Social Sciences & Humanities Open 5 (2022) 100239

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points☆

Anna Farkas, Renáta Németh *

Eötvös Loránd University, Faculty of Social Sciences, 1117, Budapest, Pázmány Péter sétány 1/A, Hungary

ARTICLE INFO

Keywords:
Machine bias
Gender bias
Machine translation
Machine learning
Occupational segregation

ABSTRACT

In this paper—as a case study—we present a systematic study of gender bias in machine translation with Google Translate. We translated sentences containing names of occupations from Hungarian, a language with gender-neutral pronouns, into English. Our aim was to present a fair measure for bias by comparing the translations to a real-world oriented non-biased machine translator. When assessing bias, we used the following reference points: (1) the distribution of men and women among occupations in both the source and the target language countries, as well as (2) the results of a Hungarian survey that examined if certain jobs are generally perceived as feminine or masculine. We also studied how expanding sentences with adjectives referring to occupations affects the gender of the translated pronouns.

As a result, we found bias against both genders, but biased results against women are much more frequent. Translations are closer to our perception of occupations than to objective occupational statistics. Finally, occupations have a greater effect on translation than adjectives.

1. Introduction

In the recent years, there has been a growing interest in the research of machine bias, also referred to as algorithmic bias. The term "machine bias" describes the phenomenon that machine learning algorithms are prone to reinforce or amplify human biases (Prates et al., 2020). Machine learning algorithms are written by humans and draw conclusions from data that was created, collected, cleaned, and stored by humans. Thus, human error and bias can impact the algorithm and the results it generates (Ságvári, 2017). Nowadays, machine learning is used in a wide variety of sectors, including insurance, crime prevention, recruitment, healthcare, search engines, news outlets, online advertising, and recommendation systems among others (Burrell, 2016; Goodman & Flaxman, 2016; Ságvári, 2017; Sandvig et al., 2014). Since machine learning algorithms have a great deal of influence on many aspects of life, it raises concerns when they exhibit bias or discrimination. Over the past few years, researchers and journalists have discovered many cases when algorithms created biased results against certain social groups, thus, making socially unjust decisions in terms of race, gender, age, or religion—a few examples of these are: gender bias in hiring algorithms (Chen et al., 2018; Dastin, 2018; Schwarm, 2018), ageist and racist ad targeting (Angwin et al., 2017; Barocas & Selbst, 2016; Chen et al., 2018), and the lack of using regional dialects in the training corpus of Natural Language Processing (NLP) algorithms (Jurgens et al., 2017).

One of the most researched areas of machine bias is related to NLP models and bias in machine translation. Our study mainly focuses on the latter. Machine Translators (MTs) are trained on large corpora, and their translation results largely depend on the content of these texts. Bias is inherent in human texts, and it is not necessarily a source of unfairness. As Garrido-Muñoz et al. (Garrido-Muñoz et al., 2021) put it in their formal definition of bias, a language model behind the MT estimates the probability of a sequence of words. This allows, for example, to estimate the next most probable word in a sequence of words. The bias in this framework is some undesired variation of the conditional probability distribution of certain words according to certain prior words. Illustrated by our topic, those prior words could be "she" and "he," and we expect the conditional probability of words related to occupations after "she" to be very similar to that after "he." For instance, there is bias towards males if the conditional probability of "doctor" after "he" is higher than the conditional probability of "doctor" after "she." This bias can occur simply because there are more male doctors in the training texts.

☆ The paper has been posted on a pre-print server, see https://arxiv.org/abs/2011.06445.

* Corresponding author.
E-mail addresses: farkasanna98@gmail.com (A. Farkas), nemeth.renata@tatk.elte.hu (R. Németh).

https://doi.org/10.1016/j.ssaho.2021.100239
Received 16 November 2020; Received in revised form 19 June 2021; Accepted 5 December 2021
Available online 15 December 2021

It is now clear that bias is not a fault of the MT by itself, it is largely due to the data used to develop the system. However, as MT systems aim to aid or replace human translation, detection of their bias is highly relevant. Going beyond detection, explanation or correction would require, e.g., the examination of the input data, but these aims are outside the scope of our article. However, it is worth mentioning here that there have been attempts to debias language models, which were proved to be successful in reducing bias (Bolukbasi et al., 2016), but further improvement is needed (Gonen et al., 2019). Vanmassenhove et al. (Vanmassenhove et al., 2018a), Mirkin et al. (Mirkin et al., 2015), and Rabinovich et al. (Rabinovich et al., 2017) have concluded that personality-aware MTs can result in better translations.

Turning to our main topic, studies found that machine translation tools (e.g., Google Translate, Kakao translator, or Bing Microsoft Translator) can exhibit gender bias and have a tendency to provide male defaults (Cho et al., 2019; Prates et al., 2020; Rescigno et al., 2020). Gender bias in machine translation manifests in the use of gender-specific words when translating between languages that use gender-neutral and languages that use gender-specific forms of certain words (Johnson, 2020). For example, English has a pronominal gender system meaning it has masculine, feminine, and neuter forms of the third person singular pronoun, i.e., "he," "she," and "it" (Siemund, 2013). While gender-neutral languages, such as Hungarian, do not express gender in the third person singular pronoun. Translating pronouns between these languages, the machine translation tool—necessarily—provides either feminine or masculine translations to originally gender-neutral words.

In this paper, we present a case study of Google Translate[1], a widely used machine translation tool. We examine the bias of Google Translate by treating it as a black box. We focused on its translation results to detect and measure the bias of the system. We follow the results of previous studies about Google Translate regarding gender bias (Cho et al., 2019; Prates et al., 2020; Rescigno et al., 2020). These studies examined the translations of job positions, nouns, and adjectives. Prates et al. (Prates et al., 2020) analyzed the English translations of sentences written in a wide variety of gender-neutral languages, including Hungarian and Turkish. They translated sentences containing job positions and gender-neutral pronouns, like "ő egy orvos" ("he/she is a doctor" in Hungarian)—where "orvos" means "doctor"; and "ő" is a gender-neutral third person singular pronoun. In these sentences, the personal pronoun of the source language can be either translated into "he" or "she." Since, without any context, "he" and "she" are both correct translations of the third person singular pronoun "ő," Google Translate's algorithm must choose between them and provide one correct, gender-based answer for the query. It creates the possibility of gender stereotypes appearing in the algorithm. Fig. 1 shows examples of translating gender-neutral pronouns into gender-based pronouns.

This paper presents a systematic study of job-related gender bias in machine translation. As former studies captured the complex notion of bias with a single indicator (Prates et al., 2020), we aimed at providing a more complete and sociologically more grounded quantification of the phenomenon. The case study aims to measure the extent of gender bias appearing in the Hungarian–English translation of sentences including the names of occupations. We examine the translation of sentences, like "ő egy orvos" ("he/she is a doctor") or "ő egy mérnök" ("he/she is an engineer"). Analyzing these translations, we measure the extent of gender bias by comparing Google Translate to a benchmark. As a benchmark, we define the "fairest" MT possible that simply reflects the real-life distribution of men and women in each occupation, and we call it "real-world oriented MT" (RWOMT). For example, if an occupation is conducted by more women than men, our RWOMT translates a sentence containing the occupation with a female pronoun (i.e., "she"), the same logic applies for occupations that are conducted by more men than women.

