Alexandra Butler

# Evaluating Gender Bias in BERTić, a Large Language Model Trained on Croatian Data

## 1 Introduction

The body of research on large language models (LLMs) is being added to every day in the field of computational linguistics. These computational models are typically trained using artificial neural networks and large corpora of text generated by humans from literature and the news, Wikipedia entries, and other crawled web pages. Google's BERT is trained on 3.3 billion tokens, for example (Devlin et al., 2018). The applications of LLMs such as Google's BERT or OpenAI's GPT-3 cover natural language processing (NLP) tasks such as translation, question answering, and text prediction, services employed by millions of users every day. The tasks can be completed quickly and with such accuracy that the models' performance has been described as comparable to or better than that of humans (Brown et al., 2020; Thoppilan et al., 2022). Many researchers have focused on the harmful impact of LLMs, including the preservation and even accentuation of bias present in society (Bender et al., 2021; Weidinger et al., 2021).

Because of the immense quantity of text available online, the internet is the source of training data for most LLMs as they require large amounts of data for optimized performance. Biases held by internet users, including those on user-generated platforms such as Wikipedia, are encoded in the models and perpetuated by them during their output (Dinan et al., 2020). A platform like Reddit is mainly used by young, male-identifying persons (Jurafsky & Martin, 2023), and the views held by that demographic are retained by the model. Among the papers continually being published on artificial intelligence are studies which focus on evaluating and quantifying social bias in LLMs and the impact of bias on their users. Most authors have sought to measure bias in English. Bolukbasi et al. (2016) explored gender bias related to career terms in word embeddings for English with the aim of removing "undesirable" gender stereotypes associated with job roles (*receptionist - female*) while maintaining useful gender associations for NLP (*grandmother - female*). In 2017, Caliskan et al. expanded the scope to include gender and race. They also compared the results from the word embeddings to human judgments about biases in society using the Implicit Association Test (IAT). Caliskan et al. (2017) and later Kurita et al. (2019) relied on human judgments to measure the degree to which biases encoded in language models (LMs) trained on human data reflect biases held by humans. Garg et al.

(2018) used embeddings to map social change relating to attitudes about gender and ethnicity in the US over time.

Bias in other high-resource languages like Hindi or French has also recently been considered. Bhatt et al. (2022) centered the NLP fairness conversation around India-specific languages and cultures. They discussed “axes of disparities” in India related to religion, ability, gender, sexuality, caste, and region. In the case of India, language data has not been collected from every community, particularly marginalized ones, which the authors suspect may lead to an overrepresentation of powerful groups in NLP tools for Indian languages. A LLM is initially pretrained on a broad, general dataset and then fine-tuned to solve a more specific downstream task such as resume filtering or toxicity detection. Certain demographic groups are virtually always underrepresented in the models which leads to the groups’ further exclusion in downstream tasks. For this reason, research on NLP fairness is by and large aimed at removing bias in LMs. While much of the research conducted has been concentrated on English LLMs, the methods for capturing and reducing bias in English models are still developing. The tests for measuring bias in LLMs trained on low-resource languages are lagging then as well.

### 1.2 Objectives

Current work on the topic of bias in LLMs has mainly studied models and applications for English. Not much has been published regarding languages with smaller speech communities. Certain biases are limited to a small handful of languages or are more pronounced in some communities than others, as with the “axes of disparity” in India (Bhatt et al., 2022), necessitating research into unique cases. This paper evaluates bias in BERTić, a BERT-like LLM trained on Croatian in addition to other, mostly mutually-intelligible, South Slavic languages (Ljubešić & Lauc, 2021). BERTić has been trained on 8 billion tokens of Bosnian, Croatian, Serbian, and Montenegrin text. Compared to English, South Slavic languages are morphologically complex and low-resource in terms of training data which poses complex challenges in language modeling.

In this paper, gender bias in BERTić was characterized and measured for Croatian, a language spoken by 5-7 million speakers. Studying and improving LLMs for non-English languages would result in improvements in NLP tasks, allowing a larger number of speakers to benefit from reliable technology that has existed for English speakers for a number of years. Moreover, having high-performing computational models of other, lower-resource languages

would improve machine learning capabilities and provide better insight into the languages themselves. In addition to exploring fairness in BERTić, this paper also weighs the value of removing or maintaining bias in models fine-tuned for literary machine translation (MT), a topic which has not been covered to the extent of other NLP applications. The research questions posed in this paper are the following:

1. How should social biases be quantified in contextualized word embeddings?
2. Should these methods differ across language-specific models (English vs. Croatian)? If so, how?
3. Should gender bias be reduced in the case of literary machine translation?

### 1.3 Overview of methods

Gender bias was calculated across three contextual word embedding models, BERT, multilingual BERT (mBERT), and BERTić. All three models are masked language models, meaning that they are trained to assign a probability to tokens and predict a masked token given a certain context. To measure bias, the template-based approach proposed by Kurita et al. (2019) was implemented. First, English template sentences such as, “[MASK] is a programmer,” where the masked token is either *he* or *she* were inputted and the probability BERT assigned to either pronoun in that sentence calculated. Then the probability of either pronoun appearing was re-weighted using the prior bias of the model with a template sentence like, “he is a [MASK]”. The gender bias in BERT for *programmer* was finally determined by looking at the different probabilities BERT assigned to *he* or *she* appearing with *programmer*. The same procedure was conducted for mBERT using template sentences in both English and Croatian, and for BERTić in Croatian. The Croatian sentences were first directly translated, then adapted as Croatian is a highly inflected language featuring case and gender. The sentences were modified for fluency and so as not to reveal gender on the career word.

### 1.4 Contributions

This is one of the few papers to consider the impact of gender bias in a LLM trained on South Slavic language data by adapting a validated procedure for measuring bias in BERT, an English-language model. BERT has been used extensively to test language modeling performance benchmarks and has been adopted for many different high-resource languages, but little work has been done on LLMs trained on low-resource languages. The means of quantifying bias generally is discussed, as well as the means of quantifying gender bias

specifically in BERTić for Croatian, a low-resource and morphologically complex language. Gender bias is a common metric for fairness in LLMs as it exists across most cultural and linguistic contexts. This work may be later adapted to measure other kinds of bias.

