Gender Issues between Gemini and ChatGPT: The Case of English-Arabic Translation

Faiz Algobaei[1], Elham Alzain[2], Ebrahim Naji[3], Khalil A Nagi[4]

[1] Northern Border University, Saudi Arabia. E-mail: realfaiz@gmail.com

[2] King Faisal University, Saudi Arabia. E-mail: elhamalzain@gmail.com

[3] Trine University, United States. E-mail: najie@trine.edu

[4] University of Saba Region, Yemen. E-mail: khalil.naji@usr.ac

Correspondence: Faiz Algobaei, Northern Border University, Saudi Arabia. E-mail: realfaiz@gmail.com

Received: July 2, 2024 Accepted: August 15, 2024 Online Published: August 24, 2024

doi:10.5430/wjel.v15n1p9 URL: https://doi.org/10.5430/wjel.v15n1p9

Abstract

The study focuses on the gender-related issues that face English-Arabic machine translation. It aims to investigate and evaluate gender accuracy in the translations provided by two prominent large language models, Gemini and ChatGPT, recognizing the rich morphological system of Arabic that includes gender marking. The researchers develops a test suite to evaluate gender accuracy in the translation outputs of Gemini and ChatGPT. The evaluation is performed by two professional annotators. That is followed by an analysis of the patterns of the gender-related issues that appear in the translation outputs of the models under study. The results show that Gemini outperformed ChatGPT in almost every aspect when it comes to gender-related translation issues. Both the number of the annotated issues as well as the gender accuracy evaluation came in favor of Gemini. The study introduced different patterns of gender-related translation issues. It also provides recommendations for future research.

Keywords: gender-related issues, gender bias, accuracy evaluation, Gemini, ChatGPT, English, Arabic

1. Introduction

Machine Translation (MT) systems have shown remarkable development and are considered to be great tools that increase the productivity of professional translators. They are seen as a more cost-effective alternative to human translation. However, in the literature of translation studies, there are several heated arguments regarding the quality of machine translation and whether MT has achieved parity with professional human translation (Hassan et al., 2018; Barrault et al., 2019; L äubli et al., 2018; Toral et al., 2018; Freitag et al., 2021).

It is noticeable and undeniable that neural machine translation (NMT) systems have advanced greatly and provide high-quality translation outputs. However, it is also clear that there is still a gap between machine translation and professional human translation. Recent studies that performed error analysis revealed a comparatively significant list of errors (Popović, 2021; Kocmi et al., 2022; Nagi, 2023).

Now, with the advent of new large language models (LLMs), a new wave of research has started to investigate and evaluate the quality of translation provided by LLMs such as ChatGPT and Gemini (Hendy et al; 2023; Jiao et al., 2023; Siu, 2023, among others). Since the improvement of automatic translation requires more fine-grained analyses regarding translation quality (Daems et al., 2014; Popović, 2021; Kocmi et al., 2022; Rivera-Trigueros, 2022), it is, therefore, crucial to scrutinize areas of translation that form a problem for NMT systems and to evaluate the performance and quality of translation of the new LLMs .

One of the problems that faces MT is gender. It is interesting from different aspects since MT systems are known to be susceptible to bias and the variation between languages when it comes to gender and it forms a sensitive issue that has been studied to investigate and provide solutions to such issue. Certain procedures were adopted in the literature of MT to mitigate gender bias. Although these procedures have been proven to be beneficial, they are not very efficient and the issue still prevails (Vanmassenhove et al., 2018; Habash et al., 2019; Stafanovičs et al., 2020; Alhafni et al., 2020, among others).

The research team, therefore, presents this study to contribute to investigating issues related to MT and gender. However, since LLMs are still under-investigated in the literature of machine translation, the research team performs an exploratory investigation regarding what gender issues can arise in the translation outputs of ChatGPT and Gemini. It should be mentioned that although English and Arabic are both non-neutral gender languages, there are huge differences regarding gender in their grammar such as differences related to gender agreement. For instance, Arabic requires gender agreement between various syntactic objects, such as subjects and verbs, adjectives and nouns they modify, and subjects and non-verbal predicates.

