Social Bias in Machine Translation

Table of Contents

1 Introduction

Neural Machine Translation (NMT) has significantly facilitated our daily life and changed the translator's job profile. NMT systems are rich in information because they are trained on huge amounts of data found online, mostly created by society. This data reflects societal thinking and behavioural aspects in society, affecting the choices NMT systems make when translating a text into another language. At times, this can result in skewed or wrong translations, such as the current German Chancellor Olaf Scholz often being translated as a feminine chancellor (DE: "Kanzlerin") since previously, Angela Merkel had been the German Chancellor for 16 years, and from those 16 years, NMT systems learned to translate *chancellor* as "Kanzlerin" (UEPO.de 2022). These kinds of translation gaps are the focus of this paper, which examines typical biases, such as gender, racial or social bias, in machine translation. The aim is to find these biases in NMT translations and to illustrate these in a video so that students of translation studies get a better understanding of typical occurrences of bias in machine translation and how to look out for these.

2 Machine Translation

Key to understanding how NMT systems work is that they are data-driven, meaning that they are trained on large parallel corpora (Vanmassenhove et al. 2021:2204; Rama/Vanmassenhove 2021:68). This data is fed into an NMT model for training and to process this data, NMT systems create numerical representations of the source text that include all its grammatical and semantical information (Peters et al. 2018:1). An exact translation of a sequence in the target language would have a similar representation that includes the same information as the sequence in the source text (McCain et al. 2022:2). For a given source sentence, the model's translation is therefore a result, a most probable output, of what it has learned from its training data (Saunders/Byrne 2020:7724; Bahdanau/Cho/Bengio 2014:2). One aspect included in the numerical representation of language, which we analyse further in this paper, is the information about the gender marked in the words (Gonen/Goldberg 2019:1).

3 Gendered Language

This chapter aims to define the term *gendered language* to better contextualise and understand the present analysis. In general, the term *gender* describes the social distinction of humans present in many societies and which begins with the body (Kotthoff/Nübling 2018:14). It also describes which gender a person feels they belong to, and, in most cases, the gender identity (gender) corresponds to the sex assignment (sexus) made at birth (ibid.). Conventionally, people fall into two categories: male and female. However, current scientifical discussions and investigations refer to the existence of more than these conventional two genders. Traditionally, people behave according to their gender and are even expected to do so. In a broader sense, gender not only refers to a biological distinction but also to culturally and historically evolved ways of how to dress, speak, behave, consume or act (ibid.). This construct manifests itself in society not only through behaviour or attire but also in our language.

Another recent issue closely linked with gendered language is the generic masculine, which is used in many grammatical languages, such as Spanish or German to refer to people across genders (Kotthoff/Nübling 2018:91). Traditionally, the generic masculine, such as the plural German term "Lehrer" for teachers, can refer simply to a group of male teachers or it can refer

to a group of both male and female teachers. This is ambiguous and if the context is unclear, a reader cannot be certain whether only men or both men and women are referred to (Kotthoff/Nübling 2018:96f.). One could also say "Lehrer und Lehrerinnen", explicitly referring to both male and female teachers. In a few cases, there is also a generic feminine (such as "Kosmetikerinnen" for *cosmeticians*), often associated with professions still significantly dominated by women, such as florists or cosmeticians, where men are a minority (Kotthoff/Nübling 2018:121).

In recent years, using the generic masculine has been somewhat controversial because it only implicitly includes all genders. Women or non-binary people are not explicitly referred to, which can indirectly place the focus on men. To be inclusive of all gender groups, new spellings have been adopted in many languages to achieve more equality. Examples in German include the asterisk ("Kanzler*innen"), the underscore ("Kanzler_innen"), the colon ("Kanzler:innen") or writing a capital "I" as a connection ("KanzlerInnen") (Dudenredaktion 2022). These spellings are meant to explicitly refer to all genders. However, there is not yet a "golden standard", meaning that all these spellings can be accepted. It is also worth mentioning that the topic has attracted the attention of many experts in technical communication, as various ideas and approaches are currently being presented and discussed (e.g. Evers et al. (2022) from the perspective of terminology). However, the use of gendered terms remains controversial and a standardised approach is not in sight yet.

