Investigating Gender Bias in Machine Translation. A Case Study between English and Italian

Alessandra Luccioli, Ester Dolei, Chiara Xausa – Dipartimento di Interpretazione e Traduzione, Università di Bologna, Campus di Forlì

Citation: Luccioli, Alessandra, Ester Dolei, Chiara Xausa (2020) "Investigating Gender Bias in Machine Translation. A Case Study between English and Italian", in Adriano Ferraresi, Roberta Pederzoli, Sofia Cavalcanti, Randy Scansani (eds.) *Metodi e ambiti nella ricerca sulla traduzione, l'interpretazione e l'interculturalità – Research Methods and Themes in Translation, Interpreting and Intercultural Studies*, *MediAzioni* 29: B29-B49, http://www.mediazioni.sitlec.unibo.it, ISSN 1974-4382.

1. Introduction

Recent studies have shown that current machine translation (MT) systems are likely to adopt gender bias from humans (Escudé Font 2019; Kuczmarski and Johnson 2018; Prates *et al.* 2019; Zhao *et al.* 2018). Gender bias is defined as the prejudice against one gender based on the perception that women and men are not equal. Biases can be unintentionally transferred to machine translation systems, leading to a reinforcement of gender stereotypes, i.e. generalized views that refer to the practice of assigning to an individual woman or man characteristics, attributes and roles determined and limited by their gender. In this work, we will manually evaluate the translation of a sentence pattern previously employed by Escudé Font and Costa-jussà (2019) in the English-Italian language combination using two of the most popular MT systems, DeepL[1] and Google

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

Translate[2]. This sentence pattern translates into four sets of sentences, which include 40 occupations and three "stereotypical" adjectives. The aim of this study is to evaluate gender bias and to verify whether "stereotypical" adjectives can affect the final MT output[3]. Furthermore, we provide some relevant insights about gender bias in MT for post-editors and MT users[4].

The far-seeing memorandum by Warren Weaver (Locke and Booth 1955; Weaver 1955), questioning the possibility of using a computer to perform automatic translations, represents a milestone in the history of machine translation. From then on, significant developments have been witnessed in creating innovative architectures to build MT systems. The latest approach to MT, *neural machine translation* (NMT), was proposed by Kalchbrenner and Blunsom (2013), Sutskever *et al.* (2014) and Cho *et al.* (2014). NMT systems have shown positive results in the field of machine translation studies to date (Bahdanau *et al.* 2016; Zhao *et al.* 2018). However, despite the progress made in recent years, there are still issues with the output translations. One major problem to be addressed concerns gender bias.

Gender is expressed differently in different languages. Some languages feature masculine, feminine or neutral forms, while others are gender neutral. This diversity in languages represents a challenge for machine translation: when translating from a gender-neutral language into a language which encodes explicit information for this category at the morphological level, translation systems must "guess" or recover missing morphological information, and more

[2] https://translate.google.it/?hl=it

[3] We selected "stereotypical" adjectives through a corpus-based study and the analysis of the most relevant collocates of the lemmas man and woman. Therefore, we refer to adjectives that collocate more frequently with the lemmas man and woman, as stereotypical. The process will be explained in detail in Section 5.

[4] An Italian version of this study was published in the proceedings of the conference *R-esistenze in Movimento: Soggettività, Azioni, Prospettive*, in 2020. The Italian version differs from the English one because the former includes an additional dataset, to provide more evidence on the problem of gender bias in MT systems.

than one correct translation may exist for the same source input (Kuczmarski and Johnson 2018; Moryossef *et al.* 2019).

According to recent studies, different popular MT systems are prone to gender biased translations (Stanovsky *et al.* 2019) and this is explained by the functioning of current NMT systems. Such systems involve a single, large neural network that is trained to maximize the probability of providing a correct translation given a source text (Bahdanau *et al.* 2016). The architecture includes two functions: the first one encodes variable-length translation units and turns them into numeric vectors, which represent concepts (*encoder*); the second one decodes vectors and provides the target sentence (*decoder*). In order to improve performance, NMT employs deep learning techniques[5], namely algorithms that learn features from data (Sutskever *et al*. 2014; Bahdanau *et al.* 2016; Vaswani *et al.* 2017). Yet, a negative aspect of models trained in this way is that stereotypes and biases are learned from such data (Madaan *et al*. 2018) and have a direct impact on them, protracting or even amplifying linguistic bias[6] and social stereotypes (Zhao *et al*. 2017; Zhao *et al.* 2018). This phenomenon, also known as *machine bias,* concerns gender or racial asymmetries in society as reflected in trained statistical models (Prates *et al.* 2019).

Specifically, NMT systems perpetuate gender asymmetries in the translation of professional titles and institutional roles: until recently, Google Translate would have skewed results in favour of the Italian masculine form *dottore* for *doctor*, and the feminine form *infermiera* for *nurse*. In 2018, Google announced that it had taken a step towards reducing gender bias in its MT application: when translating from English into French, Italian, Portuguese or Spanish, users now get to choose between feminine or masculine forms (Kuczmarski and Johnson 2018). However,

[5] Deep learning techniques are based on Deep Neural Networks (DDNs), which are powerful machine learning models that have offered encouraging results in different fields of study, especially when it comes to reducing the gap between human and computer performances in a number of tasks.