Moreover, the goal of most research on social bias in LLMs is to reduce bias or remove it completely. Harmful forms of bias are attested in LLMs which result in the exclusion of certain groups from the model in its training or discrimination against them in its output. This is most evident in the implementation of downstream tasks such as question answering or sentiment analysis. Little work has discussed the potential value of maintaining social biases in LLMs. In the case of MT of literature, a user of the model may wish to maintain the intention of the author and the values of the time in which the work is produced. The model would not then be debiased as that information would be lost. It is currently unclear what the impact of keeping bias in such a model would look like as few papers have explored the topic. This paper considers the means of capturing bias and the possible effects of preserving bias in a model fine-tuned for literary MT.

## 2 Previous work

During the pretraining stage of a LM, it processes large quantities of text enabling it to learn the distribution of individual words in the text relative to others, including the syntactic environment a word appears in and the meaning of adjacent words (Mikolov et al., 2013). Word embeddings used for NLP are learned representations of words acquired from language models based on their distribution. The embeddings are mapped onto vectors, multi-dimensional spaces which place a word (*dog*) near other words of similar semantic meaning (*pet*) and far from words with which it does not share meaning (*lamp*). In vector semantics, word embeddings are represented by vectors which display a word as a point in a multi-dimensional space developed from the word's distribution.

Word embeddings may be static, or uncontextualized, and contextualized. Word2vec is a group of models capable of creating dense vectors made up of thousands of dimensions which optimize NLP tasks (Mikolov et al., 2013). Its embeddings are uncontextualized, only learning a static embedding for each word. GloVe is another such algorithm which produces static embeddings (Pennington et al., 2014). Word2vec can do this via the skip-gram method which assigns more weight to nearby words than distant ones (Mikolov et al., 2013). For each word, it learns two embeddings: one for the word as the target and one for the word as context for

another target. The second method is bag-of-words which instead predicts the target word from the ones surrounding it. Embeddings derived from word2vec can be used to measure semantic similarity between words or texts using the cosine that compares two vectors with the same number of dimensions.

More recently, the transformer architecture was introduced for pretraining models. Instead of only being able to look at a target word and words close by, a transformer is capable of encoding the distant context of a word from the entire input, allowing for contextualized word embeddings (Vaswani et al., 2017). The architecture is useful in downstream tasks like MT and summarization. Transformers may be left-to-right or bidirectional, meaning that they learn representations from the input from left-to-right and right-to-left. BERT (Bidirectional Encoder Representations from Transformers) is a collection of LMs introduced by Google created using the transformer architecture (Devlin et al., 2018). Many other models have descended from BERT, including mBERT and BERTić, a LLM trained on South Slavic data and the focus of this paper. BERT is known as a masked language model, trained to complete a fill-in-the-blank task. Given a sentence with a masked element, BERT computes the probability of a certain token appearing in its output. This allows for BERT to represent the meaning of a word differently each time it appears in a new context, making the word embeddings contextualized. Contextualized embeddings represent word tokens, or a certain word type in a certain context. Static embeddings like those derived from word2vec represent only word types.

### 2.1 Reducing gender bias in large language models

Word embeddings used to cluster related words such as those obtained from word2vec or BERT reveal gender biases in the models which has been a concern for researchers because of their increasing usage. In a paper on word2vec and GloVe embeddings, Bolukbasi et al. (2016) sought to remove undesirable gender associations (*receptionist - female*) from the embeddings while maintaining useful ones for NLP (*queen - female*). The static embeddings were given an analogy task to solve like, “Man is to king as woman is to *x*.” Some of the analogies inputted reflected associations deemed neutral by the authors (*brother - male*). Other analogies revealed gender bias the authors viewed as potentially harmful in downstream tasks (*petite - female; lanky - male*). To debias gender in the embeddings, the authors first quantified bias using the cosine, and then equalized inherently gender-neutral terms that are socially associated with a gender (*softball - female*).

The authors described some nouns and pronouns in English as being morphologically marked for gender (*actress)* which affects bias in the model. Language modeling in Croatian is complicated by the fact that in addition to many nouns and pronouns, verbs are also marked for gender. In English, there are more words to refer to male persons than female, and more words which refer to female persons than male in a way that sexualizes them (Stanley, 1977), which also comes through in word2vec and GloVe (Bolukbasi et al., 2016). The word embeddings produced by the models were compared to human responses crowdsourced on Amazon Mechanical Turk. The participants answered questions about analogies which in their opinion reflected social biases regarding gender and careers but which they themselves did not necessarily endorse. These human judgments were in line with the output of the word embeddings, demonstrating that the model did reflect social biases. Bolukbasi et al. (2016) referred to the phenomenon of stereotypically feminine traits being viewed more positively (*helpful, petite, charming)* as “benevolent sexism” and not a feature that the authors argued should be maintained in language modeling.

However, the distinction between desirable and undesirable bias is not made clear in the paper. Some biases are of course undesirable. Algorithms used to predict repeat criminal offenders have been proven to show racial bias (Bolukbasi et al., 2016; ProPublica, 2016), partly as a result of language data not being collected for certain demographic groups. If a group adopts a non-standard language variety like African American Vernacular English, it may not be represented in NLP tools. In the case of gender bias, an association like *grandmother - female* is helpful in implementing a downstream task like family-tree generation but the discussion of where the training data comes from and which stereotypes should be maintained and which discarded would need to be addressed before fine-tuning any LM. NLP is widely used by companies to filter large numbers of job applicants for the sake of efficiency. Therefore, an issue requiring immediate attention according to the authors are those embeddings revealed in gendered occupation analogies (*housewife, nurse - female; computer programmer, doctor - male*). Such word embeddings could be used in an employer’s web search of computer programmers. Because the role is male-dominated, the search will be more likely to pull up male programmers.

BERT and other contextual word embeddings are currently most often relied upon for various NLP tasks. Unlike word2vec which produces static embeddings whose biases may be analyzed using an analogy task and cosine calculation, BERT produces contextualized embeddings requiring a different approach. In a paper by Kurita et al. (2019) whose

methodology this paper draws on, a template-based approach for quantifying bias was proposed. The authors' aim was to evaluate gender and racial bias in BERT using the template-based method and to apply it by measuring gender bias in a downstream Gender Pronoun Resolution (GPR) task. Every word in contextualized embeddings has a different embedding so the analogy task is not applicable.