Additionally, most of today's training data does not yet contain gendered language. Therefore, two questions for machine translation are: will NMT systems be able to recognise patterns in gendered language and correctly transfer them into the respective target language? And what would be the correct spelling, as this has not yet been agreed on by the public?

4 Social Bias in Machine Translation

Biases are priors that inform our decisions and are not per se negative (Shah/Schwartz/Hovy 2020:5248) but they can be a result of prejudices and stereotypes leading to faulty and hasty judgments and shaping or even distorting individual and social patterns of perception (Wondrak 2014). Biases inherent in our society, such as prejudices, are reflected in the language we use (Czihlarz 2019). When using language technology, i.e. natural language processing (NLP), machines trained on the language we produce, prejudice and bias can even be amplified (Vanmassenhove et al. 2021:2). NMT is no exception, as the engines "learn" to translate by matching and adopting samples from existing translations in the training data. This training data will automatically include certain patterns in language and of society that an NLP model may not be able to generalise to other author-demographics, leading to bias being an inherent property in NLP systems that, if not addressed, can have negative consequences (Shah/Schwartz/Hovy 2020:5248). Bias research in NMT is therefore not only a technical issue but also an ethical one.

4.1 Social Bias

Biases are a result of psychological heuristics, they are mental "shortcuts" that help us react faster in certain situations (Shah/Schwartz/Hovy 2020:5262). In the past, looking at people and judging them based on their appearance, skin colour, gender, age, clothing, language or accent was a survival instinct. Today, however, it's a rather flawed and problematic characteristic as people tend to jump to (sometimes wrong) conclusions based on their judgements. Social biases

are present in the personal characteristics and preferences of individuals, including their beliefs, attitudes, and choices that are shared within the group they belong to.

4.2 Gender Bias in MT

One type of social bias is gender bias or gender-related bias, which is due to formulations, mental assumptions or statistical errors that result in misrepresentations of actual gender relations (Stangl 2022). Gender manifests itself both in the agreement with other words in a sentence and the choice of context-based words or on the level of syntactic constructions, and this information is integrated into NMT systems (Vanmassenhove/Hardmeier/Way 2018:3003f.). NMT systems are trained on unbalanced data that naturally refer to more men than women, leading to translation outputs unknowingly perpetuating social biases (Costa-jussà/de Jorge 2020:26; Saunders/Byrne 2020:7724). Furthermore, NMT systems dominantly translate on a sentence-by-sentence basis, which leads to professions being translated with stereotyped genders (Basta et al. 2020:99). Figure 1 shows how gender stereotypes inherent in society (and therefore in NMT training data) lead to professions being translated predominantly as either male or female.

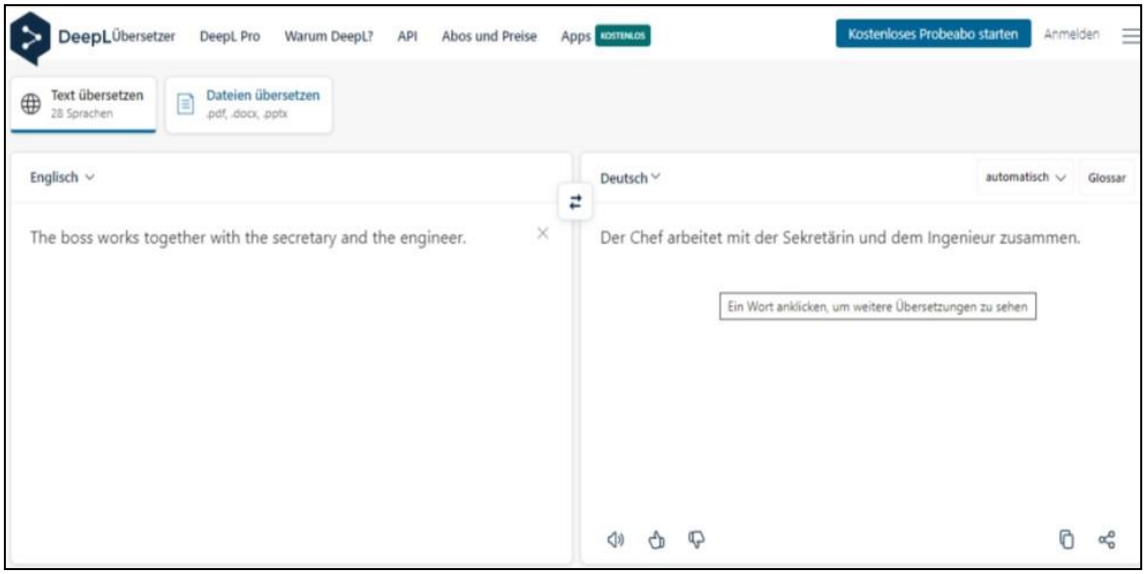

Fig. 1: DeepL translation from English into German (15 November 2022)