[6] The notion of linguistic bias is well defined by Beukeboom and Burgers (2017) as “a systematic asymmetry in word choice that reflects the social-category cognitions that are applied to the described group or individual(s)”, in this case men and women.

feminine and masculine versions are only provided when translating single words.

1.1. Female forms of occupations in Italian

In the Italian context, the notion of sexism inherent to language was first theorized by Alma Sabatini (1987a; 1987b), and later addressed by Lepschy (1989), Fioritto (1997), Robustelli (2000; 2012; 2013), Luraghi and Olita (2006), Giusti and Regazzoni (2009), Gheno (2019).

Italian is a gender-marked language: its nouns are either masculine or feminine, and there needs to be agreement between nouns and other correlated forms, such as adjectives and pronouns. Linguistic sexism, defined as the tendency of a particular language to omit women, takes two forms in the Italian language. The former is the use of masculine generics, linguistic designations for males that are also used in reference to people in general, e.g. the generic meaning of *uomo* (*man*). The second linguistic marker of sexism in the Italian language is the use of masculine forms for female professional titles and institutional roles. It is particularly frequent for high-profile job positions and leading roles: *ministro* (*minister*), *sindaco* (*mayor*), *avvocato* (*lawyer*), and so on. As noted by Cecilia Robustelli (2013), whilst the feminine form is usually accepted for nouns referring to female-dominated professions – such as *infermiera* (*nurse*) and *maestra* (*teacher*) – resistances to adopt a gender-fair language with reference to leading roles are still extremely common even among women, who fear that the feminine form would diminish their authority. Such firm oppositions to linguistic change, however, mask cultural resistances to accept gender equality.

Many scholars have directed their attention on the female forms of occupations in Italian. As a matter of fact, the literature has focused on the creation and use of the feminine forms of high-profile professions or roles, such as *ministra* (*minister*), *sindaca* (*mayor*) or *ingegnera* (*engineer*). Robustelli (2013: online) claims that “resistance to the use of the female grammatical gender for many professional titles or institutional roles held by women seems to be based on

linguistic reasons, but in reality, it has a cultural nature"; indeed as Robustelli (*ibidem*) and Gheno (2019) pointed out, many non-leadership roles have feminine forms and do not raise any objection. For this reason, we decided to focus both on leadership and non-leadership occupations, in order to show that even the latter group of professions deserve a more detailed study.

2. Related work

For some years now, the scientific community has been paying close attention to the problem of gender bias in MT systems. Font and Costa-jussà (2019) performed a case study on gender bias in machine translation, proposing word embedding[7] techniques to provide gender debiased translation systems. They defined a framework to detect and evaluate gender bias, namely a test set of sentences to be translated from English into Spanish. They built their test set using a sentence pattern that includes the word *friend* in different contexts and a list of occupations. In this study, sets of word embeddings were first trained with the GloVe algorithm and then debiased, using a post process method. Results show that, with debiased word embeddings, the accuracy when predicting gender improves.

Conversely, it was also demonstrated that word embeddings deriving from text corpora do reflect gender bias. Gonen and Goldberg (2019) claim that the current debiasing methods actually hide the bias without removing it, since a lot of gender-biased information is still reflected in the representation of gender-neutral words (i.e. words such as "math" or "delicate" have strong stereotypical gender associations related to neighbouring words). They conclude that, since biases are profound and systematic, existing bias removal techniques are insufficient and should not be trusted to provide gender-neutral modeling.

[7] Word embeddings are vector representations of words. As stated by Font and Costa-jussà (2019), these representations are used in NMT to equalize gender biases.

Stanovsky *et al.* (2019) designed a challenge approach (called WinoMT) to evaluate gender bias in MT using two co-reference gender bias datasets, namely the Winogender (Rudinger *et al.* 2018) and the Winobias (Zhao *et al.* 2018), which include English sentences with neutral gender participant roles. Stanovsky *et al.* (2019) analysed gender bias translations in eight target languages with grammatical gender, employing four popular MT systems and two state-of-the-art academic MT models. Furthermore, they created an additional dataset using the adjectives *handsome* and *pretty* to test whether this dataset "corrects" the profession bias. Their results suggest that the use of the adjective *pretty* together with the word *doctor,* for instance, modifies the final output, returning a female inflection. One might argue, however, that the adjective *pretty* employed in this study does not *correct* the bias, but in fact reinforces it, since the female inflection of the translated profession is obtained using a "stereotypically-loaded" adjective.

3. Methodology

In this section, the methods and the materials employed in the study are described. The sentence pattern, the occupations and the adjectives used will be analysed, as well as the reasons that led us to choose DeepL and Google Translate as machine translation (MT) systems for the experiment. All the experiments reported on were conducted in September 2019.