In the proposed procedure, a template sentence is first created containing a career-related attribute word (*programmer)* for which bias is calculated, and a gendered target for the bias (*she*). The target is then masked and the probability that BERT assigns to the attribute appearing in the template sentence ("she is a programmer") is computed. The likelihood of the target appearing is re-weighted using the prior bias of the model toward the target *he*. Finally, the difference between the normalized predictions for each target word can be used to measure bias for a career-related attribute word. In a highly inflected language like Croatian, *programmer* is marked for gender (masc. *programer* vs. fem. *programerica*), nullifying any attempt at calculating gender bias. The template sentences would need to be adjusted to measure bias in a model trained on Croatian (1).

(1) *Ona* *se* *bavi* *programir-anjem.*
she.NOM refl deal.PRES.3SG programming-INS
"She works in programming."

To see how well the template-based approach quantified bias in BERT, Kurita et al. (2019) applied the method to a GPR task. GPR can be used for coreference resolution where a pronoun-containing expression is paired with a referring expression. Typically, these solutions favor male entities. The task was to classify an ambiguous pronoun as either referring to a male entity, female entity, or neither. The results showed that a greater number of female pronouns were predicted to refer to no entity, and that the model would not consistently perform coreference resolution when the pronoun was feminine, especially if the topic of the text was more masculine.

In order to measure gender bias in BERT embeddings for professions, the authors used a publicly available employee salary dataset for a county in Maryland and the template sentence, "TARGET is ATTRIBUTE". They also used a dataset containing positive and negative traits and the template sentence, "TARGET is [trait from positive/negative dataset]". For bias in embeddings for professional skills, the authors relied on a dataset containing technological skills of Amazon employees. The template sentence for the task read, "TARGET can do [skill from

technological skills]". The findings demonstrated that BERT strongly prefers male pronouns, not ideal for a task like resume filtering. This was not mentioned in the paper, but it appears that aside from positive traits being more frequently associated with masculine pronouns, negative traits were as well, which may indicate that feminine pronouns were generally ignored by the model.

In addition to displaying job titles and their corresponding salaries, the Maryland dataset also revealed the first names of employees like *Sonya* or *Richard* which may have affected the results in the paper. Moreover, the Amazon dataset of technological skills included descriptions of job responsibilities which contained pronouns. In a manual search of the dataset, the pronouns *his* and *he* (usually in reference to Jeff Bezos' letters to shareholders) appear at times, and more often than *she*. The authors here were similarly concerned about the implications of using BERT word embeddings in downstream tasks. 75% of US employers use social media for recruiting and many applications are filtered using AI (Fuller et al., 2021). This is an issue if certain professions, like computer programming, are more associated with male applicants as the model could be selecting male applicants as the superior option if they are overrepresented in the training data.

### 2.2 Quantifying bias in language models trained on under-studied languages

Research on bias in NLP has spotlighted the Western context with its social biases and languages. Not much work has looked at bias in non-English languages. One study by Bhatt et al. (2022) explored NLP tooling for Indian languages. Some prior work had been done on fairness in NLP models for a few of India's widely-spoken languages but India is a country of great linguistic, religious, and cultural diversity presenting challenges in language modeling. The authors sought to a) explore bias in BERT and MuRIL, a BERT-like model trained on data from India's languages, b) create resources for measuring bias in these models, and c) show that bias is maintained in the models. Specifically, the paper investigated social biases along the "axes of disparity" in India, involving region (ethnicity), caste, gender, sexuality, religion, and ability. A dataset of identity terms for religion, caste, region etc. and single-word stereotypes associated with that identity was created. The identity terms were used in a perturbation sensitivity analysis which revealed bias by replacing terms of the same semantic category in sentences such as, "IDENTITY TERM people love food". Then, sentiment scores were recorded for each sentence when the term was switched out.

To calculate gender bias, the authors used a publicly available list of 300 Indian names. The findings of the experiments demonstrated that both BERT and MuRIL learned the gender associations of the names. An issue which was again raised was that some groups were underrepresented in the training data. Not all groups in India have equal access to education resulting in a lower literacy rate among some of them. For example, people belonging to lower castes in society are generally less-educated than people of other castes (Bhatt et al., 2022). They would then be less likely to contribute to training data to the extent of others. Women were also shown to use the internet less than men, in turn causing the views of male users to be overrepresented in NLP. The hope of the authors was for their work to be generalized to other under-studied, non-English contexts.

Like BERT and MuRIL, BERTić is a transformer language model (TLM). It is trained on 8.4 billion tokens from Bosnian, Croatian, Serbian, and Montenegrin (Ljubešić & Lauc, 2021). The motivation behind the creation of the model was to develop NLP tools for low-resource languages. The authors implemented the ELECTRA model approach to TLM of training a smaller generator model and a larger discriminator model. They argued that this type of model may be more efficient than BERT when it comes to masked LMs. The task given to an ELECTRA model is to discriminate whether a word is from the text or generated by the generator model (Clark et al., 2020). Ljubešić and Lauc (2021) evaluated the performance of BERTić using two standard token classification tasks, morphosyntactic tagging and named entity recognition (NER), along with two sequence classification tasks, geolocation prediction and commonsense causative reasoning. Across all of the metrics, BERTić outperformed the other state-of-the-art TLMs mBERT and CroSloEngual BERT trained on Croatian, Slovenian, and English. Because the model meets the NLP baselines, it may be used for experiments which quantify gender bias in its word embeddings. The results from such experiments could help improve the model and provide the Croatian speech community with better NLP tools.

### 2.3 Preserving bias in models fine-tuned for literary machine translation

Most researchers on fairness have directed their attention to downstream applications for NLP like summarization and question retrieval (Bolukbasi et al., 2016; Kurita et al., 2019) as preservation of social biases could cause real-world harm for large groups of people. In 2014, Amazon launched an automated hiring tool designed to filter job applications and select the most well-suited candidates for further human review. It was later revealed that the language model used to create the tool was trained on the company’s hiring data from the previous

decade. The model encoded patterns in the data, including gender. As a result, it was more likely to output male-identifying applicants over others for typically male roles (*software developer*), regardless of their qualifications (Reuters, 2018). Amazon has since abandoned its hiring algorithm. While LLMs can be a powerful tool for automatization, they pick up undesirable trends in human behavior which have repercussions if implemented carelessly. Tasks like resume filtering which have immediate consequences have then been the focus of most of the research examining fairness in NLP. Not as much work has investigated other applications for language models which require some consideration in terms of the social biases they capture and propagate.