Since Arabic has a rich morphological system which includes gender marking, it is expected that gender issues arise when translating from a morphologically poor language like English into Arabic despite the remarkable development of machine translation. The

researchers, therefore, build a test suite to evaluate gender issues when translating from English to Arabic using ChatGPT and Gemini. It evaluates gender accuracy and provides a focused analysis of the gender-related translation issues that occur in the translation outputs of ChatGPT and Gemini. By conducting such an analysis, the study provides an insight on the nature of errors and the possible structural reasons for them. Based on that, further specific investigations to be performed in the future are suggested.

This study may not capture the intricacies of gender representation in other languages due to the binary gender system in Arabic. The study also relies on the translation outputs of two specific language models, Gemini and ChatGPT. Furthermore, the proposed structural relationships of the occurrence of gender translation errors can specifically apply to English into Arabic translation.

2. Literature Review

This section of the study is divided into two parts. In the first part, we briefly explore the recent research trends of gender and translation in MT systems before LLMs, namely NMT systems. The second part, however, introduces the LLMs under investigation, ChatGPT and Gemini.

2.1 MT before LLMs and Gender

As it is mentioned earlier, gender forms a sensitive issue for machine translation. Therefore, various recent studies investigated gender issues in machine translation and they made use of test suites to evaluate gender bias in MT. Recent test suites included WinoMT which was composed by Stanovsky (2019) as a concatenating of WinGender, a test suite created by Zhao et al. (2018), and WinoBias, a test suite created by Rudinger et al. (2018). This test was composed of English instances that were translated into six languages with grammatical gender including Arabic. The test was used in other studies where it was updated and applied to other languages (Saunders & Byrne, 2020; Kocmi et al, 2020; Costa-jussà, 2022; Saunders et al., 2022). The efficiency of the WinoMT test suite is unquestionable. However, it is limited to cases where gender bias results from stereotyping occupation. In addition, WinoMT is based on synthetic examples and may not fully represent natural language use.

Another interesting test suite in the literature of MT is the MT-GenEval (Currey et al., 2022). This test suite is more diverse and realistic than WinoMT despite applied gender restrictions. It was applied to eight target languages which also included Arabic .

Interesting endeavors regarding Arabic MT were also made. Habash et al (2019) introduced a corpus for gender identification with a limited focus on first-person singular structures. Alhafni et al. (2020) introduced a new corpus that was limited to the first and the second grammatical persons.

Research was made in endeavor to mitigate gender bias and proposals on mitigating the issue were proposed such as gender tagging in Vanmassenhove et al. (2018) and Stafanovičs et al (2020), gender reinflection in Habash et al (2019) and Alhafni et al. (2020). However, despite of slight improvements, the problem persists especially when involving rich morphology languages which include Arabic.

Therefore, it can still be stated that gender still forms a major issue for machine translation. The issue is assumably larger when it comes to English-Arabic machine translation due to the morphological convergences between the two languages and their different gender systems. This study, therefore, investigates such aspects presenting an evaluation of the translations provided by Gemini and ChatGPT which have been among the largest language model (LLM) that has received the most remarkable growth of interest in various fields including the field of automatic translation. The study presents a focused analysis of the gender-related translation issues that appear in translating English complex clauses into Arabic using ChatGPT and Gemini .

The variation between the grammatical gender system in English and Arabic is a great factor in making the study more interesting. It can be stated that Arabic and English have different grammatical gender systems. Arabic has two genders: masculine and feminine. English has three genders: masculine, feminine, and neuter. There are no neuter nouns in Arabic. All nouns are either masculine or feminine and the gender of an Arabic noun can be determined by either its meaning or its form where a feminine noun can have a gender marking in most case but not in many others. Masculine nouns would be known by the absence of such marker taking into consideration that many feminine nouns, as mentioned, share this feature with masculine nouns. A verb in Arabic needs to match the subject in terms of gender and a modifying adjective needs to agree with the corresponding noun in terms of gender as well. In English, on the other hands, nouns generally do not have gender markings. We have some nouns that indicate the natural gender of the referee such as boy or woman. There are also a few nouns like *actor* and *actress* that show variation regarding gender. However, these nouns do not require special agreement with verbs or modifying adjectives. Most English nouns do not refer to a specific gender such as *teacher* and *doctor*. In the case of pronouns, *he* is masculine, *she* is feminine, and the rest are neuter. No gender agreement is required between the subject and the verb or the adjective and the noun it modifies. That is why it is said that English does not have a grammatical gender .