We introduce three subtypes of RWOMTs which differ in the "real-world" data they are based on. (1) The first is based on the percentage of male and female workers in each occupation in Hungary, (2) the second is based on the percentage of male and female workers in each occupation in the USA, and (3) the third is based on whether people find a particular occupation feminine or masculine. In the first two cases, real-world external data come from administrative data on social structure, while in the third case, we used self-reported survey data. RWOMT gives results that fit best to the given external data. It is important to note that RWOMT is a hypothetical translator, only used to measure bias, not to be actually implemented. It is also hypothetical because its fairness focuses on only one aspect (gender) and is defined within our given analytical framework (e.g., it translates sentence by sentence, without taking context into account).

Our paper also includes an investigation of sentences containing both occupations and adjectives that characterize them because adding an adjective to the occupations may change the pronoun in translation. We analyzed the English translation of Hungarian sentences, such as "ő egy jó orvos," "ő egy nagyon jó orvos," "ő egy rossz orvos," and "ő egy nagyon rossz orvos" ("he/she is a good doctor," "he/she is a very good doctor," "he/she is a bad doctor," and "he/she is a very bad doctor"). Rescigno et al. (Rescigno et al., 2020) conducted a similar study by analyzing EN-IT, EN-FR, EN-SP translations of adjectives and professions in a single sentence.

Before describing the case study on Google Translate, it is important to mention that for some gender-neutral languages, Google Translate has been serving both feminine and masculine translations since 2018 (Johnson, 2020; Kuczmarski, 2018). This option is called "gender-specific translation" and it has been extended to Hungarian while preparing this study (Johnson, 2020). Now it provides both feminine and masculine translations for a single sentence containing a Hungarian gender-neutral word. It translates "ő" into "he" and "she" as shown in Fig. 2.

However, the analysis of Google Translate's results remains relevant for several reasons. First, the gender-specific translation of gender-neutral words is only provided for single sentences. When users want to translate multiple sentences (see, Fig. 1) or they use Google Translate's function that translates whole documents, it still provides gender-based results. Secondly, gender-specific translations are only implemented in English translations. At the time, it is not available for, e.g., Hungarian–Spanish or Hungarian–French translations. English often serves as an intermediary language between other languages (Prates et al., 2020), suggesting that biases in Hungarian–English translations may affect Hungarian–Spanish and Hungarian–French translations. Therefore, examining Hungarian–English language pairs remains to be relevant. Thirdly, we provide a methodology that can be a fruitful approach to examine the extent of bias in other machine learning tools.

2. Previous studies

To our knowledge, Cho et al. (Cho et al., 2019) and Prates et al. (Prates et al., 2020) were the first ones to investigate gender bias in

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

[2] Source: Google Translate [Screenshot]:https://translate.google.hu/?hl=en&tab=wT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20egy%20orvos%0A%C5%91%20konferenci%C3%A1t%20szervez%0A%C5%91%20er%C5%91s%0A%C5%91%20egy%20b%C3%A9biszitter%0A%C5%91%20f%C5%91z%0A%C5%91%20sz%C3%A9p (accessed 7 August 2020) Google and Google Translate are trademarks of Google LLC and this paper is not endorsed by or affiliated with Google in any way.

[3] Source: Google Translate [Screenshot]:https://translate.google.hu/?hl=en&tab=wT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20egy%20orvos (accessed August 7, 2020).Google Translate is a trademark of Google LLC.

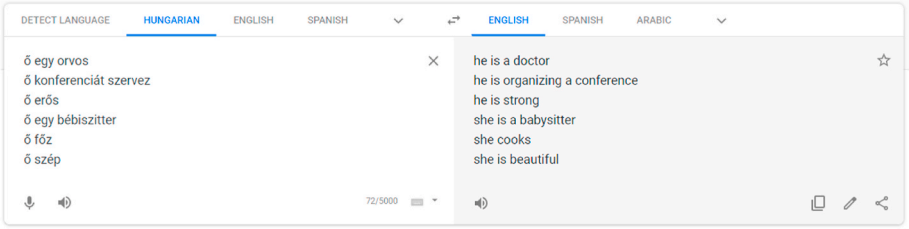

Fig. 1. Examples of translating gender-neutral pronouns into gender-based pronouns .[2].

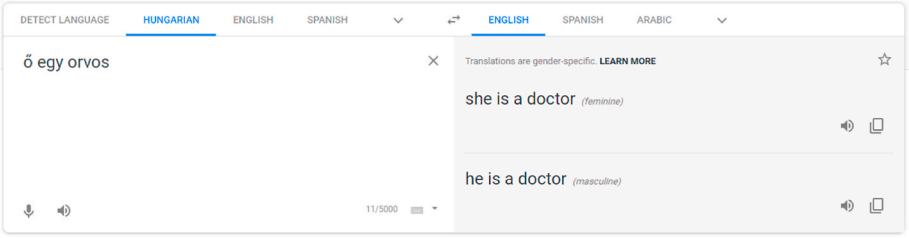

Fig. 2. Example for gender-specific translation.[3].

Google Translate from a social science perspective. Cho et al. (Cho et al., 2019) compared Google Translate's results to two other MTs (Kakao Translator and Naver Papago) measuring the difference between their performance. Prates et al. (Prates et al., 2020) compared the English translations of 12 gender-neutral languages to occupational statistics of the U.S., which allowed them to measure the gender bias of English translations in general. From a methodological point of view, Prates et al. captured the complex notion of bias with a single indicator (difference between translated results and current occupational statistics in the target language country). As compared to this study, we aimed to apply a broader approach to detect bias and present a quantitative approach generalizable to any machine bias in general.

3. Methods

3.1. Defining gender bias

To measure the extent of gender bias in Google Translate, first, we need to conceptualize gender bias in machine translation. Bias, in general, refers to the phenomenon when our observations systematically deviate from an existing or theoretical benchmark (Freedman et al., 2007). In this framework, Garrido-Muñoz et al. (Garrido-Muñoz et al., 2021) proposed a definition of bias in MT as some undesired variation of the conditional probability distribution of certain words according to certain prior words. Some researchers argue that the theoretical benchmark for machine translation is using female and male pronouns equally when the context is unknown (Cho et al., 2019). However, deviating from this 50:50 ratio does not necessarily mean the algorithm is biased. It only reflects the social world, as the distribution of male and female employees in most occupations deviates from 50:50 (Prates et al., 2020). Therefore, we define machine bias as an amplification of existing occupational gender segregation. We would emphasize here that, in our study, we examine sentence-level translations without using any contextual information.

As mentioned in the Introduction, we, as a benchmark, apply the real-life distribution of men and women in each occupation, and we call the hypothetical MT that reflects this distribution a "real-world oriented MT" (RWOMT). Prates et al. (Prates et al., 2020) followed a similar approach in their study of Google Translate. They compared the translations of occupations with the distribution of men and women in those occupations as registered in the occupational statistics of the U.S. Bureau of Labor Statistics. They compared the translations only to U.S. labor force statistics which describe the society of the target language (i.e., English), however, a comparison to the society of the source language could be justified as well.

To put our analysis into a more formal perspective, it is worth citing here Garrido-Muñoz et al. (Garrido-Muñoz et al., 2021) who emphasize that bias is the effect of the data from which the model was generated *and* the effect of the desired behavior of the model at a higher semantic level. It is this desirable behavior that we define differently when we move from the 50:50 ratio to the existing occupational gender segregation.