Thai et al. (2022) argued for the retention of social bias in models fine-tuned for literary MT. While most researchers have understandably looked at the impact of bias in LLMs on social media and in the job market, not many have considered the impact on the arts and humanities. Literary translation is a field which would benefit from the automatization of MT. Often performed very slowly due to a small number of qualified translators, the translation of literature could be improved with MT which is efficient and capable of creating tools to help train human translators (Voigt & Jurafsky, 2012; Omar & Gomaa, 2020). Though MT models have been known to produce overly literal translations, stylistic inconsistencies, discourse-level mistakes (coreference and pronoun inconsistencies), and readability issues, the quality of MT has improved with time. However, MT has not been used much for literary translation. Part of this is because the translation procedure would need to change. Typically in MT, a model will translate a document sentence-by-sentence which works for parliamentary proceedings, for example. In literary translations, translators make greater changes to the whole document for the sake of fidelity to the original text, readability, and cultural transmission (Taivalkoski-Shilov, 2019). At the level of discourse, literature is often much more complex than training manuals or legal documents that are typically the subject of MT (Thai et al., 2022).

The goals of Thai et al. (2022) were to demonstrate that state-of-the-art MT models and evaluation metrics fail when it comes to literary MT, and to improve MT by using a model trained on a corpus they created. The PAR3 (Parallel Paragraph-Level Paraphrases) is a corpus of novels originally written in 19 different non-English languages. Each paragraph is aligned with a human translation, as well as a translation produced by Google Translate. The standard BLEU, BLEURT, and BLONDE metrics for MT cannot distinguish between text outputted by Google Translate and text produced by humans, indicating they are not suitable for literary MT. This is in part because they are designed for sentence-level alignment, but human translators often combine sentences in literary translations. As part of an experiment, a group of translators were

given paragraphs in German, French, or Russian and their English translation. Some of the texts were translated by Google Translate and others by human translators. When asked to choose the better translation, the translators chose the human translation 84% of the time. Next, monolingual English experts were given just the English translation of the texts. They preferred the human translations 85% of the time.

To improve the translations outputted by Google Translate, GPT-3, a new and powerful transformer model, was fine-tuned on a post-editing task in which the model adjusted machine-translated text to a human-translated reference, correcting errors made by the model. The findings of the experiment showed that the translators preferred the translations outputted by the post-editing GPT-3 model over those outputted by Google Translate 69% of the time despite the post-editing model omitting details and making unusual stylistic changes. While the authors conceded that the translators differ in their years of translation experience, they do not provide further information about the language profiles of any of the participants. It would be important to note whether they had completed coursework in linguistics or had writing experience, for example. Because the quality of MT for literature is in question, the tests to capture social bias in such models are still developing. It is also uncertain what the purpose is of debiasing models for literary MT.

## 3 Grammatical gender in Croatian

Before outlining the experiments for this project to determine gender bias in LLMs, an explanation of grammatical gender in Croatian must be provided. English features person and number agreement which must be considered while modeling the language (Bock & Eberhard, 1993). Kurita et al. (2019) suggested using template sentences such as “[MASK] is interested in [MASK]” and “[MASK] are interested in [MASK]” to account for different forms of number and person agreement. In addition, if a masculine target word in the singular such as *he* is used to replace one of the masks, then a plural form like *men* is also recommended, and is selected from a list of words in the plural depending on which sentence construction (“are interested” vs. “is interested”) is inputted.

Like English, Croatian and other South Slavic languages feature person and number agreement. In addition, they possess gender and case agreement which results in further complications when modeling the languages. Croatian has six (Arsenijević, 2021) or seven grammatical cases (Vučković, 2004; Hrkač, 2017), and three genders: masculine, feminine, and neuter. Nouns in Croatian are typically grouped into four declension classes based on endings

for the nominative and genitive cases in the singular form as seen in Table 1. Animacy and morphological gender affect which ending a noun takes.

| | **Class I <∅, a>** | | **Class II <o/e, a>** | | **Class III <a, e>** | | **Class IV <∅, i>** | |
|---|---|---|---|---|---|---|---|---|
| | **sg** | **pl** | **sg** | **pl** | **sg** | **pl** | **sg** | **pl** |
| **NOM** | mrav-∅ “ant” | mrav-i | mor-e “sea” | mor-a | rib-a “fish” | rib-e | noć-∅ “night” | noć-i |
| **GEN** | mrav-a | mrav-a | mor-a | mor-a | rib-e | rib-a | noć-i | noć-i |
| **DAT** | mrav-u | mrav-ima | mor-u | mor-ima | rib-i | rib-ama | noć-i | noć-ima |
| **ACC** | mrav-a | mrav-e | mor-e | mor-a | rib-u | rib-e | noć-∅ | noć-i |
| **INS** | mrav-em | mrav-ima | mor-em | mor-ima | rib-om | rib-ama | noć-ju | noć-ima |
| **LOC** | mrav-u | mrav-ima | mor-u | mor-ima | rib-i | rib-ama | noć-i | noć-ima |

*Table 1. Croatian noun declension classes (Arsenijević, 2021).*

Some research suggests that the more highly inflected a language is, the more difficult it is to model (Park et al., 2017; Heitmeier et al., 2021). In Park et al. (2017), a survey of the literature found that the two most widely modeled languages are English and Chinese, neither of which is morphologically complex relative to other languages (Ku & Anderson, 2003), so questions regarding modeling morphology have not been addressed at length by researchers. Park et al. compared translations of the Bible in 92 different languages and discovered that the measure of surprisal, or processing difficulty (Attneave, 1959), is greater when the model is trained on byte pair encoding segmentation, showing that the morphology of a language may be a hurdle in computational modeling. To adapt the English-language template sentence approach for quantifying bias in a model, the morphological gender of Croatian must be considered, especially as gender agreement is marked on nouns, pronouns, adjectives, and (in some forms) on verbs as seen in example (2) taken from Tomić (2006).

(2) a. *Ova* *kuća* *se* *dugo* *gradi-la.*
this.F.SG house refl long.Adv build-PST.F.3SG
“This house was being built for a long time.”

b. *Uda-la* *mu* *se* *kćer.*
marry-PST.F.3SG 3SG.M.Dat refl daughter

“His daughter got married.”