Gender is not specified in all languages, but in the occasion that it is specified in both the source and target languages, the gender should ideally be (machine-) translated correctly as expressed in the source text. However, there are genderless languages (e.g. Turkish/Finish) or languages that only display natural but no grammatical gender (e.g. English) (Savoldi et al. 2021:847f.). At the same time, there is a tendency to use gender-neutral vocabularies (e.g. chairperson instead of chairman/-woman). Yet with all this diversity, there is a lack of established standards, both from a monolingual and multilingual perspective. In such gender-ambiguous cases, NMT systems can only assume which gender should be used in the translation. These assumptions are based on statistical likelihood resulting from the data with which the systems were trained (Forcada 2017:295).

The architecture of an NMT system is a fundamental part of how it works. However, the training data is equally important. Usually, NMT systems are trained on substantial amounts of high-quality parallel translation data. Through this training, NMT systems learn to translate data they have not seen before or that stems from another domain (Koehn 2020:37). The better (e.g. larger, of higher quality, more balanced) the training data, the more accurate the translations

will be. But this training data can also be flawed or unbalanced, which can lead to flawed translations. A simplified example: If the training data (English–German) shows that *doctor* was translated more frequently as *Arzt* (male) than as *Ärztin* (female), the system is more likely to use, even overuse due to overamplification, the male noun for its output.

The type of data NMT systems are trained with directly affects the MT output. Therefore, training datasets need to be adapted to help reduce bias in MT (Hovy et al. 2020:1689). There has been a shift in society to create more gender-balanced texts, such as the DeBiasByUs project (Daems/Hackenbuchner 2022), which aims to collect cases of gender-biased machine translations to create a database for research purposes. But as our data analysis in chapter 5 will show, this shift is not yet significantly represented in today's training datasets. Hence, many of the texts include bias, stereotypes and generalisations that are reflected in the translations. For more balanced training data, we have to take a closer look at how representative the data is and adapt it to better represent gender in language.

5 Data Analysis and Methodology

After considering social bias in machine translation theoretically, this section presents the data analysis and methodology on which the findings and results of this paper are based.

Firstly, various texts and sentences containing at least one of the biases explained in previous chapters were written in English. These texts were specifically written to test MT systems and are supposed to provoke gender bias to some extent. For example, sample text 1 uses the gender-neutral pronouns *they* and *their* somewhat interchangeably with the pronouns *she*, *he*, *his* and *her*. Though this is not necessarily grammatically incorrect, this sample text is more inconsistent in its use of pronouns than a text would be if it wasn't written for this purpose. Even though the texts contain different biases, such as social, gender or racial bias, the focus of this paper lies on gender bias.

Subsequently, the English texts were inserted into a commercial NMT system, DeepL or Google Translator, and translated into German. In this way, we can observe how these systems deal with existing bias in the source text and how they translate this bias into the target text. The results were later summarised and audio-visually presented in a video. You can find the texts on the DataLit^{MT} Repository on GitHub and the video on DataLit^{MT}-Pages.

6 Possible Solutions

When looking into the different biases that can arise when translating using an MT system, it becomes clear that an effort should be made to find a solution. There are ways to mitigate social biases while translating but unfortunately, one solution does not fit all. An explanation of why there is no overall solution will be given in this chapter. Section 6.3 presents some information about challenges one can encounter with debiasing.

6.1 User Solutions

Perhaps one of the easiest ways to help prevent bias in MT is to write unbiased source texts or debias them in pre-editing. Examples in English include using plural nouns and pronouns (they/their), replacing (singular) pronouns with generic nouns or removing unnecessary pronouns (ProEdit n.d.). Writing unambiguously can also lead to more bias-free MT outputs.

In practice, however, translators are rarely the authors of their source texts, nor do they have the time and/or permission necessary for pre-editing. And if a source text is written in a (genderless) language such as English, translators might not know or be able to disclose if a person is female, male or non-binary.