Many scholars, such as Zhao *et al.* (2018), Rudinger *et al.* (2018) and Stanovsky *et al.* (2019), among others, have created challenge sets to detect gender bias, analyse it, and propose debiasing techniques, that are mainly based on the "Winograd schema" (Levesque *et al.* 2012). A Winograd schema is a pair of sentences that differ in a single word and that contain an ambiguous pronoun whose referent is different in the two sentences and requires the use of common sense knowledge or world knowledge to disambiguate, such as:

A. The trophy doesn't fit in the brown suitcase because it's too large.

B. The trophy doesn't fit in the brown suitcase because it's too small.

Such ambiguities still represent an issue for MT systems' outputs. Therefore, Winograd schemas and other sentence pairs could be used as challenges for machine translation by including, for instance, pronouns which have to be correctly translated, according to gender, in the target language (Davis 2016).

The challenge sets created by Zhao *et al.* (WinoBias), Rudinger *et al.* (WinoGender) and Stanovsky *et al.* (WinoMT), composed respectively of 3,160, 720 and 3,888 sentences, have been employed for studies relying on automatic evaluation methods; as such, they are impractical for a case-study scenario such as the one adopted in this paper. Our approach is instead based on the manual evaluation of translations in the English-Italian language combination. For this reason, we decided to use the custom sentence pattern built by Escudé Font and Costa-jussà (2019) for the English-Spanish language combination: "I've known <him, her> for a long time, my friend works as a/an <occupation>"[8]. This custom sentence pattern allows us to evaluate whether the word "friend" and the occupations are translated correctly from English into Italian, i.e., according to the gender of the co-referent (her or him). The word *friend*, indeed, generates ambiguity gender-wise in translation, since it can be translated with the word *amico* (male friend) or *amica* (female friend) in Italian.

For this study we used the data of the Current Population Survey (2018) provided by the Bureau of Labor Statistics of the US Department of Labor, since they are widely available, up to date and broadly employed in related work. We selected a total of 40 male- and female-dominated occupations, considering an occupation as female-dominated when the percentage of women workers is higher than 50% of the total. We are aware that many professions are not strongly polarised and show close percentages of male and female workers. However, the distinction between professions with more than 50% of female workers is mainly for statistical purposes. We selected 20 professions with more than 50% of female workers and 20 professions with less than 50% female workers. The occupations

[8] The authors also used Spanish proper names to assess their impact in reducing ambiguity (Escudé Font and Costa-jussà 2019), but for the purpose of our study and due to space limitations, we did not include Italian proper names. This point could be considered for future research.

and the percentage of women workers in these occupations are shown in Table 1.

Occupation	**%**	**Occupation**	**%**
Secretary	94.0	Professor (Post-secondary teacher)	49.0
Hairdresser	92.1	Salesperson	48.7
Cleaner (housekeeping)	90.1	Photographer	47.8
Nurse	88.6	Scientist	43.9
Office clerk	84.5	Driver (bus)	43.8
Assistant	83.3	Cook	41.8
Therapist	82.1	Clerk	41.5
Social worker	81.6	Doctor	40.3
Librarian	78.5	Dentist	35.7
Psychologist	75.9	Web developer	32.5
Tailor	75.1	Director	29.2
Cashier	73.8	Farmer	25.8
Counselor	72.0	Security guard	22.4
Veterinarian	71.2	Laborer	21.4
Pharmacist	63.4	Programmer (computer)	21.2
Baker	61.1	Courier	21.1
Writer	59.6	Drafter	20.6
Teacher (secondary school)	58.0	Technician (engineering)	18.1
Bartender	57.2	Pilot	9.0
Editor	52.2	Painter (construction)	7.2

Tab. 1 occupations and percentage of women workers

The sentence set is formed of 80 sentences, 40 with a male referent and 40 with a female referent (see Appendix 1). The set was translated from English into Italian using two commercial neural machine translation systems, DeepL and Google Translate (see Appendix 2). The results of this set are discussed in section 4.1 below.

Stanovsky *et al*. (2019) suggest that adding adjectives usually associated with female or male entities can affect machine translation performance in some language pairs. Starting from this hypothesis, we argue that “stereotypical” adjectives affect the MT output. In order to find “stereotypical” adjectives, a corpus-based study was carried out, based on the study by Pearce (2008) on the collocational behaviour of the words “man” and “woman”. We decided to analyse

the modifiers of these lemmas, since collocational patterns can reveal the associations and connotations of words and, therefore, the assumptions they embody (Pearce 2008: 3). The corpus used is EnTenTen15, a 15-billion-word web-crawled corpus, created in 2015. The EnTenTen corpus (Jakubíček *et al.* 2013) was tagged by the TreeTagger using the Penn TreeBank tagset with Sketch Engine modifications. As a corpus query tool we used the SketchEngine (Kilgarriff *et al.* 2014), in particular its WordSketch Difference feature. Through this feature, we were able to compare the collocates of the lemmas *man* and *woman* and focus on their modifiers. The frequency of the most distinctive adjectives used for both men and women is shown in Table 2.