In (2)b., the verb agreeing with *kćer “*daughter” is marked for gender. In (2)a., the demonstrative and the verb are also marked for gender because the noun *kuća* “house” is feminine.

## 4 Methods

### 4.1 Experiment I: Quantifying gender bias using fixed word lists

Gender bias was measured across three contextual word embeddings models and two languages, BERT for English, mBERT for English and Croatian, and BERTić for Croatian. This was done to compare gender bias in the models and determine the best practices for adjusting previously-used template sentences for measuring bias in English-language word embeddings (Kurita et al., 2019) for Croatian word embeddings. All of the code for this project was written in Google Colaboratory notebooks using the Python programming language. To measure bias in BERT and the other BERT-like models which are all masked language models trained to compute the probability of a masked token appearing in a given context, gender, career, and family words were selected to compare how likely the models were to predict either male or female-related words in the context of career or family-related ones. Specifically, the likelihood of a target word (*he, she*) and attribute word (*programmer*) appearing in a template sentence (“TARGET is an ATTRIBUTE”) was calculated. The general methods are as follows:

1. Replicate the template-based approach in Kurita et al. (2019) to quantify gender bias in BERT and mBERT for English.
2. Quantify gender bias in mBERT and BERTić for Croatian using the same methods.
3. Adjust the template sentences to account for case and gender in the Croatian models.

Template sentences such as, “[MASK] is a doctor,” where the masked token is either a male or female-related target word were inputted and the probability that the models assigned to either gender in that sentence was computed. Next, the probability of words relating to either gender appearing in a given sentence was re-weighted using the prior bias of the models with a template sentence like, “he is a [MASK]”. The gender bias in the models for *doctor* was then calculated by comparing the difference in probabilities that they assigned to *he* or *she* occurring with *doctor*.

To measure bias in BERT and mBERT for English, the template sentences, “[MASK] likes/like [MASK]”, and “[MASK] is/are interested in [MASK]” were inputted, along with a word list related to gender, career, and family shown in Table 2.

| family | career | female sg. | male sg. | female pl. | male pl. |
|---|---|---|---|---|---|
| home | boss | she | he | women | men |
| marriage | company | wife | son | wives | fathers |
| family | salary | mother | father | | |
| house | position | mom | dad | | |
| kinship | work | sister | brother | | |
| clan | administrator | | | | |
| kin | medicine | | | | |
| wedding | law | | | | |

*Table 2. English target and attribute words.*

Plural and singular forms of two of the target words were included to match the grammatical number in the template sentences. The word list proposed in Kurita et al. (2019) is somewhat different than the one displayed in Table 2 as most of the singular male and female words are English men’s and women’s first names such as *Amy* or *John*. This list was also inputted for the purposes of this project and the results recorded. However, because the Croatian BERTić model did not have women’s names in its vocabulary, the uniform word list for both languages was changed to include only words which appear in the vocabulary of all three models for the sake of consistency. The same steps were repeated to measure bias in mBERT for English.

The template sentences were translated directly into Croatian to measure bias in BERTić and the Croatian version of mBERT as, ““[MASK] voli/e [MASK]”, and “[MASK] se zanima/ju za [MASK]”. The Croatian word list is shown in Table 3.

| family | career | female sg. | male sg. | female pl. | male pl. |
|---|---|---|---|---|---|
| dom “home” | sef "boss" | ona "she" | on "he" | zene "women" | tate "dads" |

| | | | | | |
|---|---|---|---|---|---|
| brak “marriage” | firma "company" | zena "wife" | sin "son" | on-e "they-F" | on-i "they-M" |
| obitelj “family” | placa "salary" | majka "mother" | otac "father" | | |
| kuca “house” | polozaj "position" | mama "mom" | tata "dad" | | |
| rod "kinship" | rad "work" | sestra "sister" | brat "brother" | | |
| klan "clan" | administrator | | | | |
| porodica "kin”, “family" | medicina "medicine" | | | | |
| pir "wedding" | pravo "law" | | | | |

*Table 3. Croatian target and attribute words.*

An effort was made to keep the word lists identical. Some of the entries needed to be changed for Croatian as not all of the same words were present in the Croatian models’ vocabularies as in the English ones. Nevertheless, the general categories (*career*, *family*) stayed the same. The words do not include special characters (*ć, š, ž*) because the models had difficulty processing them during input. All of the words in the Croatian list are in the nominative case which would result in ungrammatical sentences as in (3).

(3) a. **Ona se zanima za medicin-a.*
she.NOM refl interest.PRES.3SG in medicine-NOM
“She is interested in medicine.”

b. **On voli porodic-a.*
he.NOM love.PRES.3SG family-NOM
“He likes family.”

These words were included anyway to evaluate how the morphology of the language may necessitate altering the test. Some of the words which would be gender neutral in English but need to be gendered in Croatian were only inputted in their masculine form, again resulting in ungrammatical sentences as in (4).

(4) a. **Ona je šef.*

she.NOM COPL boss.M.NOM

“She is a boss.”

b. **Ona je administrator.*

she.NOM COPL administrator.M.NOM

“She is an administrator.”

**4.2 Experiment II: Quantifying gender bias using adjusted template sentences**

The same process was implemented again for the Croatian-language models, but using altered template sentences without fixed word lists. Croatian contains six or seven morphological cases and is marked for gender, making it difficult to model the language by applying the same methods used for English. For example, the suffix in the attribute word meaning *professor* would be marked for gender and appear as either *profesor* (masculine) or *profesorica* (feminine). In a template sentence such as, “[MASK] is a professor”, there is gender agreement between the pronoun and noun when translated to Croatian as in (5).

(5) *Ona je profesor-ica.*

she.NOM COPL professor-NOM.F

“She is a professor.”

Though the models have been trained on Croatian data, many attribute words only appear in the nominative form (*dizajn* "design"), or are reduced to their lemma. Some common words such as *ured* "office" and *radnik* "worker" do not exist in the BERTić vocabulary altogether. After directly translating the sentences into Croatian and testing how the models processed them, the sentences were adjusted so as to account for case and also to omit morphological gender as shown in (6).

(6) a. *On se bavi programir-anjem.*

he.NOM refl deal.PRS.3SG programming-INS

“He works in programming.”

b. *Ona voli djec-u.*

she.NOM like.PRS.3SG child.PL-ACC

“She likes children.”