6.2 Machine Solutions

Some users, particularly laypersons, of commercial MT systems may not be able to understand the target language or have the time or possibility to review texts (particularly long texts) for gender mistranslations. Ideally, NMT systems should produce unbiased or gender-correct translations, leaving little to no room for the necessity of post-editing gender.

A way to get a step closer to debias an NMT system is to train it on annotated datasets. In gender-annotated datasets, the speaker is tagged with gender information at the start of a sentence (Vanmassenhove et al. 2018:3004).

Furthermore, it might be helpful to use datasets with more female plural terms in order to create a more varied and balanced dataset so that this can be reflected in the translation (Vanmassenhove et al. 2018:3005). Saunders and Byrne, who treated gender bias as a domain adaptation problem, suggest, as they have shown in their study, that using a "small, trusted gender-balanced set" to train an NMT system could reduce gender bias in machine translation as it "allow[s] more efficient and effective gender debiasing than a larger, noisier set" (2020:7724).

If the gender of a person in a source text is ambiguous (i.e. cannot be known, such as in "My neighbour brought me flour."), an alternative is showing the user this ambiguity and letting them choose between different options (different genders), as shown in figure 2 and the video.

In cases of ambiguity, systems like Fairslator (Fairslator 2022) allow the user to manually select the gender of the person, aiming to remove biases. The user/translator can choose between different options and has the opportunity to provide more information about the source text so that the translation result is as bias-free as possible (Měchura 2022). This feature is demonstrated in the video.

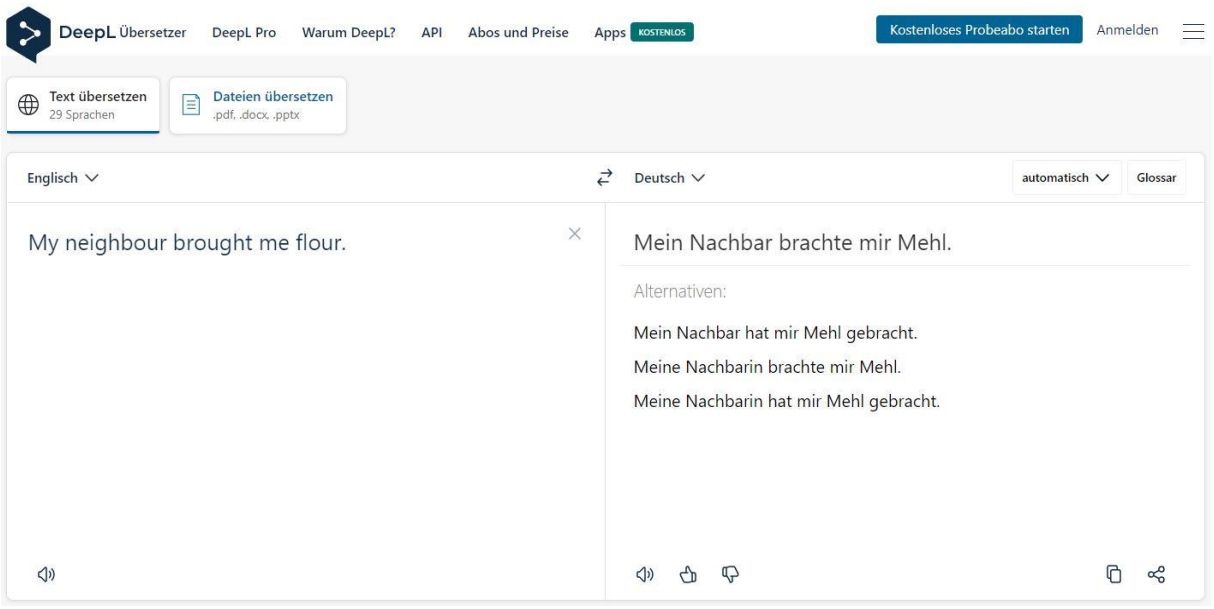

Fig. 2: Example of DeepL showing different translations for an ambiguous source sentence (24 November 2022)

Another pilot project aimed at analysing how susceptible a source text might be to bias, such as gender or racial bias, was developed by Word2Vec (Mouroum 2022). The Word2Vec team developed a (prototype) computational solution to analyse text, such as the sentence:

"As a nurse Peter had to face long working hours, forcing him to quit his job and spend more time with his children at home. Now he is considering becoming a monk or a police officer."