However, before investigating the gender-related translation issues when translating from English to Arabic using LLMs, the following part provides a quick review of ChatGPT and Gemini, the LLMs that this study investigates their performance in relation to the mentioned issues.

2.2 LLMs Translation and Gender

2.2.1 ChatGPT

ChatGPT is the most well-known artificial intelligence (AI) application recently whose popularity began even before its official launch. Companies like BBC, CNN, and People's Daily announced the forthcoming AI revolution. ChatGPT quickly rose in popularity due to its

ability to perform various tasks effectively. These tasks include generating text, classifying text, answering questions, writing codes, and translating languages (Siu, 2023) .

Regarding translation, Jiao et al. (2023) and Hendy et al. (2023) indicated that the performance of ChatGPT is on par with that of commercial translation systems such as Google Translate when it comes to the translation of high-resource European languages. However, it lags behind in translating low-resource languages. It is also indicated that ChatGPT struggles with translating biomedical abstracts or Reddit comments, but it excels at translating spoken language. Jiao et al. (2023) also noted that the GPT-3.5 model underperforms in specific domains when compared to its performance in the translation of spoken languages. Khoshafah (2023) stated that ChatGPT struggles with specialized texts such as scientific, medical, legal, or literary texts, however, it performs well with simple content. Zhu et al. (2023) also indicated that while the multilingual translation capabilities of LLMs are improving, GPT-4 still has not achieved the desired level of performance when it comes to low-resource languages. Nagi et al. (2024) also stated that ChatGPT translation outputs show a high error frequency when translating English complex sentences into Arabic.

Generally speaking, it was pointed out that ChatGPT generates gender-biased texts (Wan et al., 2023; Hada et al, 2023; among others). Recent studies also showed that gender issues are present in ChatGPT translation outputs. Ghosh & Caliskan (2023) conclude that ChatGPT sustains the same issues related to gender biases as it is the case in NMT systems such as Google Translate or Microsoft Translator. However, it is noticed that, in general, gender-related issues in LLM translation outputs are nearly untouched, and to the best of the researchers' knowledge, there is no study about these issues when a text is translated into Arabic.

2.2.2 Gemini

Gemini is the model that is developed at Google. It was introduced at the end of 2023 as a replacement for Bard that was introduced earlier. The first version, Gemini 1.0, includes Gemini Ultra (for highly-complex tasks), Gemini Pro (for enhanced performance and deployability at scale), and Gemini Nano (for on-device applications). Their architecture has been built on top of Transformer decoders which is enhanced with improvements in architecture as well as model optimization to enable stable training at scale. The pre-training dataset of Gemini models was both multimodal and multilingual. It used various data that was obtained from web documents, books, and code. It also included image, audio, and video data (Team et al., 2023) .

Regarding machine translation, Team et al. (2023) assessed a post-trained Gemini Ultra model on all language pairs in the WMT 23 translation benchmark. It was stated that it demonstrated exceptional performance in translating from English to other languages and outperformed other LLM-based translation methods when translating from other languages to English. They also pointed out that it achieved the highest LLM-based translation output quality, since it achieved an average of 74.8 BLEURT (Sellam et al., 2020) score compared to GPT-4's 73.6 score and PaLM2's 72.2 score when translating from English to the other languages. A general average of 74.4 BLEURT score was also achieved by Gemini, followed by an average of 73.8 achieved by GPT-4 and an average of 72.7 achieved by PaLM 2.

To the best of the researchers' knowledge, gender issues in the translation outputs generated by Gemini are still untouched, especially when it comes to an Arabic translation output.

3. Methodology and Results

3.1 The Test Suite

Unlike test suites that are used to test gender issues, the research team builds a realistic test suite that does not include built-up sentences but sentences that are selected from various resources. The research team does not just ensure that the test suite is realistic but also divergent. The test suite is composed of 160 complex sentences that are divided into two main parts. The first part is composed of 80 complex sentences that are selected from different news essays with various topics such as daily life, culture, sports, technology, health, etc. The other 80 complex sentences are extracted from various academic and scientific texts which include topics on education, medicine, biology, and technology.