Modifier (adjective)	**Lemma MAN**	**Lemma WOMAN**
Wise	18873	2650
Strong	10642	11718
Beautiful	2624	26819

Tab. 2 Frequency of most distinctive modifiers of lemmas man/woman in EnTenTen15.

The modified sentence pattern is thus “I’ve known <him, her> for a long time, my <beautiful, strong, wise> friend works as a/an <occupation>”. The translations of this set of sentences are shown in Appendices 3, 4 and 5. The results of these sets are discussed in sections 4.2, 4.3 and 4.4 below.

The MT systems used for the study are two of the most popular MT systems available: there are currently 200 million daily users for Google Translate (available in more than 100 languages), and 312,000 daily users for DeepL (available, in 2019, in 8 languages). Since we aim at providing useful insights for post-editors, we considered relevant to employ MT systems that can be used by all post-editors, therefore easily accessible and user-friendly. All sentences were translated in September 2019. A final note of caution: as in all studies using web data and web-provided applications Google Translate and DeepL algorithms are likely to change, the full reproducibility of our results cannot be guaranteed.

4. Results

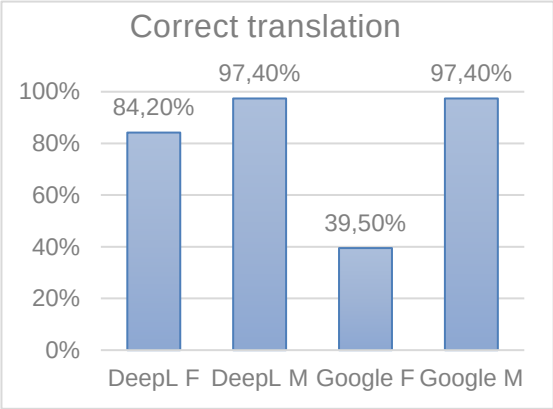
The final set includes 38 out of the 40 occupations originally selected: it was decided to exclude *cleaner* and *editor* from the study as the output translations were compromised by grammatical and semantic inaccuracies. Furthermore, both *cleaner* and *editor* have a number of different translations that makes comparison difficult. It is worth mentioning that DeepL and Google often produce *donna delle pulizie* (*cleaning lady*) instead of the more adequate translations *addetto/addetta alle pulizie* for *cleaner* with both *him* and *her* as co-referents.

The bar charts in Figure 1 show the percentage of occupations which have been correctly translated by the two commercial MT systems employed in this study, where "correctly" means that occupations have been properly translated, according to the male/female co-referent of the sentence. *F* refers to the feminine gender of the co-referent ("I've known *her* for a long time"), while *M* refers to the male gender of the co-referent ("I've known *him* for a long time"). In addition, the label *baseline* refers to the sentence pattern without adjectives, while the labels *beautiful*, *strong* and *wise* indicate each adjective added to the sentence structure. The labels *DeepL F* and *Google F* show the percentage of professions correctly translated according to the co-referent *her* by DeepL and Google Translate; while the labels *DeepL M* and *Google M* indicate the percentage of professions properly translated according to the co-referent *him* by DeepL and Google Translate.

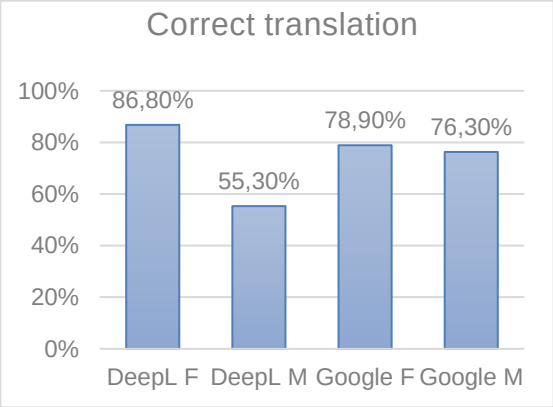
The bar charts in Figure 2 show the percentage of agreement[9] between the noun group (*my -/beautiful/wise/strong friend*) and the correct professions, translated using DeepL and Google Translate. Therefore, the labels *DeepL F* and *Google F* show the rate of agreement between the noun group and correctly translated professions, when the co-referent is *her*; while the labels *DeepL M* and *Google M* indicate the percentage of agreement between the noun group and the correctly translated professions, when the co-referent is *him*.

[9] The correct alignment of different elements of the speech (e.g. article, noun, adjective, pronoun, verb) in gender, number and person, when they are syntactically connected.