How susceptible certain words in this sentence are to gender and racial bias, can be analysed with Word2Vec (see figure 3).

The potential solutions and examples of how to analyse or disambiguate text from bias or how to better train NMT systems to debias them show that research in the field of gender bias is greatly developing. Savoldi et al. (2021) summarise research conducted on the topic of gender bias and how to reduce it in machine translation and provide a unified framework for further research in this field.

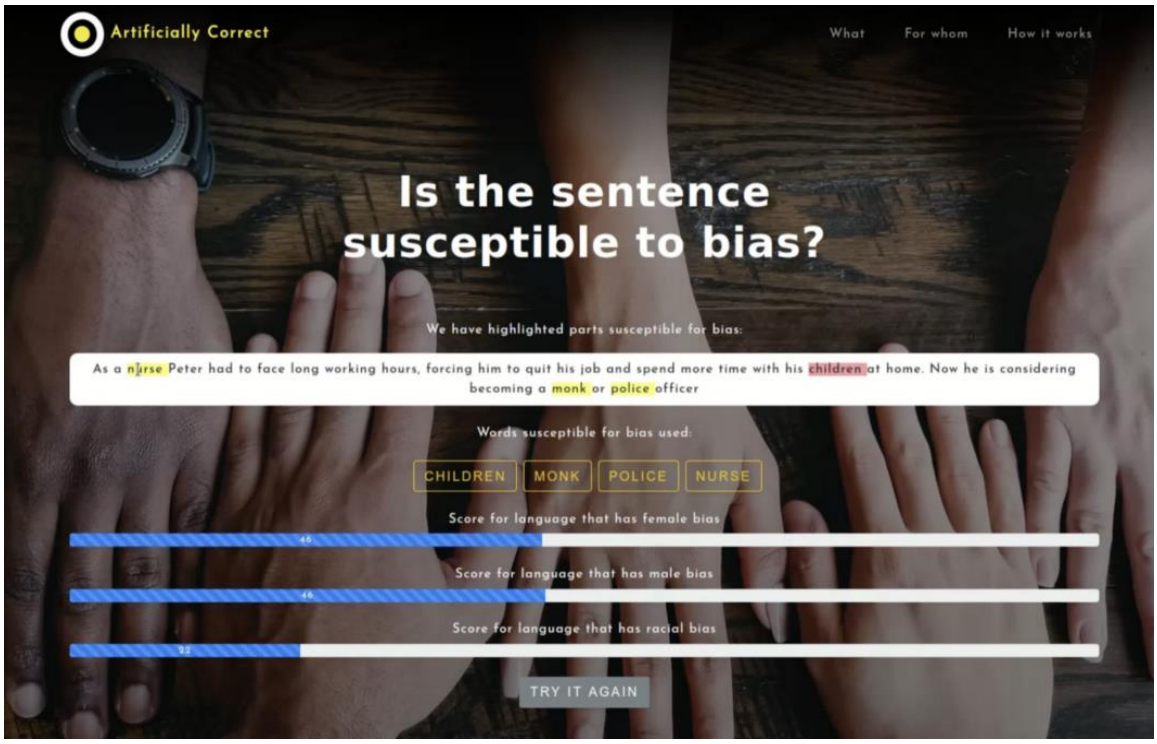

Fig. 3: Word2Vec analysing what words in a sentence are susceptible to bias (Mouroum 2022)

6.3 Issues with Debiasing

Even though it is important to debias machine translation, translators might face several types of issues when it comes to debiasing translations. Thus, debiasing, especially in the case of gendered language, may unnecessarily lengthen the text or even make it less readable for some target groups (see example: target sentence options 1 & 2).

Example:

- Source sentence: „The **receptionists** were kind."
- Target sentence:
  - option 1: "Die **Empfangsmitarbeiter:innen** sind nett."
  - option 2: "Die **Empfangsmitarbeiterinnen und Empfangsmitarbeiter** sind nett."

If the source text is written in simple or easy language and is translated into German while the language in the text is gendered, it fails in its task of being easy and understandable. For this style of writing, it is especially important to write as simply as possible, which is why Maaß (2020), for example, advocates for avoiding gender-sensitive language. Exceptions for gender-sensitive variants can be granted "in contexts where they [d]o not disproportionately burden the sentence and are concordant with the text functionality" (ibid.:247).