The test suite is built to test the performance of automatic translation systems under study in translating gender related issues from English to Arabic. The research team ensures that the test suite is composed from complex sentences rather than simple ones because they form a challenge for MT and, accordingly, certain procedures such as source text simplification have been proposed in the literature to get more accurate translation outputs (Hasler et al., 2017; Štajner & Popović, 2018; Niklaus et al., 2019; Sulem et al., 2020; Lu et al., 2021, among others). Qasmi et al. (2020) and Turganbayeva et al. (2022) stated that complex sentences form a greater challenge to MT systems when a morphologically rich language is involved. Therefore, with the rich morphology of Arabic and its requirements of gender agreement, it is natural that complex sentences form a great challenge for MT systems when it comes to gender-related issues. It is also expected that a test suite constructed of complex sentences would work great as an exploratory tool to examine the performance of the new LLMs in the aspect under study.

It should be noted here that, as opposed to test suites used to investigate issues related to gender in MT, the test suite used in this study is not limited to human referents since every noun in Arabic, even inanimate nouns, are included under the feminine class or the masculine class.

3.2 Annotation and Evaluation

Gender accuracy is the primary metric for evaluation and it is calculated by the percentage of instances in which the produced translation has no gender-related issues. This evaluation is based on the proposal of Stanovsky et al. (2019). A two-annotator team scores the translations based on the correctness of gender-marked words. The annotators are professional and native-like fluent in both English and Arabic. A detailed analysis is provided to identify patterns of translation issues related to gender.

3.3 Distribution of Gender Errors

This section presents the distribution of gender translation errors in the translated sentences of the test suite under study. It also presents an evaluation of the translation based on gender accuracy as mentioned earlier .

It should be mentioned here that Arabic is a masculine language. Therefore, when a masculine translation of a general singular or plural noun, i.e., a noun with a generic reference, is provided, the translation is not considered to be biased. However, when either a masculine or a feminine translation of a singular noun has a specific reference, the issue is pointed out as biased (tendency to masculine or tendency to feminine) when the context does not indicate the gender of the referent. It also includes pronouns like *it* and *them* when they have a specific exophoric reference. In that case, it is pointed out whether the tendency of the system is towards a masculine or a feminine translation in the results of the table below. This type of issue is classified under gender bias and the rest of the issues are included under others. The prominent pattern of the issues under others are presented later in Section 3.5.2.

The tables below present the number of gender issues annotated in the translated sentences. Table 1 presents the number of issues in the scientific and academic sentences. Table 2, on the other hand, presents the number of issues in the news sentences

Table 1. Number of Gender Issues in Scientific and Academic Sentences

Models	ChatGPT			Gemini		
Issues	Gender Bias		Other Gender Issues	Gender Bias		Other Gender Issues
	Tendency to Masculine Translation	Tendency to Feminine Translation		Tendency to Masculine Translation	Tendency to Feminine Translation	
No. of Issues	3	-	15	-	3	6
Total No. of Issues	18			9		

Table 2. Number of Gender Issues in News Sentences

Models	ChatGPT			Gemini		
Issues	Gender Bias		Other Gender Issues	Gender Bias		Other Gender Issues
	Tendency to Masculine Translation	Tendency to Feminine Translation		Tendency to Masculine Translation	Tendency to Feminine Translation	
No. of Issues	8	-	6	5	3	2
Total No. of Issues	14			10		

From the tables above, it can be stated that Gemini performs better with translating both scientific and academic sentences and news sentences when it comes to gender-related issues. As shown in Table 1, the number of detected issues in Gemini translation outputs are 9 in scientific and academic sentences compared to 18 detected issues in ChatGPT translation outputs. Table 2 also shows that the number of detected issues in Gemini translation outputs are 10 in scientific and academic sentences compared to 14 detected issues in ChatGPT translation outputs. Therefore, the total number of issues in Gemini translation outputs are 19 compared to 32 errors in ChatGPT translation outputs, which indicates that Gemini outperforms ChatGPT significantly in this aspect.

It is also noticed that ChatGPT performs better with news sentences. Gemini, on the other hand, has a very slight difference in favor of its performance with scientific and academic sentences. Both models seem to perform better in terms of gender bias when translating scientific and academic sentences. The reason behind this can be due to the context itself. It is noticed that the news sentences contain more nouns that refer to people and occupations, which causes the increase in the number of bias issues.