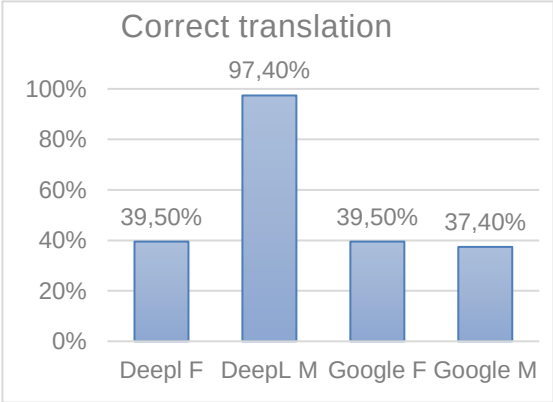
Set: baseline



Set: beautiful



Set: strong



Set: wise

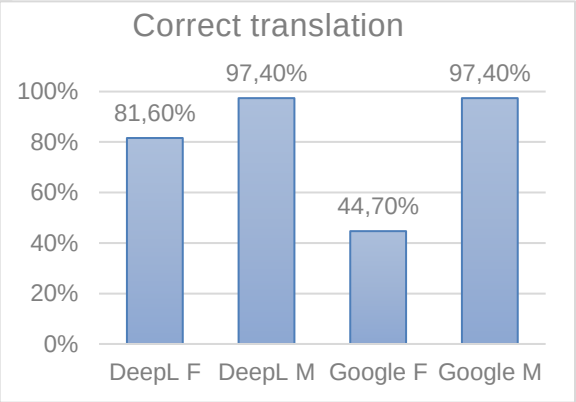
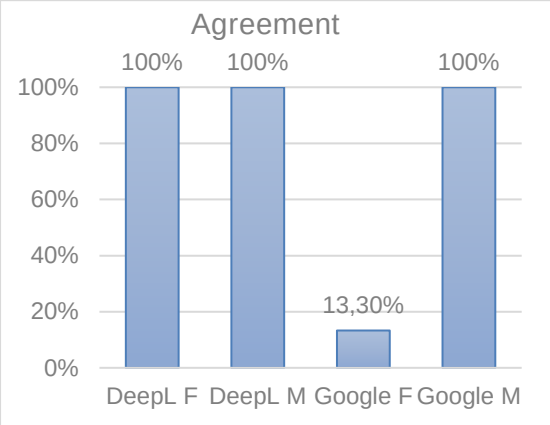
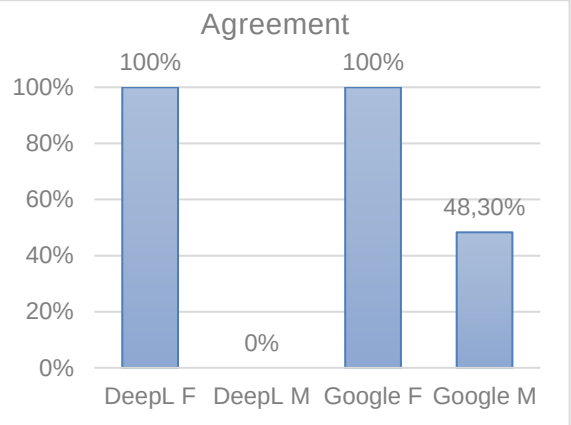


Fig. 1 Percentage of correctly translated professions.

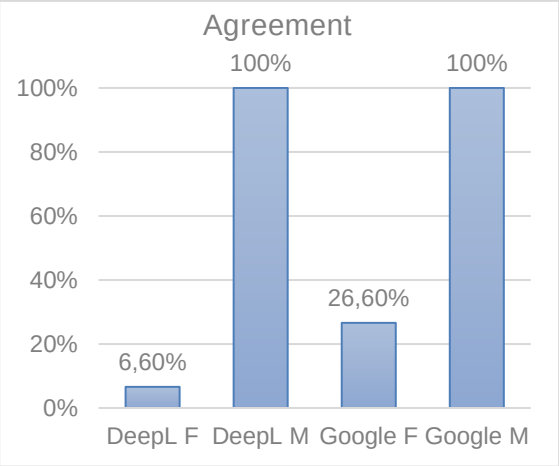
Set: baseline



Set: beautiful



Set: wise



Set: strong

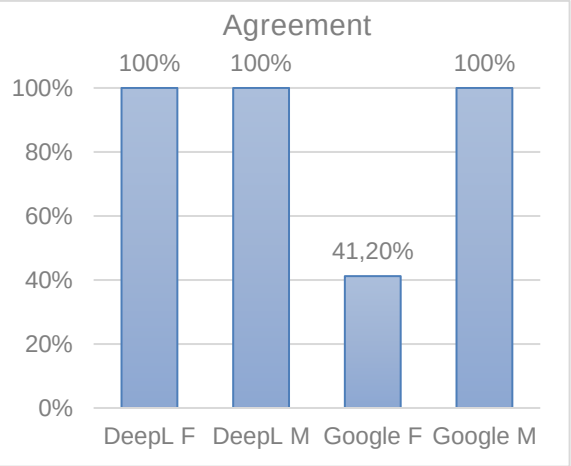


Fig. 2 Percentage of agreement between correctly translated professions and the nominal group.

4.1. Analysis of the *baseline* set

Our analysis indicates that DeepL and Google Translate have no problems in providing the correct translation both for the male and female pronoun. Both systems, however, achieve their best performance with male pronouns: if the original English sentence contains the male pronoun *him*, the systems predict the gender of the occupation with 97.4% accuracy. The only inaccuracy is related to a profession mainly held by women: *nurse* is biased towards a female translation in both systems. The gender of the nominal group *my friend* is predicted with

100% accuracy, and the gender agreement with the profession is always maintained.