The gender bias exhibited by Google Translate is largely due to the corpora used to develop the system. Those texts may have been written originally in the target language or the source language. In the case of Hungarian–English translations, the original texts may have been written either in Hungarian or in English, containing gender stereotypes and biases of either the Hungarian or an English-speaking society. Google Translate was originally trained on texts published by the United Nations and the European Parliament (Prates et al., 2020), which indicates that probably most of the texts in its training corpus were written in English. This would justify the comparison to U.S. statistics. However, since 2014, it also relies on data obtained from users (Kelman, 2014; Prates et al., 2020), which makes it reasonable to compare Google Translate's results to the society of the source language as well. Therefore, in this study, we compare the translation of occupations to both Hungarian and U.S. statistics.

To compare the translations to the male-to-female ratio of each occupation in Hungary, we used the 2011 Hungarian census data (Hungarian Population Cens, 2011), which includes information about the distribution of men and women in 485 occupational categories. To compare the translations with the male-to-female ratio of each occupation in the U.S., we used data based on the Current Population Survey conducted in 2011—retrieved from the Bureau of Labor Statistics (BLS) (Bureau of Labor Statistics, 2011). Unfortunately, the latter data had some weaknesses from our point of view. First of all, the BLS did not report the proportion of men and women in occupational categories with less than 50,000 employees, thus, we had to eliminate some U.S. occupational categories from our analysis, which means that our comparison included only 239 U.S. occupational categories. Secondly, the BLS did not provide any data on military occupations, thus, we could only compare the translations of military occupations to Hungarian statistics. Lastly, the occupational categories of the U.S. statistics provided by the BLS have a different structure than the occupational

categories of the Hungarian census. The BLS uses the SOC (Standard Occupational Classification) system; while the Hungarian census uses the so-called FEOR (Foglalkozások Egységes Osztályozási Rendszere, Standard Classification System of Occupations) system. To compare the gender bias according to Hungarian statistics with the bias according to U.S. statistics, we had to pair the Hungarian FEOR occupational categories with the American SOC occupational categories. We used the International Standard Classification of Occupations (ISCO) system to create a crosswalk between the FEOR categories and the SOC categories (Bureau of Labor Statistics, 2015; Hungarian Central Statistical Office (n.d, 2020a).

The distribution of male and female employees is not the only factor determining the gender asymmetries of the translations of occupations. The gender stereotypes and bias present in the training data of Google Translate are also influenced by the way society thinks and writes about those occupations. Earlier, we defined a fair MT that follows the real-life male-to-female ratio of occupations. Following an alternative definition for "real-life," we also defined a fair MT that follows our society's views on occupations and gender based on whether people find an occupation feminine or masculine. One of the most cited studies about gender stereotypes of occupations is Shinar's study written in 1976 (Shinar, 1975), in which participants had to rate occupations as masculine, feminine, or neutral on a scale from 1 to 8. Although several studies have been published based on Shinar's research (Beggs & Doolittle, 1993; Couch & Sigler, 2001), to our knowledge, there has not been any representative survey conducted recently neither in Hungary nor in the U.S. that aimed to measure the gender stereotypes of occupations. Consequently, we were not able to compare the translations to any data or survey results indicating what occupations are considered feminine and masculine in the USA. On the other hand, to be able to conduct the same comparison of translations with Hungarian data, we designed a survey based on Shinar's study (Shinar, 1975). The questionnaire measured gender stereotypes of occupations on a Likert scale. Respondents had to answer the question: "On a scale from 1 to 6, where 1 means it is very masculine and 6 means it is very feminine, how much do you consider the occupations below to be masculine or feminine?" (The original question was written in Hungarian for the Hungarian participants). The survey was part of an online omnibus survey with a sample size of 1000 participants representing the Hungarian adult population. In total, one hundred occupations were included in the survey. To avoid bias coming from respondent fatigue, we randomly divided the occupations into five groups. 200 participants were asked to evaluate each group of occupations by labeling the occupations with a value from the masculine–feminine continuum (1–6). Thus, each participant had to label 20 occupations. The order of the 20 occupations was also randomized in order to reduce response bias.

When comparing the survey responses with the translations, we need a measure of masculinity/femininity for each occupation. As the original Likert-scale does not give such a score, we transformed it by keeping the distances between the original answer values unchanged. Responses 1 and 6 ("very masculine" and "very feminine") were transformed to 2.5, responses 2 and 5 ("masculine" and "feminine") were transformed to 1.5, and responses 3 and 4 ("somewhat masculine" and "somewhat feminine") got a value of 0.5. Masculinity score of an occupation emerges as a sum of values for responses (1–3) divided by the sum of values for all responses (1–6); while femininity score of an occupation emerges as a sum of values for responses (4–6) divided by the sum of values for all responses (1–6). An example for the calculation is shown in Tables 1 and 2: the masculinity score of the occupation "carpenter" is (425 + 18+3.5)/454 = 98%. Over the above, the survey included an additional question to assess the basis on which respondents decided whether they considered an occupation to be masculine or feminine.

3.2. Measuring gender bias

When evaluating the bias of a MT, we have to compare its functioning to the "reality" by using some real-world data. Following the above considerations, we compared results of Google Translate to three kinds of external real-world data: (1) the proportion of male and female workers of each occupation in Hungary, (2) the proportion of male and female workers of each occupation in the USA, and (3) the masculinity/femininity score of occupations. However, such a raw comparison is not fair: results of a MT cannot fit exactly to the external data. This is an issue former studies (Prates et al., 2020) failed to consider. To create a valid comparison, we defined a "fair" MT that fits as well to the external data as possible.

Table 1
Frequency of the answers given for the question regarding the occupation "carpenter" →"On a scale from 1 to 6, where 1 means it is very masculine and 6 means it is very feminine, how much do you consider the occupations below to be masculine or feminine?" (With weighting by gender, age, education, type of residence, and region.).

	1 – very masculine	2	3	4	5	6 – very feminine	Total
carpenter	170	12	7	3	4	0	196

In the following, we explain how to create a fair MT that best fits the male-to-female ratio of employees, and how to measure gender bias of a particular MT by comparing it to the fair one. Supposing that 60% of the employees of occupation "A" are women and 40% are men, a real-world oriented MT that strictly follows this ratio translates occupation "A" with "she" in 60% of the time and with "he" in 40% of the time. We call this MT a *probabilistic RWOMT,* because it translates a given occupation into "he" and "she" randomly with probabilities equaling its male-to-female ratio. Google Translate, on the other hand, is not a probabilistic MT but a deterministic one. A *deterministic RWOMT* translates occupation "A" with the pronoun "she," because it is conducted by more women than men. However, this translation does not represent 40% of the employees of occupation "A" who are men, so the translation gets 40 error points. The error of a deterministic RWOMT is called *optimal error* (E_o). "Optimal," because this is the smallest error achievable by a deterministic MT. A fair evaluation of the performance of Google Translate is based on a comparison to this optimal error. If Google Translate relates occupation "A" with "she," its error (E_t) will also be 40 error points, that is, it works in a fair way. Meanwhile, if Google Translate relates occupation "A" with the pronoun "he," its error will be 60 error points since it does not represent 60% of the employees, who are women. The 20 point difference can be benchmarked against the optimal error. In the example, the comparison gives 20/40 = 0.5, that is, the MT is 50% more biased than a deterministic RWOMT. Formally, the extent of gender bias (B) in Google Translate is:

$$B = \frac{E_t - E_o}{E_o}$$

If Google Translate translates a sentence with the adequate pronoun, its error will equal the error of a deterministic RWOMT and the extent of gender bias will be 0. Thus, if there are more female employees in a given occupation than male employees and the MT translates it with a female pronoun accordingly, we consider the translation to be unbiased. Accordingly, if the MT translates the sentence with an inadequate pronoun, we consider it to be biased. The score measuring the extent of gender bias can be any positive number. We have to highlight here that when we talk about "adequate" or "inadequate" translations we talk about it in relation to our hypothetical RWOMTs.