However, it can also happen that gender is overemphasised in translations and thus becomes more important than it should be (see example: target sentence option 2). The main message of the source sentence in the example is not *who* is working there but *how* they are working. In the gendered German translation, the emphasis appears to be on *who* is working at the reception.

This issue is linked to overcomplicating language that is meant to be easy to understand. If a text is more precise after translating the source text, errors and misunderstandings can find their way into the text or make the text seem unreadable. Nevertheless, it is important to gender in translations or to remove other biases because it is a crucial part of developing a more gender-equal language and fair society. It can be seen as a learning process, even if it may sometimes seem awkward and cumbersome (Burel 2021).

7 Conclusion

The analysis conducted for this paper was very enlightening, as it showed stereotyping and evident gender bias in machine translation. Possible methods were discussed to reduce gender bias in MT, but these are not yet sufficient. However, both commercially and in research, MT systems are constantly being improved, also to help reduce gender bias. Nevertheless, further research is required to present feasible solutions on how to reduce (gender and racial) bias in machine translation.

References

Bahdanau, Dzmitry/Cho, KyungHyun/Bengio, Yoshua (2014): Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015 Conference. Published in *CoRR: 1–15*.

Basta, Christine/Costa-Jussà, Marta R./Follonosa, José A. R. (2020): Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information. In: Cunha, Rossana; Shaikh, Samira; Varis, Erika; Georgi, Ryan; Tsai, Alicia; Anastasopoulos, Antonios; Raghavi Chandu, Khyathi (eds): *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. Association for Computational Linguistics: 99–102. https://aclanthology.org/2020.winlp-1.25/.

Burel, Simone (2021): Gendergerechte Sprache: weniger Emotionen, mehr Fakten. *congree*. https://www.congree.com/blog/gendergerechte-sprache-weniger-emotionen-mehr-fakten (25 November 2022).

Costa-Jussà, Marta R./de Jorge, Adrià (2020): Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. In: Costa-jussà, Marta R.; Hardmeier, Christian; Radford, Will; Webster, Kellie (eds): *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics: 26–34. https://aclanthology.org/2020.gebnlp-1.3.

Czihlarz, Jochen (2019): Kann gendergerechte Sprache Vorurteile verringern? *ANTI-BIAS*. https://www.anti-bias.eu/anti-bias-strategien/praxisbeispiele/gendergerechte-sprache/ (25 November 2022).

Daems, Joke/Hackenbuchner, Janiça (2022): DeBiasByUs: Raising awareness and creating a database of MT bias. In: Macken, Lieve/Rufener, Andrew/Van den Bogaert, Joachim/Daems, Joke/Tezcan, Arda/Vanroy, Bram/Fonteyne, Margot/Barrault, Loïc/ Costa-jussà, Marta R./Kemp, Ellie/Pilos, Spyridon/Declercq, Christophe/Koponen, Maarit/Forcada, Mikel L./Scarton, Carolina/Moniz, Helena (Eds.): *Proceedings of the 23rd annual conference of the European Association for Machine Translation*. European Association for Machine Translation, 289f. https://biblio.ugent.be/publication/8761022.

Dudenredaktion (2022): Geschlechtergerechter Sprachgebrauch. *Duden*. https://www.duden.de/sprachwissen/sprachratgeber/Geschlechtergerechter-Sprachgebrauch (29 November 2022).

Evers, Elisabeth/Früh, Beate/Schmitz, Klaus-Dirk (2022): Gendern terminologisch betrachtet. In: *technische kommunikation* 44(6), 34–38. https://technischekommunikation.info/ (29 November 2022).

Fairslator (2022): Fairslator.com. https://www.fairslator.com/ (29 November 2022).

Forcada, Mikel L. (2017): Making sense of neural machine translation. In: *Translation Spaces* 6(2), 291–309. https://doi.org/10.1075/ts.6.2.06for.

Gonen, Hila/Goldberg, Yoav (2019): Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv*. https://doi.org/10.48550/arXiv.1903.03862.