3.4 Gender Accuracy

This section provides an evaluation based on the proposal of Stanovsky et al. (2019) where the accuracy is calculated by the percentage of translated sentences in which the produced translation has no gender-related issues. Table 3 below presents the number of correct sentences and their percentage out of the number of sentences included in the test suite.

Table 3. Correct Sentences and Gender Accuracy

	Correct Scientific and Academic Sentences	Accuracy in Scientific and Academic Sentences	Correct News Sentences	Accuracy in News Sentences	Total of Correct Sentences	Total Accuracy
ChatGPT	63	78.75%	68	85%	134	81.875%
Gemini	72	90%	71	88.75%	142	89.375%

In concordance with the number of detected issues, the number of correct sentences and the gender accuracy also show that Gemini performs better than ChatGPT when it comes to gender-related issues. The table above shows that Gemini performs better than ChatGPT with scientific and academic sentences with 72 correct sentences out of 80 and an accuracy of 90% compared to 63 correct sentences out of 80 and an accuracy of 78.75% in ChatGPT. Gemini also performs better with news sentences with 71 correct sentences out of 80 and an accuracy of 88.75% compared to 68 correct sentences out of 80 and an accuracy of 85% in ChatGPT. The results show again that Gemini performs slightly better when translating scientific and academic sentences (72 correct sentences and 90% accuracy) than when translating news sentences (71 correct sentences and 88.75% accuracy). ChatGPT, on the other hand, performs better when translating news sentences (68 correct sentences and 85% accuracy) than when translating scientific and academic sentences (63 correct sentences and 78.75% accuracy).

3.5 Patterns of Gender Errors

3.5.1 Gender Bias in ChatGPT and Gemini

The gender-related issues are divided into two main categories: gender bias and other issues. Gender bias, as noted by Savoldi et al. (2021), is an "umbrella term" that refers to "a wide array of undesirable phenomena" (p. 856). However, in this study, gender bias is limited and it refers to issues related to stereotyping of certain nouns / pronouns in contexts that does not indicate the gender of that noun / pronoun reference. This issue raised due to the fact that English has neutral nouns and pronouns which is not the case in Arabic. It should be noted that ChatGPT provides 100% masculine translations of all related terms. English nouns like little one, customer, clerk, relative, interviewer, and employer are translated into their masculine equivalents when there is no indication of the gender in the translated sentences. Gemini also seems to follow the "masculine tendency" of Google Translate that Schiebinger (2014) refers to in her essay where most of the English related nouns are translated into their Arabic masculine counterparts. It performs a little better in terms of masculine tendency where little one and clerk are translated into their feminine counterparts regardless of unavailability of any indication of the gender of the reference.

Pronouns like *it* and *them* are neutral in English where they refer to both masculine and feminine. In Arabic, however, there are feminine and masculine versions of them. In the translation of these pronouns in sentences where the antecedent is not available and there is no indication of the gender of the reference, each model takes a different path. Whereas ChatGPT translates all examples of these pronouns into their Arabic masculine counterparts, Gemini translates all examples of these pronouns into their Arabic feminine counterparts.

3.5.2 Other Gender Issues in ChatGPT and Gemini

In this part, we discuss the patterns of other gender-related issues concerning some structural features of the positions where they appear and their relationship to other syntactic objects. It is noticed that most issues can be put under the following points.

Inanimate nouns, abstract nouns, and scientific terms (nouns of non-human referent): Even inanimate nouns, abstract nouns, and scientific terms are considered to be masculine or feminine and they require gender agreement with the related syntactic objects according to Arabic grammar. The word *window*, as a subject, requires a verb with a feminine marking whereas the word *door* requires a verb with a masculine marking. ChatGPT has performed poorly in this aspect and many gender-related issues are annotated where inanimate nouns, abstract nouns, and scientific terms do not agree in gender (as subjects) with the corresponding verbs or (as antecedents) with their anaphoric pronouns. Fifteen related issues are spotted in ChatGPT translation outputs. Gemini has performed poorly as well. However, it has performed much better than ChatGPT in this aspect where 4 issues of the kind are spotted in its translation outputs.