A similar approach and formula were used to compare Google Translate's results to a RWOMT that is based on society's view on gender and occupations. If we suppose that occupation "A" is seen as rather feminine by 60% of the participants of the survey, and Google Translate relates it with the inadequate pronoun "he," Google Translate's error will be 60 error points and it will be evaluated to be 50% more biased than a RWOMT based on society's opinion.

Some examples are shown in Table 3 displaying the score measuring

Table 2
Sum of values for answers given for the question regarding the occupation "carpenter" → "On a scale from 1 to 6, where 1 means it is very masculine and 6 means it is very feminine, how much do you consider the occupations below to be masculine or feminine?" & the calculated masculinity/femininity scores.

	1 – very masculine	2	3	4	5	6 – very feminine	Total	masculinity score (% of total)	femininity score (% of total)
Carpenter	425	18	3.5	1.5	6	0	454	98%	2%

Table 3
The calculation of the gender bias score of occupational categories.

occupation	FEOR category	pronoun	female employees (%)	male employees (%)	gender bias score of the occupation	gender bias score of the category
Statistician	Statisticians	he	73	27	1.7	1.7
Dancer	Dancers and Choreographers	she	58	42	0	0.2
Choreographer	Dancers and Choreographers	he	58	42	0.4	

the extent of gender bias when defined in relation to FEOR occupational statistics. As can be seen, for example, the strongly female-dominated "statistician" is translated with the pronoun "he," and its gender bias score is 1.7. There are some occupational categories, like "Dancers and Choreographers'' which contain multiple occupations. Since we only had data about the male-to-female ratio of the occupational categories—and not individual occupations—, the gender bias score of occupational categories containing multiple occupations was calculated by averaging the bias over all occupations within the category.

The scores measuring gender bias were calculated by comparing the translation of every occupation to occupational statistics (FEOR, SOC) and attitude survey results. Besides analyzing the extent of gender bias of occupations, we created larger occupational groups, so that we could examine the average gender bias arising in different occupational fields. We specified 18 larger groups of occupations with the help of the beforementioned FEOR, SOC, and ISCO occupational categorizations. We aimed to create groups that include similar occupations, however, are not too broad, as too broad grouping may cover up distortions in the case of certain groups. Number of occupations per groups is shown in Table 4. Since the structure of FEOR and SOC categories differ and the BLS did not provide any information about the male-to-female ratio of several SOC categories, the number of occupational categories a group involves may differ in the case of Hungarian and U.S. occupational categories.

Table 4
Number of occupations and occupational categories within employment sectors.

Employment sectors	number of occupations according to FEOR categories	number of occupations according to SOC categories
Managers	31	19
Science and Engineering	69	28
Social Science	13	6
Healthcare	28	23
Culture, Arts, and Sports	27	13
Education	14	8
Business and Finance	18	15
Legal Occupations	5	3
Office and Administrative Occupations	27	20
Building Industry	20	14
Crafts and Light Industry	23	6
Industry, other	18	7
Service	46	37
Sales	14	13
Agriculture	15	4
Machine Operators and Assemblers	38	15
Transportation	15	8
Military	3	0

By grouping occupations, we were able to measure the extent of gender bias in different employment sectors, such as Healthcare or Education. The extent of gender bias in a given employment sector equals the weighted average of gender bias in occupations within the group, with weights equaling the number of people in the given occupation. Since the ratio of female- and male-dominated occupations vary among these employment sectors, it would not be fair to compare the average bias score of the male-dominated and the female-dominated sectors. To resolve this problem, we analyzed the bias of the female- and male-dominated occupations of a sector separately. Analyzing the average gender bias of male- and female-dominated occupations in different employment sectors helped us to explore which sectors are affected by gender stereotypes the most.

3.3. *A step forward: when adjectives characterize the occupations*

Adding an adjective such as "good" to the noun "doctor" might make the male form more likely. Therefore, in a complementary study to our research, we examined sentences containing both occupations and adjectives. There have been previous studies that investigated the gender bias of sentences containing adjectives (Cho et al., 2019; Prates et al., 2020; Rescigno et al., 2020). Prates et al. (Prates et al., 2020) and Cho et al. (Cho et al., 2019) analyzed the translation of sentences, like "he/she is happy" or "he/she is sad"; Rescigno et al. (Rescigno et al., 2020) combined adjectives with names of professions, like "I am a famous pianist."

We chose two adjectives, "good" and "bad," as attributives and created sentences that followed this structure: "ő egy jó orvos," "ő egy nagyon jó orvos," "ő egy rossz orvos," "ő egy nagyon rossz orvos" ("he/she is a good doctor," "he/she is a very good doctor," "he/she is a bad doctor," "he/she is a very bad doctor"). As there is no existing benchmark for determining the gender of "good" and "bad doctors," we could not compare the results of these sentences to a fair MT that does not exhibit gender bias. Instead, we examined how adjectives change the gender of the pronoun used in the original sentences that contained only occupations. The comparison allowed us to measure the role of names of occupations and adjectives in determining the gender of the pronoun in the translation.

4. Data

To see how Google Translate translates gender-neutral pronouns from a source language to gender-based pronouns of the target language, we had to create a list of occupations and put them into sentences that followed the structure of "ő egy …" (meaning "he/she is a …"). To create

a list of occupations, we used the Hungarian Standard Classification of Occupations of the Hungarian Central Statistical Office (FEOR classification, (Hungarian Central Statistical Office (n.d, 2020b)). Using the FEOR classification for making a list of occupations proved to be effective for several reasons. First of all, the source language of our translations was Hungarian, so it was obvious to use a Hungarian classification. Secondly, using the FEOR system allowed us to compare the translations to the Hungarian census-based occupational statistics that are given also according to the FEOR system. Thus, we had data on the male-to-female ratio of every FEOR occupational category. Thirdly, as mentioned earlier, we created a crosswalk between the FEOR and the SOC systems, which allowed us also to assign the American male-to-female ratio to our list of occupations. There were some cases when the official SOC category associated with a FEOR category did not describe a given occupation accurately. In those cases, we changed the SOC category to a more appropriate one, using the occupation code finder developed by the U.S. Department of Labor (National Center for O*NET Development, 2020).

Some of the occupational categories of the FEOR system describe one job position, while others contain multiple occupations. For the latter, it was necessary to divide the category into several subcategories. For instance, the category "Adatbázis-tervező és -üzemeltető" (Database designer and operator) was divided into "adatbázis-tervező" (database designer) and "adatbázis-üzemeltető" (database operator), as shown in Table 5. We strived for selecting occupations that are well-known and, thus, more likely appearing in the training corpus of Google Translate. Therefore, occupations that correspond to special positions mentioned only in the FEOR system were not included in the final list of occupations. Additionally, some job positions in Hungarian define the gender of the person holding the position, e.g., "védőnő" refers to a female healthcare practitioner. Other occupations have different forms for male and female workers, like "színész" and "színésznő" (meaning "actor" and "actress"). Although "színész" can refer to both female and male actors, it is somewhat more common to use it for male actors. These occupations had to be excluded from the list, because they determine the gender of the pronoun in the translation. The same reasoning applies to religious occupations. Religious occupations determine (or at least are strongly related to) gender, and therefore, they are not included in our list. The final list includes 742 occupations. A few of them are shown in Table 5 together with their FEOR and SOC categories.