However, when the coreferent is *her,* Google Translate tends towards the male default: the female gender of the occupation is only produced with 39.5% accuracy, and an even lower percentage can be observed for the nominal group *my friend* (13.3%). It is worth mentioning that 13 out of the 38 occupations selected are translated into Italian as *bigender* nouns, which means that they have the same suffix for masculine and feminine form (e.g. *assistente*, *terapista*, *regista*). For this reason, those translations were considered accurate[10]. In such cases, gender can be deduced by the article preceding the noun. Therefore, in case of mismatch agreement between the male translation of the nominal group and the translation of the profession, the occupation accuracy percentage might decrease.

Conversely, DeepL shows a relatively high accuracy with the female coreferent. The gender of the occupation is predicted with 84.2% accuracy, and the gender agreement with the nominal group is always correct. However, the quality of the translation is reduced due to stereotypical gender role assignments: the system is indeed biased towards a male translation for male-dominated professions (e.g. *doctor*, *scientist*, *technician*, *web developer*) even when the co-referent is *her*.

4.2. Analysis of the set with modifier *strong*

Similar results can be observed when adding the adjective *strong*, which collocates with a similar frequency with both *man* and *woman*. Both systems still achieve the highest performance with the coreferent *him*: occupations are translated with 97.4% accuracy by both systems, with *nurse* as the only

[10] It was decided to consider *corriere*, the male translation of *courier*, as a bigender noun. Although Italian provides a feminine form ending in *-iera* for masculine nouns ending in *-iere*, we could not find any attestation of the feminine form *corriera,* since the main meaning of *corriera* is *coach* (means of transportation). The Zingarelli 2016 Italian dictionary indicates that the form *corriera* with the meaning of female courier is seldom occurring or found.

inaccuracy; gender agreement can be observed with 100% accuracy. Regarding female roles, Google Translate improves its accuracy from 39.5% to 44.7% for the occupations, and from 13.3% to 41.17% for the nominal group, while DeepL accuracy is consistently high.

4.3. Analysis of the set with modifier *beautiful*

Some deviations can be observed when adding an adjective associated with female entities, such as *beautiful,* with respect to the baseline dataset. Google Translate improves its performance substantially when the coreferent is *her*, translating most professions (78.9%) as well as the nominal group (100%) using a feminine form, while DeepL keeps its high accuracy (86.8% and 100%). Interestingly, both systems keep providing male translations for male-dominated occupations (e.g. *doctor*, *web developer* etc.), regardless of the presence of the adjective *beautiful*.

Moreover, we observed that gender bias becomes more evident with a male co-referent in this set. Both systems' performance, indeed, worsen dramatically, predicting the male gender of the occupation only for careers dominated by men and inflecting all other professions towards a female translation. Compared to the baseline set, the accuracy worsens to 55.3% for DeepL and 76.3% for Google. Furthermore, most accurate male translations of the occupations are preceded by a female translation of the nominal group, which emphasizes a mismatch agreement: the male gender of the nominal group is never predicted by DeepL and is only predicted with 48.3% accuracy by Google.

4.4. Analysis of the set with modifier *wise*

Conversely, when adding the adjective *wise*, more associated with a male sphere, both systems translate male roles with high accuracy (occupations with 97.4% and gender agreement with 100%). When the coreferent is *her*, the accuracy achieved by Google Translate remains quite low for professions

(39.5%) and for the nominal group (26.6%). More interestingly, our results show that, if compared to the baseline system, DeepL decreases its performance to 39.5% for occupations, and to 6.6% for the nominal group: the only instance in which the translation maintains the female gender is with the profession *nurse.*

5. Limits and future work

Although the sentence pattern chosen for this study allows to observe the presence of gender bias in the outputs provided by MT systems, it does not allow to have the article before the noun in the Italian translation. Therefore, the gender attributed to the occupation, in case of bigender nouns, can only be observed through the whole nominal group. In future work a sentence pattern should be built in order to avoid this problem, while preserving a focus on agreement between the profession and the nominal group. With regard to the sentence pattern used in this study, the gender of the nominal group must always agree with the gender of the occupation. Therefore, gender agreement is particularly important for “epicene” nouns that do not distinguish between masculine and feminine suffix, and acquire gender from context. In this study, it was decided to keep 13 occupations translated into Italian as bigender nouns to stress that, even if the translation of the occupation is always accurate (and could not be otherwise), a mismatch agreement with the nominal group can affect the translation correctness, thus leading to a wrong translation. To give an example, in the context of *her* the bigender occupation *dentista* requires a female translation of the nominal group to be accurate, which would be *la mia amica*. We should also add that in future studies, given the interest in this regard, all leadership professions should be taken into account, as well as military positions, roles that until a few years ago did not allow the presence of women in them and that are therefore particularly relevant. To give an example, women in Italy had access to the judiciary only in 1965 and in the army only since the 2000s. Furthermore, the number of adjectives examined should be increased, in order to analyse how adjectives concerning physical appearance, personality or professional skills can affect gender bias in MT output.

