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

Argentina Anna Rescigno[1], Eva Vanmassenhove[2], Johanna Monti[1], Andy Way[3]

[1]UNIOR NLP Research Group, University of Naples L'Orientale
[2]Department of CSAI, Tilburg University [3]ADAPT Centre Dublin City University
a.rescigno1@studenti.unior.it, e.o.j.vanmassenhove@tilburguniversity.edu, jmonti@unior.it, andy.way@adaptcentre.ie

Abstract

This paper presents the results of an evaluation of Google Translate, DeepL and Bing Microsoft Translator with reference to natural gender translation and provides statistics about the frequency of female, male and neutral forms in the translations of a list of personality adjectives, and nouns referring to professions and bigender nouns. The evaluation is carried out for English→Spanish, English→Italian and English→French.

1 Introduction

Gender manifests itself in a language in many ways, and different languages use different linguistic devices to mark (or sometimes 'not mark') gender. When dealing with language, three types of gender come into play: natural gender, grammatical gender and social gender. Natural gender is generally based on the sex of a person or an animal realised by means of the male/female polarity or on the absence of sex for neutral nouns. Grammatical gender, instead, is not always coherent with the semantic categorization of a word and can vary from language to language since it depends on the representation of objects in the world on the basis of specific properties attributed to them in a specific cultural context. Social gender is used in relation to the properties of a word on the basis of which the speakers of a language associate the natural gender of a person to a word (Hellinger and Motschenbacher, 2015): this mainly happens with names of professions, such as for instance *doctor* or *nurse* which are interpreted according to social stereotypes concerning the roles of males and females in the society. Gender is present in the data we use to train MT systems due to the demographic features of the human training data and because of the nature of stereotypes and biases we communicate in our day-to-day communications. As most state-of-the-art MT systems handle translations on the sentence-level, gender phenomena are, usually, resolved on statistics inferred from the training data. Mistranslations of gender information occur more frequently when translating from gender-neutral languages, such as English[1], into morphological-rich languages, such as Italian or French, which explicitly mark gender and require additional information to correctly translate gender phenomena. When such additional information or context is not provided, the system will pick the most likely variant. A recent study by Prates et al. (2018) showed how Google Translate (GT) yields more male defaults than what ought to be expected when looking at demographic data on its own, alluding that there might be a phenomenon they refer to as machine bias (Prates et al., 2018; Vanmassenhove et al., 2019). In this paper, we systematically evaluate: (a) single-word queries, containing personality adjectives and profession nouns, and (b) bigender nouns[2] in an EN → IT, FR, ES translation setting for GT, DeepL (DL) and Bing Microsoft Translator (BMT) to verify the diversity in translations provided by these MT providers.

2 Related Work

As recent years have seen an increase in literature on bias in NLP, we focus particularly on work that has attempted controlling the seemingly random fluctuations in terms of gender in the translations provided by large MT providers, with a specific focus on neural approaches. Rabinovich et al. (2016) conducted a more elaborate series of experiments very similar to the work by Bawden et al. (2016). Their work on preserving original author traits fo-

[1] Aside from pronouns such as 'she' and 'he' or some exceptions such as 'actress' vs 'actor'.

[2] Bigender nouns do not have a fixed grammatical gender; their gender is determined by the context and without any further context, they are valid for both male and female referents.

cuses particularly on gender. They treated personalizing PB-SMT systems as a domain-adaptation task where the female and male gender are two separate domains. In NMT, Vanmassenhove et al. (2018) experimented with the insertion of an artificial token at the beginning of the sentence, indicating the gender of the speaker (Vanmassenhove and Hardmeier, 2018). This approach is similar to Sennrich et al. (2016) who added an 'informal' or 'polite' tag indicating the level of politeness expressed to the training sentences.

The work by Elaraby et al. (2018) presents a technique for the translation of speech-like texts focusing particularly on English-to-Arabic. They train a baseline on generic data (4M sentences) and use a set of gender-labelled sentences (900k) in order to tune the system towards generating translations with correct gender agreement.