The 100 occupations—included in the survey that aimed to measure the femininity and masculinity of these occupations—were chosen from the above list of 742 occupations. We selected occupations that are included in the FEOR categories as separate categories, so that we could compare the male-to-female ratio of occupational categories with survey results. In addition, we tried to include male- and female-dominated occupations equally.

Before analyzing the translation of sentences, we had to create sentences from the occupations. These sentences followed the structure of "ő egy orvos" ("he/she is a doctor"). It is important to mention that the sentences chosen to be translated can have different structures. They can begin with a capital letter ("Ő egy orvos") or follow the structure of "ő orvos" (the indefinite article "egy" meaning "a(n)" can be dropped from the Hungarian version of the sentence). These minor changes have an effect on the translations as shown in Fig. 3. For our case study, we decided to use "ő egy orvos" because previous studies used the same structure (Prates et al., 2020). To create sentences with each occupation from our list, we used a code written in the programming language Python. These sentences were translated by using a special function of Google Translate with which it is possible to translate whole documents. At the time, the function translating documents provides only one gender-based pronoun for every sentence. It is important to mention that Google Translate's algorithm might have changed since we obtained the translations for this case study. All translations were retrieved in April 2020, hence, our results show the extent of gender bias of translations at that time.

All data collected and created for this case study can be accessed at: https://genderbiasdata.000webhostapp.com/

Table 5
Examples from the occupational list (translated into English) and their FEOR and SOC categories.

occupation	FEOR category	SOC category
[a] veterinarian	Állatorvos	Veterinarians
[b] database designer	Adatbázis-tervező és -üzemeltető	Database designers and administrators
[b] database operator	Adatbázis-tervező és -üzemeltető	Database designers and administrators
[c] hairdresser	Fodrász	Hairdressers, Hairstylists, and Cosmetologists
[c] barber	Fodrász	Barbers

[a] The FEOR category describes one occupation.
[b] The FEOR category contains multiple occupations.
[c] The FEOR category is associated with multiple SOC categories, thus, multiple occupations were assigned to the FEOR category.

5. Results

5.1. Bias defined in relation to the male-to-female ratio of occupations

First, we present results on Google Translate's bias defined in relation to the male-to-female ratio of occupations according to the Hungarian census data. Compared to Hungarian occupational statistics, 36% of the occupations were translated with an inadequate pronoun, of which 76% were translated with "he" instead of "she." It suggests that although there is a bias against both genders, biased results against women are much more common. If the occupations were supposed to be translated with "she," Google Translate translated the occupation with "he" in 67% of the time. Meanwhile, if the adequate pronoun was supposed to be "he," Google Translate used "she" in only 14% of the time. The probability of producing an inadequate translation for occupations that are female-dominated in Hungary is much greater than the probability of producing an inadequate translation for male-dominated occupations.

When the translations are compared to the U.S. occupational statistics, a similar pattern appears. Google Translate translated the U.S. SOC occupational categories with an inadequate gender in 44% of the time, which is slightly greater than the error ratio of the translations compared to the Hungarian FEOR occupational categories. When the translations were inadequate compared to U.S. statistical data, Google Translate should have used a female pronoun in 71% of the time and a male pronoun in 29% of the time. 73% of the female-dominated and 22% of the male-dominated occupations were translated inadequately. These results indicate a tendency similar to the bias measured against Hungarian occupational statistics.

We measured the extent of bias of Google Translate compared to a RWOMT, with a score ranging from 0 to any large positive number. Bias based on the Hungarian and the U.S. statistics showed a similar picture. Where we found gender bias compared to Hungarian statistics, the extent of bias ranged from 0.001 to 173.44 with a median of 1.01. Where we found gender bias compared to U.S. statistics, the bias score ranged from 0.01 to 88.91 with a median of 0.85.

Study results for employment sectors can be seen in Figs. 4 and 5. Since the employment sectors have varying numbers of male- and female-dominated occupations, and employment sectors containing a lot of female-dominated occupations have a greater chance to get a

[4] Source: Google Translate [Screenshot]:https://translate.google.hu/?hl=en&tab=wT#view=home&op=translate&sl=hu&tl=en&text=%C5%91%20orvos%0A%C5%90%20orvos%0A%C5%91%20egy%20orvos (accessed 7 August 2020).Google Translate is a trademark of Google LLC.

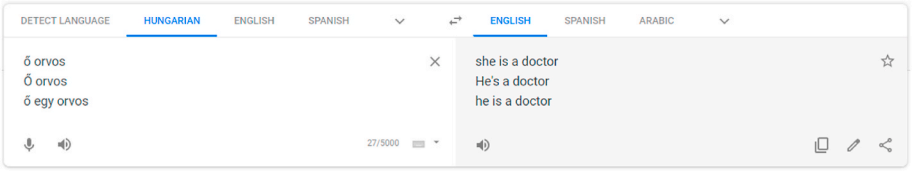

Fig. 3. Different options for sentence structure alter the pronoun used in translation.[4].

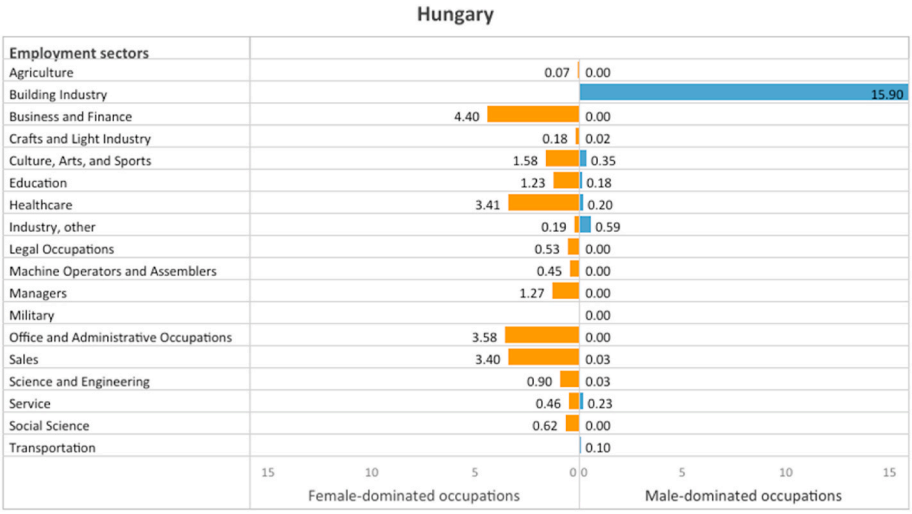

Fig. 4. Weighted average of gender bias in female- and male-dominated occupations in Hungary by employment sector.