Hovy, Dirk/Bianchi, Federico/Fornaciari, Tommaso (2020): "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In: Jurafsky, Dan/Chai, Joyce/Schluter, Natalie/Tetreault, Joel (Eds.): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1686–1690. https://doi.org/10.18653/v1/2020.acl-main.154.

Koehn, Phillip (2020): *Neural Machine Translation*. Cambridge: University Press.

Kotthoff, Helga/Nübling, Damaris (2018): *Genderlinguistik: Eine Einführung in Sprache, Gespräch und Geschlecht*. Tübingen. Narr Francke Attempto.

Maaß, Christiane (2020): *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*. Berlin: Frank & Timme. https://doi.org/10.26530/20.500.12657/42089.

McCain, Miles/Garg, Rhythm/Barakaiev, Igor (2022): Lasts4Ever: Language-Agnostic Semantic Text Similarity for "Every" Language. Stanford CS224N Custom Project. https://web.stanford.edu/class/cs224n/reports/custom_116626702.pdf (29 November 2022).

Měchura, Michal (2022): About this project. *Fairslator*. https://www.fairslator.com (29 November 2022).

Mouroum, Marvin (2022): How we fight bias, sexism and racism with AI… *Medium*. https://medium.com/@mouroum.m/how-we-fight-bias-sexism-and-racism-with-ai-3f228478aba3 (29 November 2022).

Peters, Matthew E./Neumann, Mark/Iyyer, Mohit/Gardner, Matt/Clark, Christopher/Lee, Kenton/Zettlemoyer, Luke (2018): Deep contextualized word representations. *arXiv*. https://doi.org/10.48550/arXiv.1802.05365.

ProEdit (n.d.): Tips for Avoiding Gender in Bias in Writing. *ProEdit*. https://proedit.com/tips-for-avoiding-gender-bias-in-writing/ (29 November 2022).

Rama, Arbnor/Vanmassenhove, Eva (2021): A Comparison of Different NMT Approaches to Low-Resource Dutch-Albanian Machine Translation. In: Ortega, John/Ojha, Atul Kr./Kann, Katharina/Liu, Chao-Hong (Eds.): *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*. n.p.: Association for Machine Translation in the Americas, 68–77. https://aclanthology.org/2021.mtsummit-loresmt.7/.

Saunders, Danielle/Byrne, Bill (2020): Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. *arXiv*. https://doi.org/10.48550/arXiv.2004.04498.

Savoldi, Beatrice/Gaido, Marco/Bentivogli, Luisa/Negri, Matteo/Turchi, Marco (2021): Gender Bias in Machine Translation. In: Roark, Brian/Nenova, Ani (Eds.): *Transactions of the Association for Computational Linguistics* 9. Cambridge, Massachusetts: MIT Press, 845–874. https://doi.org/10.1162/tacl_a_00401.

Shah, Deven/Schwartz, H. Andrew/Hovy, Dirk (2020): Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In: Jurafsky, Dan; Chai, Joyce; Schluter, Natalie; Tetreault, Joel (eds): *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics: 5248–5284. https://aclanthology.org/2020.acl-main.468.

Stangl, Werner (2022): Gender Bias. *Online-Lexikon für Psychologie und Pädagogik*. https://lexikon.stangl.eu/30369/gender-bias (29 November 2022).

UEPO.de (2022): Gender Bias: Google Translate und DeepL übersetzen „el canciller“ mit „die Bundeskanzlerin“. *Uepo.de*. https://uepo.de/2022/01/17/gender-bias-google-translate-und-deepl-uebersetzen-el-canciller-mit-die-bundeskanzlerin/ (29 November 2022).

Vanmassenhove, Eva/Hardmeier, Christian/Way, Andy (2018): Getting Gender Right in Neural Machine Translation. In: Riloff, Ellen/Chiang, David/Hockenmaier, Julia/Tsujii, Jun’ichi (Eds.): *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, 3003–3008. https://doi.org/10.18653/v1/D18-1334.

Vanmassenhove, Eva/Shterionov, Dimitar/Gwilliam, Matthew (2021): Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. *arXiv*. https://doi.org/10.48550/arXiv.2102.00287.

Wondrak, Manfred (2014): Wie lautet die Unconscious Bias Definition? Wie unterscheiden sich Stereotype von Vorurteilen? *ANTI-BIAS*. https://www.anti-bias.eu/wissen/definitionen/unconsciousbias-definition/ (29 November 2022).