

Troles, J. S. and U. Schmid (2021), Extending challenge sets to uncover gender bias in machine translation. impact of stereotypical verbs and adjectives, *Proceedings of the Sixth Conference on Machine Translation*, p. 531–541.

Vanmassenhove, E. (2024), Gender bias in machine translation and the era of large language., *arXiv e-prints*.

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

Wikimedia Foundation, . (n.d.), Wikimedia downloads. https://dumps.wikimedia.org.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang (2018), Gender bias in coreference resolution: Evaluation and debiasing methods, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2 (Short Papers), Association for Computational Linguistics, pp. 15–20.

Zhu, S., B. Du., J. Zhao, Y. Liu, and P. Liu (2024), Do plms and annotators share the same gender bias? definition, dataset, and framework of contextualized gender bias, *Proceedings of the Fifth Workshop on Gender Bias iN Natural Language Processing (GeBNLP).*, pp. 20–32.

Appendix A. ChatGPT - Prompt for Neutral Words

As explained in section 3.1, we prompted ChatGPT to provide us with gender-neutral seed words. This is due to the fact that previous research in this field provides lists of terms that are female- or male-inflected, but not a list of terms that is 'neutral' or non-gender inflected. We applied the following basic prompt: *Can you give me a list of 50 hyponyms of 'person' that are considered gender-neutral? (not feminine or masculine skewed like nurse or doctor).*

Appendix B. Gender: Seed Word Labels

As explained in Section 4.1.1, only a slight association was measured between the annotators' own gender and their annotations (choice of gender label for a given seed word).

Visualised in Figure 9 is an overall average of which gender labels were chosen to annotate the seed word, chosen from male, female, N/A or non-binary.

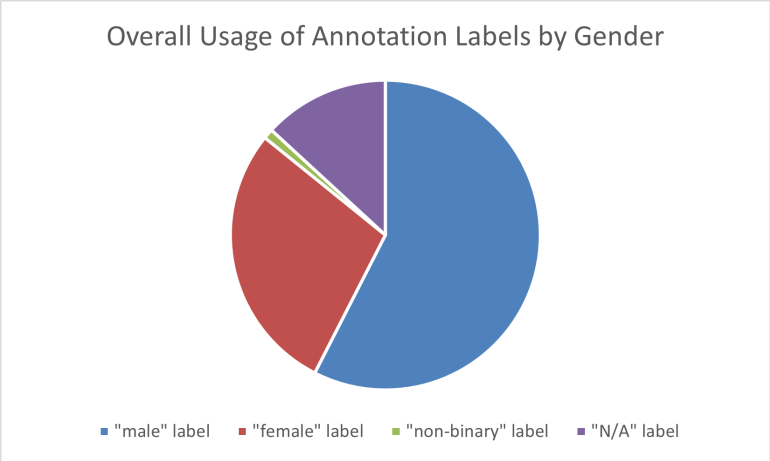

Figure 9: Overall usage of labels used to annotate the seed word, by gender: male, female, N/A or non-binary.

Visualised in Figure 10 are the average gender labels that a male, female or non-binary annotator chose.

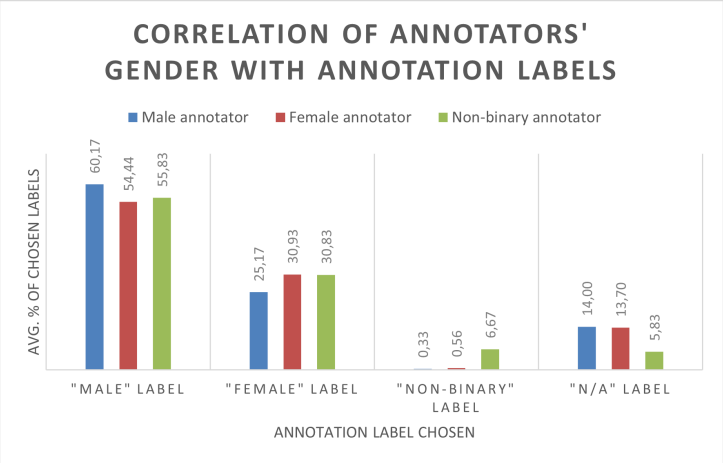

Figure 10: Correlation of what annotation labels male, female or non-binary annotators chose for the seed words in question.

Appendix C. Gamma Agreement: Continuum

As explained in Section 4.3.1, a gamma agreement is calculated both based on a dissimilarity and on a continuum. A continuum stores the set of annotations produced by several annotators for each annotated file, as explained in the gamma documentation, `https://pygamma-agreement.readthedocs.io/en/latest/principles.html`.

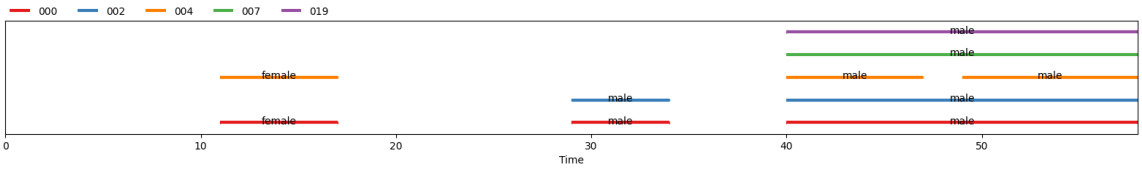

Figure 11: An example of a continuum, based on which span agreement gamma can be calculated. This example continuum shown here has been computed for five annotators that highly agreed for the sentence *Kensington massage therapist jailed for sexually assaulting clients*, where the trigger words annotated were *sexually assaulting*, *jailed* and *massage*.

Appendix D. Gender-inflections vs. Machine Translations and Annotations

As explained in Section 5, DeepL primarily translated the gender-ambiguous sentences from English into the generic masculine in German. However, with the exception of a few sentences, DeepL opted for a female translation. All seed words in Figure D have either a **male-inflected** word embedding or were considered to be **neutral**. All sentences shown were translated into a female gender by DeepL. This shows that MT does take cues in the sentence context (e.g., pronouns or names referring to other persons in the sentence) into account when translating. All sentences were equally annotated as 'female' by a majority of the top 9 annotators.

seed word (gender-inflection)	**gender-ambiguous source sentence**	**DeepL**	**Top 9 Ann.**	**All Ann.**
officer (male)	I am also the chief executive officer of Global Women Network, a United Kingdom-based Non-governmental Organisation with roots in Nigeria.	female	female 100%	female 100%
musician (male)	In an Instagram video posted last month, the "All Too Well" musician can be seen collaborating with producer Jack Antonoff on the piano.	female	female 66.6%	female 60%
opponent (male)	On Thursday evening, finally, she stepped out onto the court against a top 10 opponent for just the second time of her life.	female	female 100%	female 95.23%
companion (neutral)	In 2018, the billionaire said he couldn't be happy if he wasn't in love with a long-term companion.	female	female 100%	female 85%
friend (neutral)	After a friend suggested she try it, Ann said, "Sure!"	female	female 100%	female 85%
guard (male)	The reserve guard stepped up in the absence of fellow rookie guard Jordan Nixon, who injured her hamstring during warmups.	female	female 63.63%	female 62.5%

Table 2: English gender-ambiguous sentences that DeepL translated into German with a female-inflection or that the majority of annotators annotated as 'female', where the original seed word either had a 'male' or 'neutral' gender-inflection.