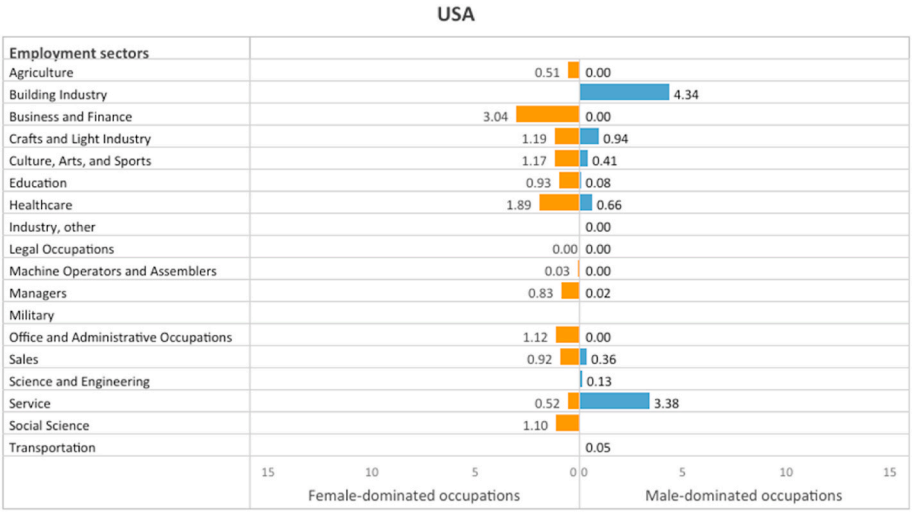

Fig. 5. Weighted average of gender bias in female- and male-dominated occupations in the USA by employment sector.

higher average bias score, we analyzed male- and female-dominated occupations per sector separately. We plotted the bias of the employment sectors averaged over the female- and male-dominated occupations within the groups (with weights equaling the number of women/men in the given occupation).

Fig. 4 shows the average bias of sectors defined in relation to Hungarian statistics. The Building Industry has the highest average bias score. There is a remarkable bias against men when translating the occupations of the Building Industry sector. In all other sectors there is a bias against women, or the bias against men and women are nearly equal. It is worth mentioning that none of the translations of Military occupations proved to be biased. Fig. 5 shows the average bias results defined in relation to U.S. statistics. Bias against men is noticeable in the Building Industry and Service sectors. However, in most sectors, the extent of gender bias against women is greater than or nearly equal to the gender bias against men. All in all, it can be concluded that the average gender bias, when defined in relation to either Hungarian or U. S. statistics, mainly derives from bias in female-dominated occupations with a few outlier sectors.

5.2. Bias based on the perceived femininity and masculinity of occupations

In this chapter, we present the results of measuring gender bias in Google Translate when defined in relation to society's opinion about the femininity and masculinity of occupations. Fig. 6 shows the femininity

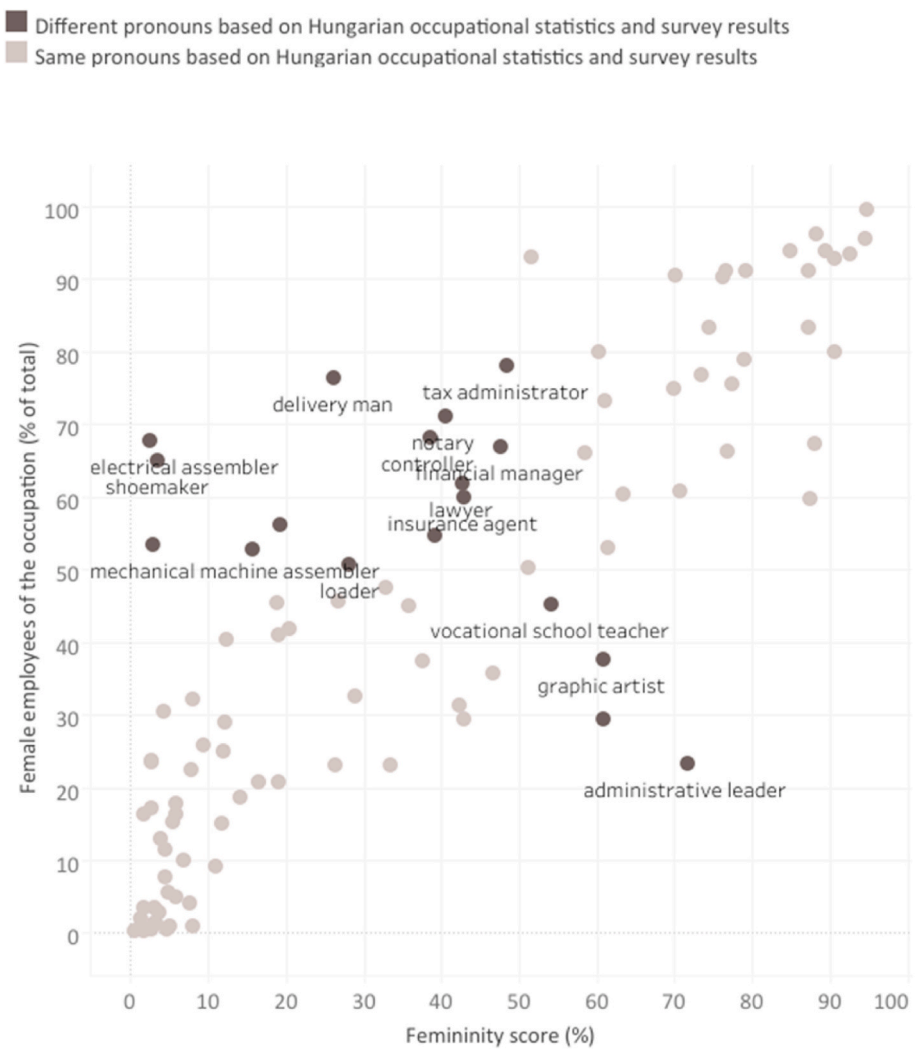

Fig. 6. Scatter plot of the perceived femininity of the occupations in the survey and the proportion of their female employees according to Hungarian census data.

score of occupations (based on the percent of survey respondents who consider a given occupation feminine). It also gives information about how consistent the femininity score is with the percent of female employees of the given occupation. Occupations above/below the 45-degree diagonal have a higher/lower proportion of women than what could be expected from survey results. There is a moderate strong positive association between the perception of occupations and their real-life male-to-female ratio. There are a few outliers, like "shoemaker" or "administrative leader," as the figure shows.

The discrepancy between the perception and real-life male-to-female ratio of those occupations can be assessed by using the additional survey question on the respondents' motivation. The question asked why the respondents decided whether they considered the occupations to be masculine or feminine, with four possible answers: (1) based on the proportion of men and women in the occupations, (2) based on personality traits related to the occupation, (3) based on physical abilities required for the occupation, and (4) based on other factors (an open-ended category). Only 20% of the respondents said that they based their answers on the proportion of men and women of occupations, while 30% of them made a decision based on personality traits related to the occupation, and 47% of them made a decision based on physical abilities required for the occupation. Some respondents mentioned that they took more than one aspect into account when they made a decision, and some had difficulty deciding, because they felt that "nowadays there are no longer jobs that are masculine or feminine" (quote from a respondent translated into English).

In the following, we analyze the extent of gender bias based on the survey. Google Translate results proved to be closer to the answers of the respondents than to the real-life male-to-female ratio of occupations. It translated only 20% of the occupations with the inadequate pronoun. In comparison, the same value is 30% when compared to Hungarian occupational statistics. Out of 20 cases, where the algorithm was inadequate compared to survey results, the sentences were supposed to be translated with a female pronoun in 17 cases and with a male pronoun in only 3 cases, as shown in Fig. 7. Fig. 7 also shows the gender bias score of the inadequately-translated 20 occupations.

If the sentences were supposed to be translated with a female pronoun according to the survey results, Google Translate translated the sentence with the inadequate pronoun half the time. If the adequate pronoun was supposed to be "he," it used "she" in only 4,5% of the time. Although the 100 occupations included in the survey are not a representative sample of the full list of occupations, the tendencies of gender bias based on survey results show a similar pattern to the tendencies of gender bias based on occupational statistics.