Because the aim is to measure gender bias, the assumption made is that morphological gender may impact the results. The same sentences which omitted morphological gender in Croatian were translated and inputted into the English-language models to see if and how the bias scores changed. P-values from permutation tests and their effect sizes were calculated for the fixed word lists, as were probability bias scores for the adapted template sentences as proposed in Kurita et al. (2019).

## 5 Results

### 5.1 Experiment I: P-values from permutation tests and their effect sizes

Permutation tests were implemented for all three models to calculate the difference in means between the models predicting either female or male-related target words given a template sentence in English or Croatian. In this kind of statistical test, the data for two groups are added randomly, and then split into two new groups. Next, the mean difference between the new groups is computed many times over to create a distribution of mean differences. A p-value is assigned to the proportion that the new mean difference between the groups is different from the original mean difference. For the purposes of this project, a p-value smaller than 0.01 was selected as indicating statistical significance. If the p-value were greater than 0.01, it may be concluded that there is no statistically significant difference between the two group means. This would suggest that a model would not be more likely to predict a male or female target word.By contrast, a p-value of less than 0.01 would indicate a difference between targets, indicating bias in the model. The p-values for the permutation tests are shown in Table 4.

| | English | Croatian |
|---|---|---|
| BERT | 0.18 | -- |
| mBERT | 0.1 | 0.48 |
| BERTić | -- | 0.23 |

*Table 4. Permutation test p-values.*

None of the p-values fall below 0.01, suggesting that none of the models revealed gender bias. Overall, the result is unexpected based on prior work which shows that most LLMs, and certainly English-language ones, have gender bias encoded in them (Bolukbasi et al., 2016;

Garg et al., 2018; Bhatt et al., 2022; Ulčar et al., 2022). It is expected however, that the Croatian-language models would be less likely to reveal gender bias given the limited number of tokens on which the models were trained. Effect sizes are more relevant in the event that a p-value falls below 0.01. They are shown in Table 5.

| | English | Croatian |
|---|---|---|
| BERT | 1.331 | -- |
| mBERT | 1.526 | 0.3526 |
| BERTić | -- | 0.6257 |

*Table 5. Effect sizes.*

According to Leppink et al. (2016), a value of 0.2 would be a small effect size, and 0.8 would be large. As stated above, because none of the p-values from the permutation tests indicated a statistically significant difference between the models having a bias towards one or the other gender, the effect sizes are not so important here.

### 5.2 Experiment II: Probability bias scores for adjusted template sentences

Kurita et al. (2019) proposed a method of calculating the probability bias score as the difference between two target words (*he, she*) being predicted by the model in a given context. A set of template sentences were inputted first in the Croatian-language models which were grammatical, accounted for case, and did not include attribute words which were marked for morphological gender so as to avoid confounding gender bias results. Next, the sentences were translated into English and inputted into the English-language models. They are shown in (7).

(7) a. *On/a* *se* *bavi* *programir-anjem.*
he/she.NOM refl deal.PRS.3SG programming-INS
“He/she works in programming.”

b. *On/a* *se* *bavi* *medicin-om.*
he/she.NOM refl deal.PRS.3SG medicine-INS
“He/she works in medicine.”

c. *On/a* *se* *bavi* *prav-om.*

he/she.NOM refl deal.PRS.3SG law-INS

“He/she works in law.”

d. *On/a* *voli* *djec-u.*

he/she.NOM like.PRS.3SG child.PL-ACC

“He/she likes children.”

e. *On/a* *voli* *obitelj.*

he/she.NOM like.PRS.3SG family.SG.ACC

“He/she likes family.”

The preliminary results for the adjusted template sentences are more promising. They are shown in Table 6.

| | BERT | mBERT (En) | mBERT (Cro) | BERTić |
|---|---|---|---|---|
| He/she works in programming. | 3.74 | 2.32 | 8.03 | 0.17 |
| He/she works in law. | 4.72 | 1.23 | 40.5 | 0.09 |
| He/she works in medicine. | 4.29 | 1.7 | 15 | 0.077 |
| He/she likes children. | 1 | 1.25 | 1.93 | 0.21 |
| He/she likes family. | 1.71 | 1.66 | 3.59 | 0.24 |

*Table 6. Probability bias scores.*

If the probability bias score is equal to 1, there is no difference in the probability of a model predicting one gendered target word over another. A score above 1 suggests a model is more likely to predict a masculine target word and a score below 1 indicates it is more likely to predict a feminine one. More template sentences need to be inputted in order to assess how effective it is to adjust them in this manner to quantify gender bias. However, the early scores do reveal certain trends. Based on the findings, BERT may be said to be biased as it mostly predicted masculine target words except in the sentence, “He likes children”. mBERT also predicted masculine target words for every sentence in both English and Croatian. BERTić

appears to have been more likely to predict feminine target words in every case, but these results are unusual for any LLM.

**6 Analysis**

Overall, the results show that gender bias has been encoded in BERT and mBERT for both languages and was more easily detected using the adjusted template sentences rather than the fixed word lists. When the exact word list borrowed from Kurita et al. (2019) was inputted into the English models, the p-values and effect sizes obtained were in line with expectations. The word list was very similar to the one created for this project but included men's and women's first names as target words. This word list was ultimately not selected as women's first names are not part of BERTić's vocabulary. The p-values and effect sizes for the original word list are included here for comparison with the results in Tables 4 and 5 which do not reveal interesting findings, suggesting that the word list created for this project may need to be changed for future experiments. The values for the original word list may be seen in Table 7.

| | p-value | effect size |
|---|---|---|
| BERT | 0.003 | 1.3313 |
| mBERT (En) | 0.0003 | 1.526 |

*Table 7. Permutation test p-values and effect sizes for the word list from Kurita et al. (2019).*

A p-value less than 0.01 shows statistical significance (Kurita et al., 2019), and an effect size greater than 0.8 indicates a meaningful effect (Leppink et al., 2016). By choosing men's and women's first names as target words, the English models did reveal gender bias. These findings, which are more in line with expectations than the ones obtained through the experiment designed specifically for Croatiant, demonstrate that the word list without first names is not the best-suited test for the question of gender bias. To accurately decide whether a fixed word list may be used, one would need to be created which contains words present in the vocabularies of all the models compared and produces reasonable results.