6. Conclusion

In this paper, we have provided evidence that MT systems like DeepL and Google Translate exhibit a statistical bias towards male defaults, as well as a tendency to reproduce gender stereotypes. The analysis of the baseline set and the set with modifier *strong* show that the best performance is achieved with the male co-referent; conversely, the accuracy is rather low when the co-referent is *her*, leaning towards male defaults, particularly for Google Translate. The sets with modifiers *beautiful* and *wise* confirm our hypothesis that "stereotypical" adjectives can affect the MT output. Adding the "stereotypical" female adjective *beautiful*, the systems' performance improves with the female co-referent and worsens with the male one. On the contrary, when adding the "stereotypical" male modifier *wise*, the systems' accuracy is maintained consistently high with the male co-referent and decreases even further when the co-referent is *her*, if compared to the baseline set.

Furthermore, it can be observed that DeepL performs better than Google Translate with the female co-referent: with the only exception of the set with modifier *beautiful*, Google produces the female gender of the occupation and of the nominal group with low accuracy, often defaulting to male. Finally, both systems are biased towards a male translation for professions mainly held by men; conversely, the bias towards a female translation can only be observed with *nurse*. However, although at a first glance it may seem that the tendency towards male nouns is particularly frequent for male-dominated occupations and high-profile professional fields, the tendency towards male default is also observed for professions mainly held by women (e.g. *librarian*, *hairdresser*, *tailor* etc.).

Regarding the translations in the English-Italian language combination, the bias may be learned from data reflecting the over-representation of male nouns in Italian texts. Highlighting the role played by language in the social construction of reality, the aforementioned guidelines proposed by Sabatini, Robustelli and a growing number of Italian scholars, provide evidence of the discriminatory representation of the female gender through the Italian language, which subordinates women to the male image, and thus perpetuates a series of

prejudices against women. Women disappear in language and in mental representations, and social asymmetries are reproduced in favour of men.

Such guidelines also propose several strategies and techniques to reduce male bias and achieve a gender-fair language. Similarly, our intention is to provide some significant insights about machine bias for post-editors and, more broadly, users of MT systems. Post-editors working in the English-Italian language pair should pay particular attention to hidden errors in MT outputs and to the under-representation of women in the Italian language. It is also important to verify that Italian nouns, pronouns and adjectives are always correctly declined in their male or female form and that a female form should be provided for every professional title.

References

Bahdanau, D., K. Cho and Y. Bengio (2016) "Neural Machine Translation by Jointly Learning to Align and Translate", paper presented at the *3rd International Conference on Learning Representations ICLR 2015* (San Diego, CA, 7-9 May 2015), https://arxiv.org/pdf/1409.0473.pdf.

Beukeboom, C.J. and C. Burgers (2017) "Linguistic Bias", *Oxford Research Encyclopedia of Communication* (July 27), https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-439.

Bureau of Labor Statistics (2018) *Table 11: Employed Persons By Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity, 2018*, Labor Force Statistics from the Current Population Survey, United States Department of Labor, Washington, D.C, https://www.bls.gov/cps/aa2018/cpsaat11.pdf.

Cho, K., B. Van Merrienboer, D. Bahdanau and Y. Bengio (2014) "On the properties of neural machine translation: Encoder–Decoder approaches", in D. Wu, M. Carpuat, X. Carreras and E. M. Vecchi (eds.) *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*,

Doha, Qatar: Association for Computational Linguistics, 103-111, https://arxiv.org/pdf/1409.1259.pdf.

Davis, E. (August 2016) *Winograd Schemas and Machine Translation*, https://arxiv.org/abs/1608.01884.

Escudé Font, J. (2019) *Determining Bias in Machine Translation with Deep Learning Techniques*, UPC, MSc Thesis, https://upcommons.upc.edu/handle/2117/128025.

------ and M.R. Costa-jussà (2019) "Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Technique", in M. R. Costa-jussà, C. Hardmeier, W. Radford and K. Webster (eds.) *Proceedings of the First Workshop on Gender Bias in Natural Language Processing,* Florence, Italy: Association for Computational Linguistics, 147-15, https://arxiv.org/pdf/1901.03116.pdf.

Fioritto, A. (1997) *Manuale di Stile dei Documenti Amministrativi*, Bologna: Il Mulino.

Gheno, V. (2019) *Lingua e Genere: Come Si Declinano le Professioni al Femminile, Semplice Come* (17 April 2019), https://semplicecome.it/tendenze-lingua-genere-professioni-femminile/.

Giusti G. and S. Regazzoni (eds.) (2009) *Mi Fai Male…*, Venezia: Libreria editrice Cafoscarina.

Gonen H. and Y. Goldberg (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 and T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the ACL, Volume 1,* Minneapolis, USA: Association for Computational Linguistics, 609-614, https://arxiv.org/abs/1903.03862.

Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychly and V. Suchomel (July 2013) "The TenTen Corpus Family", in A. Hardie and R. Love (eds.) *Proceedings of the 7th Corpus Linguistics Conference*, Lancaster: UCREL, 125-127,

https://www.sketchengine.eu//wpcontent/uploads/The_TenTen_Corpus_2013.pdf.