5.3. *The effect of adjectives in the sentences to Be translated*

In this chapter, we present our results of the complementary study of occupations and adjectives, where we examined sentences, like "he/she is a good doctor," "he/she is a very good doctor," "he/she is a bad

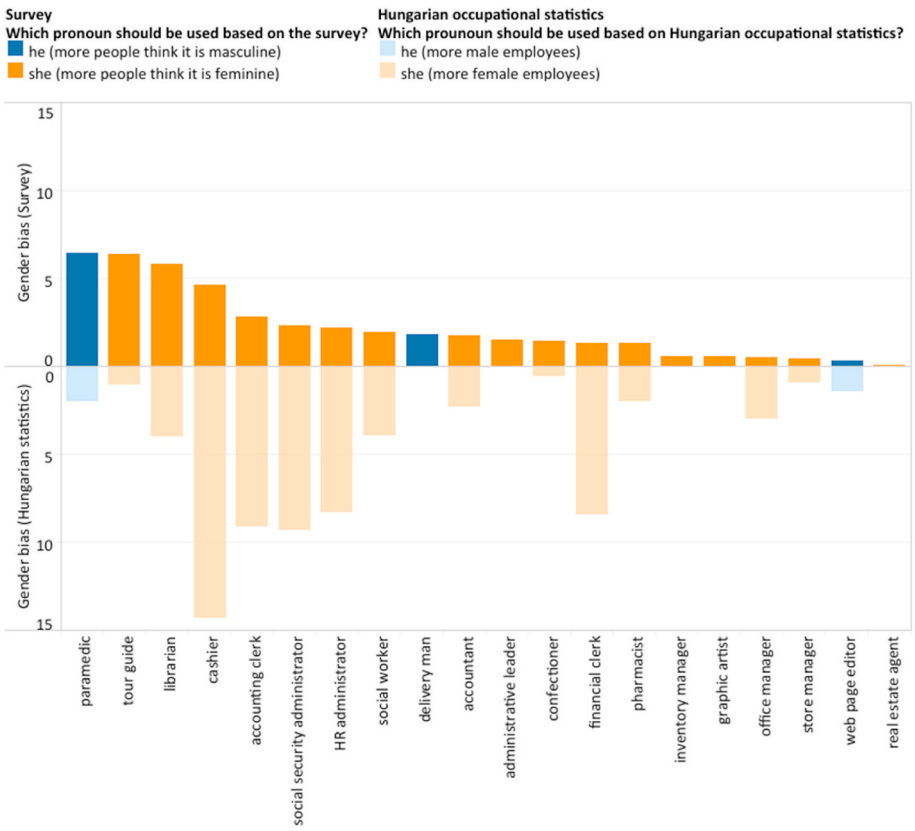

Fig. 7. Bar chart of the 20 inadequately-translated occupations according to survey results showing the extent of their gender bias defined in relation both to the survey data and to Hungarian statistics.

doctor," "he/she is a very bad doctor." Table 6 shows how the pronouns of the original sentences changed when adjectives were used. Although pronouns did not change in the majority of the sentences; when they did change, with one exception, they changed from "she" to "he." Nevertheless, adjectives altered the original pronoun of only 5–12% of the occupations, depending on the adjective used. This suggests that, in the examined sentences, occupations have a greater effect on the translations than adjectives.

Table 7 shows the proportion of male and female pronouns used in the sentences containing occupations and adjectives characterizing them. We found that 79% of the occupations were translated with a male pronoun and only 21% of them were translated with a female pronoun when only occupations were used. This tendency was even more asymmetrical for sentences where adjectives were used as well. When adjectives modified the original pronoun, the pronoun changed from "she" to "he" in almost every case as seen in Table 6.

6. Discussion

Building on previous studies that found gender bias in Google Translate and other machine translation tools, we presented an approach in this paper for a systematic study of gender bias in machine translation algorithms. To give a complete picture, we tried to cover the fullest possible range of occupations, and measured machine bias both for particular occupations and also for larger employment sectors.

With the aim to define a fair measure for bias, we compared the translations not to the actual gender composition of occupations (as e.g., Prates et al., (Prates et al., 2020)) but to a real-world oriented non-biased MT. For Hungarian–English translations of occupations, we established three benchmark criteria that a fair Hungarian–English MT can be based on. The first two RWOMTs aim to reflect the male-to-female ratio of occupations in Hungary and the USA, while the third RWOMT reflects the perceived femininity and masculinity of occupations. Comparing Google Translate's results to all three types of RWOMTs, we confirmed the findings of previous studies that discovered gender bias in the algorithm (Cho et al., 2019; Prates et al., 2020). Comparing translations either to the Hungarian or the U.S. occupational statistics, we find a remarkable bias against both genders, but biased results against women are much more frequent. Also, in larger employment sectors, gender bias derives mainly from bias against women. We examined the average bias score of 18 employment sectors, e.g., Healthcare, Business and Finance, and Education. We found that in

Table 6
The proportion of pronouns changed due to the use of adjectives.

Changes compared to the original sentences (original sentence → sentence with an adjective) (% of total)						
Adjective	she →she	he → he	she → he	he → she	did not change	changed
good	16%	79%	5%	0%	95%	5%
very good	14%	79%	7%	0.1%	93%	7%
bad	10%	79%	11%	0%	89%	11%
very bad	9%	79%	12%	0%	88%	12%

Table 7
The proportion of masculine and feminine translations.

Adjective	she	he
NONE (only occupation)	21%	79%
good	16%	84%
very good	14%	86%
bad	10%	90%
very bad	9%	91%

almost every sector, the average bias against women was bigger than the average bias against men—or they were nearly equal.

When comparing the translations to perceived masculinity/femininity of occupations instead of objective occupational statistics, we found the bias to be much lower, but still dominantly present against women. The result that Google Translate gives translations that are closer to our perception of occupations than to objective occupational statistics is plausible, as (1) they are perceptions and not actual social structures that are presented in the training corpus of Google Translate, and (2) changes in perceptions follow the actual structural changes with a delay.

We hypothesize that a more apparent bias against women can be explained by the lack of female data in the training corpus. Google Translate uses hundreds of millions of texts collected from the Internet in order to determine its translations (Kuczmarski, 2018), but researches show that online texts contain more male-related words than female-related ones (Bolukbasi et al., 2016; Schiebinger, 2014). Vanmassenhove and others (Vanmassenhove et al., 2018a, 2018b) have shown that the online available Europarl dataset[5]—used for machine translation purposes—has a male bias. They found that two-thirds of the sentences in the English–French subcorpora of the Europarl dataset are produced by men. It can be hypothesized that Google Translate uses male pronouns more frequently because male-related words are overrepresented in its training corpus. If the probability of using a male pronoun is higher, it will be likely to use a male pronoun even for occupations that are female-dominated. Naturally, machine translation algorithms are more complicated than the process described above, but the overrepresentation of male-related words in the dataset can cause a tendency where using male pronouns in translation is more common compared to using female pronouns.

Similarly to previous studies (Cho et al., 2019; Prates et al., 2020; Rescigno et al., 2020), our findings reinforce the assumption that male pronouns are more regular in Google Translate. The predominant use of male pronouns is more noticeable in sentences containing both occupations and adjectives (see, Table 7). This implies that a more complex sentence structure raises the probability of using a male pronoun—this theory, though, needs further proof and investigation.

The score created to measure the extent of gender bias proved to be efficient for quantifying gender bias in machine translation. Our methodology raises several possibilities for further research. Based on the formula of the gender bias score, follow-ups could be planned to investigate whether the structural changes in the job market cause changes in the translation of occupations. As another theoretical benchmark, the prestige scores associated with occupations could be also used to assess bias in translation. Finally, gender bias in other fields beyond occupations can be also quantified.