More recently, Moryossef et al. (2019) presented a simple yet effective black-box approach to control the NMT system's translations in case of gender ambiguity. Instead of appending a token, they concatenate unambiguous artificial antecedents with information on the speaker and the interlocutors to ambiguous English sentences. Some recent studies have addressed the problem of the scarcity of publicly available corpora and create corpora specifically designed to evaluate or to test MT performance with respect to gender translation (Font and Costa-Jussa, 2019; Di Gangi et al., 2019). Finally, Monti (2020) provides an overview of outstanding issues and topics related to gender in MT and Sun et al. (2019) a literature review of work related to gender bias in the field of NLP.

3 Experimental setup

For the experiments, we compiled a dataset with the translation of both English single-word queries and short sentences into Italian (IT), French (FR) and Spanish (ES) with GT, BMT and DL. We experimented with both single-words and short sentences as e.g. GT provides, since end 2018[3] gendered translations for single-word queries (limited to nouns and adjectives) for EN–FR, EN–ES and EN–IT. Similarly, BMT and DL provide users with alternative translations on the user interface. Differently from GT, these are simply alternative translations for the word in question. As such, GT is the only system that currently, to some extent, deals with gendered variants in a systematic way.

3.1 Compilation of Datasets

The datasets[4] are compiled on the basis of a list of nouns and adjectives collected from different sources (see Table 1). The translations generated and their manual evaluations are also part of the datasets. The setup for this experiment consists of both words and sentences. We collected a set of 136 personality adjectives and 107 nouns of professions from three different sources and 30 of the most common bigender nouns in the Italian language. We tested these separate sets and analysed the behaviour of the major state-of-the-art MT systems, comparing the translations of the three language pairs. The first two sets of words have been assessed alone, without any context, while the last set has been examined within the sentence level.

Alongside with the number of adjectives and nouns retrieved, Table 1 provides more detailed information on the sources and the original language in which the data was retrieved.

	#	Sources
Adjectives	136	(I, 2019a); (II, 2019a);(III, 2019)
Professions	107	(I, 2019b); (II, 2019b)
Bigender	30	(Cacciari et al., 1997); (Cacciari et al., 2011) (Thornton and Anna, 2004)

Table 1: Overview of adjectives, profession and bigender nouns along with the sources from which they were retrieved

3.2 Description of the MT Systems

To evaluate the gender issues in translation we used three state-of-the-art, freely available, MT systems:

GT: Launched in 2003 as a statistical MT system, GT has switched to a NMT system in 2016 (Monti, 2017). The translations are generated at the sentence level. Since 2018, Google provides two alternatives when translating 'ambiguous' or underspecified English words into languages that have male\female alternatives for various languages (e.g. Italian, French and Spanish). The male\female variants are listed alphabetically (i.e. first the female variant, then the male one).

[3] https://www.blog.google/products/translate/reducing-gender-bias-google-translate/

[4] available at https://github.com/argentina-res/gender_project.git

BMT: MT system owned by Microsoft that originally used a statistical approach (Monti, 2017) but more recently switched to a neural system (Almahasees, 2018). Unlike GT, BMT does not provide alternatives in the translation box itself. However, it does give synonyms in the "Other ways to say" section and provides examples of usage in the "How to use..." section, where sometimes the female form of a word is listed (by chance rather than in a consistent way).

DL: The most recent platform launched in August 2017 by a German company, DeepL GmbH. DL uses convolutional neural networks based on the Linguee database (Morán Vallejo and others, 2019). Even though it only supports nine languages (all Indoeuropean), DL is, according to a recent study, outperforming the other competitors (Morán Vallejo and others, 2019). The layout of the interface is similar to that of GT. Nevertheless, its suggestions for alternatives are not systematically morphological variants (although they are often somehow included in the alternatives provided). Underneath the actual translations, there is also a dictionary-like section when translating words in isolation.

4 Results

In this section, the results obtained with our experiments and our subsequent evaluation will be presented. A more in-depth analysis and a discussion of some concrete examples is provided in Section 4.1. All the manual evaluations were conducted in November, 2019[5].

The **adjectives and profession nouns** were both evaluated in single-word settings. The **bigender nouns** were evaluated in short sentences. Bigender nouns in Italian, such as for example the nouns *giornalista, pianista* are by themselves not marked for gender. As such, a single-word query would not reveal any gender marking. However, the articles, adjectives, verbs, etc. that agree with these bigender nouns are (often) marked for gender. Therefore, these specific nouns were evaluated in short sentence settings.