Coordinated subjects: Only one issue is annotated in the translation outputs of each model. However, that does not mean that ChatGPT and Gemini have performed well regarding this aspect. The reason can simply be that there are not many coordinated subjects in the test suite.

Relative pronouns and resumptive pronouns: Relative pronouns and resumptive pronouns need to agree in gender with the relative head in Arabic relative clause. The relative pronouns also need to agree in gender with the relative head and the verb in the case of Arabic subject relative clauses. In regard to this issue, ChatGPT has performed better than Gemini where one issue of this type is annotated in ChatGPT translation outputs as compared to three annotated issues in Gemini translation outputs.

Other issues are spotted. Examples of these issues include the case where the subject is a noun phrase that is headed by an expression of quality like many or more. In addition, nouns that can be both countable and uncountable can be problematic if the gender of the singular noun is different from that of the plural noun. Such issues appear to be problematic for ChatGPT.

4. Discussion and Conclusion

From the results above, it is concluded that Gemini performed better than ChatGPT in almost every aspect when it comes to gender-related issues. Both the number of issues, as well as gender accuracy evaluation, came in favor of Gemini .

It is also clear that the constructed test suite was able to shed a light on the gender-related issues that may occur translation outputs generated by LLMs. The research team came with what can be considered patterns of these issues. However, since the investigation has an exploratory nature, the mentioned patterns are not either inclusive or conclusive. Based on the results of this research, more focused and detailed realistic test suites should be built. That should be followed by extensive evaluation and deliberate training of LLMs.

The results also showed that there are many errors in the context of nouns of non-human referents. This does not come only from the fact

that these nouns in English do not have gender which is not the case in Arabic. It also comes from the complicated gender-related nature of these nouns. These nouns do not have an overt gender marking. It is also known that in Arabic the plural of a singular masculine non-human referent noun can be feminine, and as a subject, it requires a verb with a feminine marking. Therefore, it can be said that ChatGPT has a significant problem in this aspect and it requires more focused training. This also forms a challenge for Gemini, but it seems that the issue is not as severe as it is in ChatGPT .

It should be mentioned that it was noticed that both ChatGPT and Gemini produced translation outputs that showed various other issues. Therefore, a fine-grained analysis is to be performed.

Acknowledgments

Not applicable.

Authors' contributions

The authors have contributed equally to this study.

Funding

This research received grant no. (129/2023) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

Competing interests

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.

Informed consent

Obtained.

Ethics approval

Not applicable

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Data sharing statement

No additional data are available.

References

Alhafni, B., Habash, N., & Bouamor, H. (2020). Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing* (pp. 139-150). Retrieved from https://aclanthology.org/2020.gebnlp-1.12

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 1-61). https://doi.org/10.18653/v1/W19-5301

Costa-jussà, M. R., Escolano, C., Basta, C., Ferrando, J., Batlle, R., & Kharitonova, K. (2022). Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 11855-11863). https://doi.org/10.1609/aaai.v36i11.21442

Currey, A., Nădejde, M., Pappagari, R. R., Mayer, M., Lauly, S., Niu, X., ... Dinu, G. (2022). MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4287-4299). https://doi.org/10.18653/v1/2022.emnlp-main.288

Daems, J., Macken, L., & Vandepitte, S. (2014). On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship. In *9th International Conference on Language Resources and Evaluation (LREC)* (pp. 62-66). European Language Resources Association (ELRA). Retrieved from http://hdl.handle.net/1854/LU-4418636

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., &Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460-1474. https://doi.org/10.1162/tacl_a_00437

Ghosh, S., & Caliskan, A. (2023). ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 901-912). https://doi.org/10.1145/3600211.3604672

Habash, N., Bouamor, H., & Chung, C. (2019). Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (pp. 155-165). https://doi.org/10.18653/v1/W19-3822

Hada, R., Seth, A., Diddee, H., & Bali, K. (2023). " Fifty Shades of Bias": Normative ratings of gender bias in GPT generated English text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.115

Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., & Byrne, B. (2017). Source sentence simplification for statistical machine translation. *Computer Speech & Language*, *45*, 221-235. https://doi.org/10.1016/j.csl.2016.12.001

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567*. https://doi.org/10.48550/arXiv.1803.05567