In this study, we discussed the issue of social bias found in Machine Translators. Bias is not a fault of the model by itself, it is largely due to the data used to develop the system, which contain some imbalances. However, as Machine Translators are widely used tools, we consider detection of their bias to be relevant in itself. Our paper described a methodology to measure gender bias in machine translation through a case study of Google Translate. We think that by investigating bias in machine learning and finding tools to measure its extent, this paper may provide relevant findings to the ongoing discussion of the societal consequences of machine learning algorithms and may contribute to the mitigation of the resulting social risks.

CRediT authorship contribution statement

Anna Farkas: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Formal analysis, Methodology, Software. **Renáta Németh:** Conceptualization, Methodology, Writing – review & editing, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the Inspira Group research company for generously incorporating our questions into their omnibus survey.

ELTE Eötvös Loránd University, Budapest, Hungary supported the proofreading of the paper.

References

Angwin, J., Scheiber, N., & Tobin, A. (2017). *Facebook job ads raise concerns about age discrimination.* The New York Times, 20 December https://www.nytimes.com/2017/12/20/business/facebook-job-ads.html. (Accessed 17 August 2020).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

Beggs, J. M., & Doolittle, D. C. (1993). Perceptions now and then of occupational sex typing: A replication of Shinar's 1975 study. *Journal of Applied Social Psychology, 23*(17), 1435–1453.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, & I. Guyon (Eds.), *NIPS'16: Proceedings of the 30th international conference on neural information processing systems* (pp. 4349–4357). NY, United States: Curran Associates Inc.

Bureau of Labor Statistics. (2011). *Labor force statistics from the current population survey*. https://www.bls.gov/cps/aa2011/cpsaat11.htm. (Accessed 11 August 2020).

Bureau of Labor Statistics. (2015). *Crosswalk between the 2008 international standard classification of occupations to the 2010 SOC*. https://www.bls.gov/soc/soccrosswalks.htm. (Accessed 11 August 2020).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Soc., 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14). New York, NY: Association for Computing. https://doi.org/10.1145/3173574.3174225.

Cho, W. I., Kim, J. W., Kim, S. M., & Kim, N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the first workshop on gender bias in Natural Language processing* (pp. 173–181). Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3824.

Couch, J. V., & Sigler, J. N. (2001). Gender perception of professional occupations. *Psychological Reports, 88*(3), 693–698. https://doi.org/10.2466/pr0.2001.88.3.693

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters, 10 October https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. (Accessed 11 August 2020).

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W. W. Norton & Company.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., & Ureña-López, L. A. (2021). A survey on bias in deep NLP. *Applied Sciences, 11*(7), 3184. https://doi.org/10.3390/app11073184

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 609–614). Minneapolis, Minnesota: Association for Computational Linguistics.

Goodman, B., & Flaxman, S. (2016). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Hungarian Central Statistical Office (n.d.), A hazai (FEOR-08) és a nemzetközi (ISCO-08) foglalkozási nómenklatúrák közötti fordítókulcs [Translation key between national (FEOR-08) and international (ISCO-08) occupational nomenclatures], https://www.ksh.hu/docs/osztalyozasok/feor/fordkulcs_feor_isco_hu.pdf (accessed 9 November 2020).

Hungarian Central Statistical Office (n.d.), Foglalkozások Egységes Osztályozási Rendszere [Standard Classification of Occupations] (FEOR–08), https://www.ksh.hu/docs/szolgaltatasok/hun/feor08/feorlista.html (accessed 9 November 2020).

Hungarian Population Census. (2011). *A foglalkoztatott népesség FEOR-08 kategóriák szerint [Employed population by FEOR-08 categories]*.

Johnson, M. (2020). A scalable approach to reducing gender bias in google translate. In *Google AI blog*. https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html. (Accessed 16 June 2021).

[5] The Europarl dataset contains corpus from the proceedings of the European Parliament (Koehn, 2005).

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In R. Barzilay, & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics* (Vol. 2, pp. 51–57). https://doi.org/10.18653/v1/P17-2009. Short Papers), Vancouver, Canada.

Kelman, S. (2014). *Translate community: Help us improve Google translate!* Google Blog. https://search.googleblog.com/2014/07/translate-community-help-us-improve.html. (Accessed 12 August 2020).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Conference proceedings: The tenth machine translation summit* (pp. 79–86). Phuket, Thailand: AAMT.

Kuczmarski, J. (2018). Reducing gender bias in google translate. In *Google blog.* https://www.blog.google/products/translate/reducing-gender-bias-google-translate/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+GoogleTranslateBlog+%28Translate+%7C+Google+Blog%29&utm_content=FeedBurner. (Accessed 12 August 2020).

Mirkin, S., Nowson, S., & Perez, J. (2015). Motivating personality-aware machine translation. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in Natural Language processing* (pp. 1102–1108). Lisbon, Portugal: Association for Computational Linguistics. C.

National Center for O*NET Development. (2020). *Business and financial operations occupations*. https://www.onetcodeconnector.org/find/family/title?s=13. (Accessed 12 August 2020).

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with google translate. *Neural Computing & Applications, 32* (10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Rabinovich, E., Nath Patel, R., Mirkin, S., Specia, L., & Wintner, S. (2017). Personalized machine translation: Preserving original author traits. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Vol. 1. Proceedings of the 15th conference of the European chapter of the association for computational linguistics* (pp. 1074–1084). Valencia, Spain: Long Papers, Association for Computational Linguistics.

Rescigno, A. A., Monti, J., Way, A., & Vanmassenhove, E. (2020). A case study of natural gender phenomena in translation: A comparison of google translate, bing Microsoft translator and DeepL for English to Italian, French and Spanish. In S. O'Brien, & M. Simard (Eds.), *Workshop on the impact of machine translation (iMpacT 2020), association for machine translation in the Americas, virtual* (pp. 62–90). https://www.aclweb.org/anthology/2020.amta-impact.4/. (Accessed 16 June 2021).

Ságvári, B. (2017). Diszkrimináció, átláthatóság és ellenőrizhetőség [Discrimination, transparency and verifiability]. *REPLIKA, 103*, 61–79.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). In *Auditing algorithms: Research methods for detecting discrimination on Internet platforms, paper presented at data and discrimination: Converting critical concerns into productive inquiry, a preconference of the 64th annual meeting of the international communication association. Seattle, WA*. https://www.semanticscholar.org/paper/Auditing-Algorithms-%3A-Research-Methods-for-on-Sandvig-Hamilton/b7227cbd34766655dea10d0437ab10df3a127396. (Accessed 6 November 2020).

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature, 507*(9). https://doi.org/10.1038/507009a

Schwarm, A. (2018). *Amazon, machine learning, and gender bias*. LinkedIn, 31 October https://www.linkedin.com/pulse/amazon-machine-learning-gender-bias-alex-schwarm/. (Accessed 12 August 2020).

Shinar, E. H. (1975). Sexual stereotypes of occupations. *Journal of Vocational Behavior, 7. 1*, 99–111.

Siemund, P. (2013). *Varieties of English: A typological approach*. Cambridge: Cambridge University Press.

Vanmassenhove, E., & Hardmeier, C. (2018). Europarl datasets with demographic speaker information. In J. A. Pérez-Ortiz, et al. (Eds.), *Proceedings of the 21st annual conference of the European association for machine translation* (p. 371). Alacant, Spain: Universitat d'Alacant.

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in Natural Language processing* (pp. 3003–3008). Brussels, Belgium: Association for Computational Linguistics.