We manually evaluated all the outputs and will report on the percentage of male (M), female (F) and 'neutral' (N) or 'covered' (C) (i.e. no explicit gender in the target language) variants for the single-word and sentence queries. Additionally, we report on the errors (e.g. untranslated or mistranslated words).

As mentioned earlier, GT is currently the only system that provides both male and female variants when given a single-word ambiguous query (EN→FR, ES, IT). The alternatives are limited to adjectives and nouns[6]. DL and BMT provide the user with multiple alternatives underneath the primary translation for both nouns and adjectives but not in a systematic way. As such, these alternatives do not necessarily consist of morphological variants in terms of gender.

For our evaluation, only the main translations offered by the MT systems were considered, i.e. we evaluated both gendered forms for GT, but did not evaluate the list of alternative translations provided by BMT and DL as they can be alternatives of any kind and they differ depending on the query.

In the interest of clarity and order, we will separately consider the different test-sets, single words (i.e. adjectives and nouns), and sentences with bi-gender nouns. Each set has been tested in the previously stated systems and language pairs. In the tests, we were especially investigating – apart from any translation error – the occurrence of female forms, to see if there is somehow a bias towards the gender. From a practical point of view, we will consider only the first output as a valid result, since the systems also give "alternatives" for single-word translations. For Bing, it is quite straightforward: adjectives do not present alternatives, while nouns sometimes do. Google Translate produces now two different results, marked with the gender (feminine/masculine, in this sequence for alphabetical order). DeepL, in particular, provides at least three alternative translations, for both nouns and adjectives. However, even though we are not considering alternatives, we will explain or anyway mention the ones that are significative in this research. All the results have been recorded in November 2019. However, the systems continually improve, so results may vary also in a short time.

Table 2 presents the results for the single-word translations consisting of adjectives: for all three systems, the male variant was the most common. This was especially so for BMT, where only 1.5% of the adjectives were translated into a female variant. The 'other' category consists of translations:

[5] The translations were evaluated by a native Italian speaker\linguist for Italian, and by a professional linguist for French and Spanish.

[6] In the target languages certain verbs are also sometimes marked for gender, e.g. reciprocal verbs, passive constructions, past participles...

(a) that were ambiguous (e.g. *mean* was translated as a verb instead of as an adjective), (b) words that were not translated by the systems and (c) errors.

ADJ	GT	BMT	DL
F	37.3	1.5	22.8
M	**39.2**	**58.8**	**45.6**
N	20.7	33.1	26.5
Other	2.8	6.5	5.1
Total	100	100	100

Table 2: Results in % for male (M), female (F) and neutral (N) adjectives generated for EN → IT for GT, BMT and DL. The "Other" label includes all results obtained that do not correspond to the "adjective" category

We ought to note that, none of the single words were mistranslated by GT. The only single-word queries that caused a divergence between male and female forms were queries that were a compound noun in either the source or the target (e.g.*good-tempered*). They are not treated as a single unit by GT and thus the system fails to render both variants. From the Table 2, it is clear that GT performs best in terms of balanced single-word adjective translations. Table 3 presents a similar set of results but for the nouns indicating professions. Like Table 2 GT generated the most diverse translations, while BMT the least. As far as the set with sentences is concerned, we used bigender nouns from the Italian language. We used 30 common bigender nouns in two different contexts: (a) first, in a minimal sentence that would allow us to infer the gender based on the article in the target language "I am a(n).." and (b) with a referring adjective. We used *beautiful, efficient, intelligent, sad* and *famous*. In Table 4, our results are presented for the bigender nouns on minimal sentences ("I am a(n)...") and in combination with the aforementioned adjectives ("I am a(n) + adj..."). In the results, we oppose the translation where we added *beautiful* as an adjective as they differed considerably from the others. Table 4 presents the results for the translations generated by BMT for bigender nouns in sentences for Italian, French and Spanish. It can be noted that the male translation is the most common in the simple sentences that do not contain an adjective in all three languages. However, when adding the adjective *beautiful* to the phrase, the female forms are the most common for all three languages. An example of such sentences is given below:

(a)	EN:	"I am a pianist"	(N)
	IT:	"Sono un pianista."	(**M**)
	FR:	"Je suis pianiste."	(N)
	ES:	"Soy pianista."	(N)
(b)	EN	"I am a beautiful pianist."	(N)
	IT	"Sono **una bellissima** pianista."	(**F**)
	FR	"Je suis **une belle** pianiste."	(**F**)
	ES	"Soy **una hermosa** pianista."	(**F**)
(c)	EN	'I am a famous pianist."	(N)
	IT	"Sono **un famoso** pianista."	(M)
	FR	"Je suis **un** pianiste célèbre."	(M)
	ES	"Soy un pianista **famoso**."	(M)

The results obtained for DL are very similar to the ones obtained with BMT except for the fact that DL generates overall more female forms than BMT. Interestingly, among all systems, GT is the most biased towards male forms when evaluating entire sentences for all three language pairs, with the male forms being the dominant ones for all categories and for several set-ups we observe more than 90% male variants.

NOUN	GT	BMT	DL
F	35.8	0.9	7.5
M	**46.1**	**60.4**	**60.4**
N	17.6	28.3	28.3
Other	0.6	10.5	3.7
Total	100	100	100

Table 3: Results in % for male (M), female (F) and neutral (N) nouns generated for EN → IT for GT, BMT and DL. The "Other" label includes all results obtained that do not correspond to the "noun" category

4.1 Analysis

In the analysis we compare the results of the three systems, comparing the occurrences of the female-gendered translated forms in terms of the different systems and the different languages.

GT: for the single nouns and adjectives, GT provides both male and female forms. The system, however, produces more male outputs as sometimes the alternative is not provided (e.g. for compound nouns). One of the provided nouns was ambiguous (*printer*) in English. The system translated this as the object instead of the profession.

BMT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	10.0	**86.7**	Q^*	10.0	**63.3**	26.7	3.3	**66.7**	30.0
beautiful	**63.3**	36.7	0.0	43.3	**56.7**	0.0	**66.7**	33.3	0.0
other adj.	13.3	**83.3**	Q^*	3.3	**96.7**	0.0	6.7	**93.3**	0.0
DL	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj	30.0	**70.0**	0.0	20.0	**63.3**	16.7	3.3	**76.6**	20.0
beautiful	**83.3**	16.7	0.0	**73.3**	26.7	0.0	**96.7**	3.3	0.0
other adj.	**53.3**	43.3	Q^*	13.3	**83.3**	3.3	6.7	**93.3**	0.0
GT	IT			FR			ES		
	F	M	N	F	M	N	F	M	N
no adj.	6.7	**93.3**	0.0	6.7	**90.0**	3.3	3.3	**66.7**	30.0
beautiful	43.3	**56.7**	0.0	**80.**	20.0	0.0	**80.0**	20.0	0.0
other adj.	3.3	**96.7**	0.0	3.3	**96.7**	0.0	3.3	**96.7**	0.0

Table 4: Results in % for male (M), female (F) and neutral (N) forms generated for EN → IT, FR and ES for BMT, DL and GT

Whenever an ambiguous word was translated accurately, yet not in the way we intended it to be translated, we included it into the 'other' category. Considering the adjectives, we observed one incorrect translation where the adjective *supportive* was translated into an Italian noun ('*supporto*' meaning *support*). Two other adjectives were ambiguous (*mean* and *kind*) and were translated into a verb and a noun respectively by GT. For the sentence-evaluation, GT has the strongest preference for translations using male-endings.