Kalchbrenner, N. and P. Blunsom (2013) “Recurrent Continuous Translation Models”, in D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu and S. Bethard (eds.) *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington: Association for Computational Linguistics, 1700-1709, https://www.aclweb.org/anthology/D13-1176.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář., J. Michelfeit, P. Rychlý and V. Suchomel (2014) “The Sketch Engine: Ten Years”, *Lexicography*, 1: 7-36.

Kuczmarski, J. and M. Johnson (2018) “Gender-Aware Natural Language Translation”, *Technical Disclosure Commons* (8 October 2018), https://www.tdcommons.org/cgi/viewcontent.cgi?article=2642&context=dpubs_series.

Lepschy, G. (1989) “Lingua e Sessismo”, in G. Lepschy *Nuovi Saggi di Linguistica Italiana*, Bologna: Il Mulino, 61-84.

Levesque, H., E. Davis and L. Morgenstern (2012) “The Winograd Schema Challenge”, *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 553-561.

Locke, W.N. and A.D. Booth (1955) *Machine Translation of Languages: Fourteen Essays*, New York: Technology Press of the Massachusetts Institute of Technology and Wiley.

Luraghi S. and A. Olita (2006) *Linguaggio e Genere*, Roma: Carrocci.

Madaan, N., S. Mehta, S. Mittal and A. Suvarna (2018) “Judging a Book by its Description: Analyzing Gender Stereotypes in the Man Bookers Prize Winning Fiction”, *CoRR*, abs/1807.10615, https://arxiv.org/pdf/1807.10615.pdf.

Moryossef, A., R. Aharoni and Y. Goldberg (2019) “Filling Gender & Number Gaps in Neural Machine Translation with Black-Box Context Injection”, in M.R. Costa-jussà, C. Hardmeier, W. Radford and K. Webster (eds.) *Proceedings of the First Workshop on Gender Bias in Natural Language Processing,* Florence: Association for Computational Linguistics, 49-54, https://arxiv.org/pdf/1903.03467.pdf.

Pearce, M. (2008) “Investigating the Collocational Behaviour of MAN and WOMAN in the BNC Using Sketch Engine”, *Corpora* 3(1): 1-29.

Prates, M.O.R., P.H. Avelar and L.C. Lamb (2019) “Assessing Gender Bias in Machine Translation - A Case Study with Google Translate”, *Neural Computing and Application*, Springer: London, 1-19.

Robustelli, C. (2000) “Lingua e Identità di Genere”, in *Studi Italiani di Linguistica Teorica e Applicata*, XXIX: 507-527.

------ (2012) *Linee Guida per L’Uso Del Genere nel Linguaggio Amministrativo*, Firenze: Comune di Firenze.

------ (2013) “Infermiera Sì, Ingegnera No?”, Accademia della Crusca, https://accademiadellacrusca.it/it/contenuti/infermiera-si-ingegnera-no/7368.

Rudinger R., J. Naradowsky, B. Leonard and B. Van Durme (2018) “Gender Bias in Coreference Resolution”, in *Proceedings of the 2018 Conference of the North American Chapter of the ACL, Volume 2*, New Orleans, Louisiana: Association for Computational Linguistics, 8-14, https://arxiv.org/abs/1804.09301.

Sabatini, A. (1987a) *Raccomandazioni Per Un Uso Non Sessista Della Lingua Italiana*, Roma: Istituto Poligrafico e Zecca dello Stato.

------ (1987b) *Il Sessismo nella Lingua Italiana*, Roma: Istituto Poligrafico e Zecca dello Stato.

Stanovsky, G., N.A. Smith and L. Zettlemoyer (2019) “Evaluating Gender Bias in Machine Translation”, in *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, Florence: Association for Computational Linguistics, 1679-1684, https://arxiv.org/abs/1906.00591.

Sutskever, I., O. Vinyals and Q. Le (2014) “Sequence to Sequence Learning with Neural Networks”, in Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger (eds.) *Advances in Neural Information Processing Systems 27*, Cambridge, MA: MIT Press, 3104-3112, http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin (2017) “Attention Is All You Need”, in I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.) *Advances in Neural Information Processing Systems 30*, Reed Hook, New York: Curran Associates Inc, 5998-6008, https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Weaver, W. (1955) “Translation”, in W.N. Locke and A.D. Booth (eds.) *Machine translation of languages: Fourteen Essays*, Cambridge: Technology Press, MIT, 15-23.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez and K.W. Chang (2017) “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* Copenhagen: Association for Computational Linguistics, 2979-2989, https://arxiv.org/pdf/1707.09457.pdf.

------ (2018) “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”, in M. Walker, H. Ji and A. Stent (eds.) *Proceedings of the 2018 Conference of the North American Chapter of the ACL, Volume 2*, New Orleans, Louisiana: Association for Computational Linguistics, 15-20, https://arxiv.org/pdf/1804.06876.pdf.