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... Awadalla, H. H. (2023). How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*. https://doi.org/10.48550/arXiv.2302.09210

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023*). Is ChatGPT a good translator? A preliminary study*. arXiv preprint arXiv:2301.08745, 1(10). https://doi.org/10.48550/arXiv.2301.08745

Khoshafah, F. (2023). ChatGPT For Arabic-English Translation: Evaluating the Accuracy. *ResearchSquare*. https://doi.org/10.21203/rs.3.rs-2814154/v1

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT) (pp. 1-45). Retrieved from https://aclanthology.org/2022.wmt-1.1

Kocmi, T., Limisiewicz, T., & Stanovsky, G. (2020). Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 357-364). Retrieved from https://aclanthology.org/2020.wmt-1.39

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791-4796). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512

Lu, Y., Zeng, J., Zhang, J., Wu, S., & Li, M. (2021). Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1288-1298). https://doi.org/10.18653/v1/2021.acl-long.103

Nagi, K. A. (2023). Arabic and English relative clauses and machine translation challenges. *Journal of Social Studies*, *29*(3), 145-165. https://doi.org/10.20428/jss.v29i3.2180

Nagi, K. A., Alzain, E., & Naji, E. (2024). Informed prompts and improving ChatGPT English to Arabic translation. *Al-Andalus Journal for Humanities & Social Sciences*, *98*(11). https://doi.org/10.35781/1637-000-098-007

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2019). Transforming complex sentences into a semantic hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3415-3427). https://doi.org/10.18653/v1/P19-1333

Popović, M. (2021). On nature and causes of observed MT errors. In Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track) (pp. 163-175). Retrieved from https://aclanthology.org/2021.mtsummit-research.14

Qasmi, N. H., Zia, H. B., Athar, A., & Raza, A. A. (2020). SimplifyUR: unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3484-3489). Retrieved from https://aclanthology.org/2020.lrec-1.428

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, *56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 8-14). https://doi.org/10.18653/v1/N18-2003

Saunders, D., & Byrne, B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7724-7736). https://doi.org/10.18653/v1/2020.acl-main.690

Saunders, D., Sallis, R., & Byrne, B. (2022). First the worst: Finding better gender translations during beam search. In *Findings of the*

Association for Computational Linguistics: ACL 2022 (pp. 3814-3823). https://doi.org/10.18653/v1/2022.findings-acl.301

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics, 9*, 845-874. https://doi.org/10.1162/tacl_a_00401

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature, 507*(7490), 9-9. https://doi.org/10.1038/507009a

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*. https://doi.org/10.48550/arXiv.2004.04696

Siu, S. C. (2023). Chatgpt and GPT-4 for professional translators: Exploring the potential of large language models in translation. *Available at SSRN 4448091*. https://doi.org/10.2139/ssrn.4448091

Stafanovičs, A., Bergmanis, T., & Pinnis, M. (2020). Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 629-638). Retrieved from https://aclanthology.org/2020.wmt-1.73

Štajner, S., & Popović, M. (2018). Improving machine translation of english relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)* (pp. 39-48). https://doi.org/10.18653/v1/W18-7006

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1679-1684). https://doi.org/10.18653/v1/P19-1164

Sulem, E., Abend, O., & Rappoport, A. (2020). Semantic structural decomposition for neural machine translation. In *Proceedings of the ninth joint conference on lexical and computational semantics* (pp. 50-57). https://aclanthology.org/2020.starsem-1.6

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://doi.org/10.48550/arXiv.2312.11805

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 113-123). Retrieved from https://aclanthology.org/W18-6312

Turganbayeva, A., Rakhimova, D., Karyukin, V., Karibayeva, A., & Turarbek, A. (2022). Semantic connections in the complex sentences for post-editing machine translation in the Kazakh language. *Information, 13*(9), 411. https://doi.org/10.3390/info13090411

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3003-3008). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1334

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). "Kelly is a Warm Person, Joseph is a Role Model": Gender biases in LLM-generated reference letters. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.findings-emnlp.243

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 15-20). https://doi.org/10.18653/v1/N18-2003

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., ... Huang, S. (2023). Multilingual machine translation with large language models: Empirical results and analysis. *arXiv e-prints, arXiv-2304*. https://doi.org/10.48550/arXiv.2304.04675