BMT: for the nouns, BMT has a strong tendency to output male variants. The only two exceptions are *nurse* for IT, FR and ES and *makeup artist* for FR and ES. Besides, words such as *newsreader*, *translator*, *warder* were not translated by BMT and others were mistranslated, e.g. *garbage man* and *window cleaner*, for which the translations provided were too literal in IT (*uomo spazzatura*) and *pulizia finestre* where "pulizia" is the equivalent of "cleaning"). Concerning the adjectives, unlike DL and GT, BMT rarely offers alternatives and the majority of the translations generated are in the male form. Exceptional female variants were found in: (a) Italian for: *devious/subdol**a*** and *joyful/gioios**a***; (b) Spanish for: *artistic*, *bossy*, *calm*, *diplomatic*, *dynamic*, *extrovert*, *humorous*, *industrious*, *placid*. We ought to note that a considerable amount of the adjectives (33.1% for Italian and 29.4% for Spanish) have a translation with a bigender adjective. These words have the same form for both genders and are included in the 'N' (neutral) group. A small number of adjectives were translated with an expression, such as *good-tempered* → IT *di buon umore*, FR *de bonne humeur*, SP *de buen humor*. These expressions can be assigned to both male/female referents and are thus considered covered/neutral. We did not observe any errors except for the adjective *frank (without a capital letter)*, which was left untranslated by DL (as if it were the first-name "Frank"). However, the most appropriate translation was among the alternatives suggested. Moreover, sometimes the adjectives are ambiguous, therefore the system has opted for a "non-gendered" alternative (e.g. *mean* was translated as a verb instead of an adjective).

DL: DL provides multiple options, for both nouns and adjectives, but for only 7.5% of nouns the female form is the first result, as, for example, *assistant* and *nurse* for French and Italian, *doctor*, *secretary* and *shop assistant* for Spanish and Italian, *soldier* and *teacher* for Italian. For the adjectives, instead, the number of female forms increases to 22.8%. Similarly to some of the observations for BMT and GT, some of our intended nouns were ambiguous ('model' can be a noun or a verb) and the system opted for the verb translation. The only error we observed is the translation of *tailor* which was incorrectly translated into the adjective *sartoriale*.

5 Conclusions and Future Work

In future work, we would like to conduct a larger evaluation comprising of more language pairs and a more diverse set of words. Furthermore, we aim to compile a challenge set focusing specifically on gender phenomena in language that can be used and automatically evaluated. We also envisage training our own state-of-the-art MT system to verify how and whether machine bias indeed influences the output of the translations generated.

References

Zakaryia Mustafa Almahasees. 2018. Assessment of google and microsoft bing translation of journalistic texts.

Rachel Bawden, Guillaume Wisniewski, and Hélène Maynard. 2016. Investigating gender adaptation for speech translation.

Cristina Cacciari, Manuel Carreiras, and Cristina Barbolini Cionini. 1997. When words have two genders: Anaphor resolution for italian functionally ambiguous words. *Journal of memory and language*, 37(4):517–532.

Cristina Cacciari, Paola Corradini, Roberto Padovani, and Manuel Carreiras. 2011. Pronoun resolution in italian: The role of grammatical gender and context. *Journal of Cognitive Psychology*, 23(4):416–434.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2012–2017. Association for Computational Linguistics.

Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6, April.

Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Marlis Hellinger and Heiko Motschenbacher. 2015. Gender across languages. the linguistic representation of women and men, volume 4. amsterdam & philadelphia.

Adjectives Sources I. 2019a. Personality adjectives source i. `https://www.esolcourses.com/content/exercises/grammar/adjectives/personality/words-for-describing-personality.html`.

Professions Source I. 2019b. Professions source i. `https://www.scribd.com/doc/82021393/List-of-Common-Jobs`.

Adjectives Source II. 2019a. Personality adjectives source ii. `https://www.esolcourses.com/content/exercises/grammar/adjectives/personality/more-words-for-describing-personality.html`.

Professions Source II. 2019b. Professions source ii. `https://www.vocabulary.cl/Basic/Professions.htm`.

Adjectives Sources III. 2019. Personality adjectives source iii. `https://7esl.com/adjectives-that-describe-personality/`.

Johanna Monti. 2017. Questioni di genere in traduzione automatica. In *Al femminile. Scritti linguistici in onore di Cristina Vallini*, pages 411–431. Cesati.

Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge.

Alberto Morán Vallejo et al. 2019. The translation of spanish agri-food texts into english and italian using machine translation engines: A contrastive study.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. *arXiv preprint arXiv:1903.03467*.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2018. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July.

ANNA Thornton and M Anna. 2004. Mozione. *La Formazione Delle Parole in Italiano*, pages 218–225.

Eva Vanmassenhove and Christian Hardmeier. 2018. Europarl datasets with demographic speaker information.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland, August.