As for the probability bias scores for the adjusted template sentences, BERT and mBERT for both languages generally showed gender bias. The expectation was that BERT would either predict *she* in the context of *children* and *family* or would be equally likely to predict *she* or *he*, which was partially the case. mBERT predicted *he* would appear in every instance. In

addition, some of the probability bias scores obtained from mBERT for Croatian (8.04, 40.5) were unusually high, suggesting an issue with the test or model. Finally, the adjusted sentences did not produce reasonable scores from BERTić. A possible explanation is that gender bias is not present in the model to a measurable degree, but this is unlikely considering prior work on the topic. These unexpected results may alternatively be explained by the word list which may have been too short, or the fact that some of the models had relatively small vocabularies.

Another limitation may have been the template sentences or the specific words in the word list. Changing the word list produced different results for English as seen in Tables 4, 5, and 7. Moreover, a wide range of probability bias scores could be generated using different template sentences. It is possible that the template-based approach itself is not ideal for detecting and quantifying social biases in LLMs. Kwon and Mihindukulasooriya (2022) tested the probability bias score measure outlined in Kurita et al. (2019) on a benchmark NLP dataset. They found that the template sentence approach is not appropriate for contextual word embeddings as it is too sensitive to individual words rather than the meaning of the whole sentence. The solution offered in the study was to recalculate the probability bias score after paraphrasing the sentences a number of times for a more robust test.

## 7 Discussion

Some of the limitations of the experiments included the small training set that the Croatian models were trained on, as well as the small word list and number of template sentences used to quantify gender bias in the models. The choice of words and sentences themselves affected the results, as demonstrated by the reasonable results found by inputting the word list offered in Kurita et al. (2019) into BERT and mBERT. In a future experiment, the paraphrasing method recommended in Kwon and Mihindukulasooriya (2022) may be used to counteract the issue. Prior studies have compared biases in LLMs with human judgements about biases present in society by crowdsourcing responses from real people (Caliskan et al., 2017). Having human judgements to weigh against biases in LLMs would be especially important for an understudied linguistic and cultural context where less work has been done to assess which social biases may be encoded in a LLM.

## 8 Literary machine translation: An argument for the preservation of bias

With the growing discussion around equality in NLP, many have argued for debiasing LLMs trained to perform downstream tasks such as sentiment analysis and job applicant filtering (Weidinger et al., 2021; Bhatt et al., 2022). Considering the impact such models have on users

and those underrepresented or stereotyped in the training data, the argument stands that characterizing and reducing bias in the models may make them more useful and less harmful. Models fine-tuned for the task of machine translation have been grouped with models which should be debiased (Tomalin et al., 2021). Some of the fears are that only the prestigious form of a language may be available for translation (Savoldi et al., 2021), or that references to a female or non-binary person in the source text may be outputted as masculine in the translation due to a greater number of tokens being associated with men in the training data (Prates et al., 2018). While the issue of bias is important for MT generally, the concerns raised again relate to downstream tasks where the translation has a functional, sometimes short-term purpose such as that of a user manual. The purpose of literary MT may be distinguished in that regard from other kinds of translations, just as human literary translations are distinct from legal or medical translations.

Researchers on the topic of MT who have a background in translation studies have argued that MT will always fall short of human translation as it consistently outputs less reliable and cohesive translations (Taivalkoski-Shilov, 2019; Abdulaal, 2022). While MT is a useful tool for trained translators, some assert that a translator's goal should be to produce work as close to the source text as possible while adapting the language and cultural references for a new audience, a feat that MT will never be able to do as well as human translators (Hutchins, 1997). With the objective of preserving the source text, the social biases of the time and place in which the text was originally produced should also be preserved. The impact of poor translations has famously led to wildly varying translations of religious texts. The word "witch" in Exodus 22:18, "Thou shalt not suffer a witch to live," was often translated as "poisoner" or "wrongdoer" before the 17th century. Paranoia over witchcraft and consequent executions of supposed witches in England and Scotland have in part been attributed to the modified translation (Slaughter, 2020).

Though there are proponents of the efficiency and acceptability of MT (Locke & Booth, 1955; Besacier, 2014), the quality and usability of MT for literature is still highly contested. Omar and Gomaa (2020) have acknowledged the shortcomings of literary MT as it stands today, but also praise it for its potential as a pedagogy tool for translators-in-training, a sentiment echoed by Voigt and Jurafsky (2022). Besacier (2014) has also asserted that while MT does not equal human translation, readers of machine-translated literary works benefit from getting access to translations sooner, and authors benefit from reaching a wider audience. Taivalkoski-Shilov (2019) contested this view, stating that translators "must be experts of literary language in the source and target languages" to be able to generate a reliable and enjoyable translation.

Taivalkoski-Shilov (2019) focused on the ethical and creative implications of literary MT, arguing that its advent has meant that trained translators have been overall receiving lower pay and tighter deadlines. Some clients may opt to simply hire native speakers of a language for post-editing of a machine-translated text, which affects the creativity and quality of the translation. Passages being translated via crowdsourcing could also produce lower-quality, less-cohesive translations. Because MT translates sentence by sentence and cannot consider the entire work (Thai et al., 2022), Taivalkoski-Shilov states that it is not suitable for literature at all; a text as a whole needs to be translated which involves recreating the narrative in a way that transmits the cultural references to the reader. Creatively, a drawback of MT is that the author's and translator's voices are less perceptible. Readers often do not consciously notice the translator's voice but it is present in footnotes and stylistic choices which are tailored to the enjoyment of the audience (Kenny & Winters, 2020).

Irvine Welsh's novel *Trainspotting* features young male characters in Edinburgh struggling with substance abuse, poverty, and aimlessness. The dialogue incorporates Scottish English vocabulary, and speech is sometimes phonetically transcribed which may be difficult to understand even for readers of other English variants. A Croatian translator made the pertinent decision to use the Kajkavian dialect of Croatian spoken in the capital city in reconstructing the young, urban characters through their speech. With MT, most authors would be translated in much the same way, diminishing the voice of the author and erasing the creative potential of the translator. Furthermore, reducing bias in a MT LLM may dilute the intent and effect of the original work.

Quantifying bias in LLMs can be difficult. As discovered in the experiments for this project, the undertaking may involve manipulation of word lists and template sentences, especially for low-resource languages. Though LLMs trained for certain tasks should be debiased, it is still unclear how best to do this. Tomalin et al. (2021) compared various means of debiasing models fine-tuned for MT. The general consensus has been to debias a model prior to training and fine-tuning but Tomalin et al. show that that method produces low-quality translations. Many researchers in translation studies have recommended against debiasing literary MT LLMs as they produce translations far removed from the source text.

## 9 Conclusion

While the results in Kurita et al. (2019) were replicable for BERT and English mBERT, further experiments testing gender bias in Croatian models will need to be performed to assess how best to capture bias. The original word list was not applicable as words were missing from

BERTić's vocabulary, and the modified word list may not have worked because it did not account for case and gender in the language. However, even the template sentences which were grammatical and constructed so as not to include gender agreement were not suitable tests for gender bias in Croatian mBERT and BERTić. The models may not have a measurable amount of bias in them, which would be unusual, or the tests may need to be redesigned to accurately determine the extent of bias in the models.

Possible next steps to continue the project include implementing the paraphrase method outlined in Kwon and Mihindukulasooriya (2022), crowdsourcing human responses about biases in society to compare to biases found in the models, and studying other kinds of biases in the same Balkan context (ethnic, religious, linguistic etc.). Research on bias in LLMs trained on low-resource languages would impact smaller speech communities by making NLP tools more accessible to and equitable for users. On the topic of literary MT, further research should be conducted, not only on the ethical and aesthetic consequences of machine translating works of art, but also on the value of removing the lens through which the original text was written. Such work could help develop sustainable, well-suited tools for translation.

## 10 References

Abdulaal, M. A. A. D. (2022). Tracing machine and human translation errors in some literary texts with some implications for EFL translators. *Journal of Language and Linguistic Studies.*

Arsenijević, B. (2021). No gender in 'gender agreement': On declension classes and gender in Serbo-Croatian. *Balcania et Slavia*, (1). https://doi.org/10.30687/bes/0/2021/01/001

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Besacier, L. (2014). Machine translation for literature: a pilot study (Traduction automatisée d'une oeuvre littéraire: une étude pilote) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 389–394, Marseille, France. Association pour le Traitement Automatique des Langues.

Bhatt, S., Dev, S., Talukdar, P., Dave, S., & Prabhakaran, V. (2022). Re-contextualizing Fairness in NLP: The Case of India. *arXiv preprint arXiv:2209.12226*.

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. https://doi.org/10.1080/01690969308406949

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & Kaplan, J. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *ICLR.*

Dastin, J. (2018, October 10). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

Devlin, J. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv Preprint arXiv:1810.04805*.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.23

Fuller, J., Raman, M., Sage-Gavin, E., Hines, K., et al (September 2021). Hidden Workers: Untapped Talent. Published by Harvard Business School Project on Managing the Future of Work and Accenture.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16). https://doi.org/10.1073/pnas.1720347115

Heitmeier, M., Chuang, Y.-Y., & Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: Considerations and design choices. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.720713

Hrkač, M. Akuzativ u suvremenom dramskom tekstu. Master’s thesis, University of Zadar, 2017.

Hutchins, J. (1997). From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947-1954. A Chronology.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing*. Pearson Education.

Kenny, D., & Winters, M. (2020). Machine translation, ethics and the literary translator’s voice. *Fair MT*, *9*(1), 123–149. https://doi.org/10.1075/ts.00024.ken

Ku, Y.-M., & Anderson, R. C. (2003). Development of morphological awareness in Chinese and English. *Reading and Writing*, *16*(5), 399–422. https://doi.org/10.1023/a:1024227231216

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. https://doi.org/10.18653/v1/w19-3823

Kwon, B. C., & Mihindukulasooriya, N. (2022). An Empirical Study on Pseudo-log-likelihood Bias Measures for Masked Language Models Using Paraphrased Sentences. In *Proceedings of the*

*2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 74–79, Seattle, U.S.A.. Association for Computational Linguistics.

Lapesa, G., & Evert, S. (2014). A large-scale evaluation of distributional semantic models: parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, *2*, 531–546. https://doi.org/10.1162/tacl_a_00201

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, May 23). *How we analyzed the compas recidivism algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Leppink, J., O'Sullivan, P., & Winston, K. (2016). Effect size – large, medium, and small. *Perspectives on Medical Education*, *5*(6), 347–349. https://doi.org/10.1007/s40037-016-0308-y

Ljubešić, N., & Lauc, D. (2021). BERTi\'c--The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *arXiv preprint arXiv:2104.09243*.

Locke, W. N., & Booth, A. D. (1975). *Machine translation of languages: Fourteen essays*. Greenwood Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomić, O. M. (2006). Balkan Sprachbund Morpho-Syntactic Features. *Studies in Natural Language and Linguistic Theory.*

Omar, A., & Gomaa, Y. A. (2020). The machine translation of literature: Implications for translation pedagogy. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(11), 228. https://doi.org/10.3991/ijet.v15i11.13275

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, *9*, 261–276. https://doi.org/10.1162/tacl_a_00365

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*, 6363-6381.

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.

Slaughter, Lashonda, "King James and the Intellectual Influences of the Witchcraft Phenomenon in England and Scotland." Dissertation, Georgia State University, 2020. doi: https://doi.org/10.57709/20206634

Stanley, J. (1977). Gender-Marking in American English: Usage and Reference. *Sexism and Language*.

Taivalkoski-Shilov, K. (2019). Free indirect discourse: an insurmountable challenge for literary MT systems?. In *Proceedings of the Qualities of Literary Machine Translation*, pages 35–39, Dublin, Ireland. European Association for Machine Translation.

Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., & Kulshreshtha, A. (2022). LaMDA: Language Models for Dialog Applications. *arXiv Preprint arXiv:2201.08239*.

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology*, *23*(3), 419–433. https://doi.org/10.1007/s10676-021-09583-1

Ulčar, M., & Robnik-Šikonja, M. (2023). Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.932519

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Voigt, R. & Jurafsky, D. (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Vučković, K. Padežne gramatike i razumijevanje hrvatskoga jezika. Dissertation, University of Zagreb, 2004.

Welsh, I. (1993). *Trainspotting*. Editorial Anagrama.

Welsh, I. (1996). *Trainspotting*. Preveo Vladimir Cvetković Sever. LORA d.o.o. Koprivnica / KATARINA ZRINSKI d.o.o.