Gender bias in word embeddings of different languages

Thijs Raymakers
Delft University of Technology

June 21, 2020

Abstract

Word embeddings are useful for various applications, such as sentiment classification (Tang et al., 2014), word translation (Xing, Wang, Liu, & Lin, 2015) and résumé parsing (Nasser, Sreejith, & Irshad, 2018). Previous research has determined that word embeddings contain gender bias, which can be problematic in certain applications such as résumé parsing. This research has addressed the question whether gender bias is present in word embeddings of different languages. Gender bias has been measured on word embedding of 26 different languages with the help of the Word Embedding Association Test by Caliskan, Bryson, and Narayanan (2017). The results show that most of the tested languages seem to have bias towards male, while a few languages seem to have a bias towards female. This result is in line with previous literature.

Keywords: natural language processing, gender bias, word embedding, WEAT, language

1 Introduction

Word embeddings have become a useful tool in the field of natural language processing. They can be used to represent the semantics and the meaning of words as a vector of numbers, which is used for all sorts of applications, such as recommendation systems (Grbovic & Cheng, 2018) or résumé parsing (Hoang, Javed, Mahoney, & McNair, 2017; Nasser et al., 2018). Word embeddings can be used to solve analogies. An example of this would be that the vector that represents the word $\overrightarrow{queen}$ would be equal to $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$. This analogy would capture an 'opposite gender' relation. Other examples of these analogies would be $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Germany} \approx \overrightarrow{Berlin}$, capturing a 'capital of' relation, and $\overrightarrow{cars} - \overrightarrow{car} + \overrightarrow{apple} \approx \overrightarrow{apples}$, which would capture a 'pluralisation' relation (Vylomova, Rimell, Cohn, & Baldwin, 2016).

However, these word embeddings should be used with care because they have been shown to contain biases (Caliskan et al., 2017). Caliskan et al. (2017) showed that words related to science are more related to male terms and that words related to arts are more related to female terms. This is a form of gender bias that can be problematic when the word embedding is used in a context where gender bias should be avoided, such as automatic résumé recommendation. A recent example of a failure of such a system is Amazons recruiting tool that penalized female candidates, because they included words related to women on their résumé (Dastin, 2018).

Recently, efforts have been made to reduce the effect of bias in word embeddings. Research showed that debiasing word embeddings is possible to some extent (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016), but another study showed that this approach may hide the gender

bias, instead of removing it (Gonen & Goldberg, 2019).

This paper aims to take a step back and look into whether there is a difference in gender bias between word embeddings of different languages. Some existing related research has studied this difference before, looking at gender bias on Wikipedia across six different languages (Wagner, Garcia, Jadidi, & Strohmaier, 2015), and looking at biases when different languages, models and training sources are used (Lauscher & Glava, 2019).

Previous efforts have generally been focused on word embeddings of the English language. However, it might be the case that some languages are inherently less biased. This insight could be useful in the search towards a novel debiasing algorithm. Furthermore, this could explain why debiasing is easier or harder for certain languages.

The aim of this research is to investigate possible differences in gender bias in word embeddings of different languages. This difference has been researched by studying to what extent word embeddings of different languages are biased towards gender. This research question will be answered with the help of the WEAT method by Caliskan et al. (2017).

2 Methodology

The word embeddings provided by previous research are used for this study (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). These word embeddings have been trained for various different languages on texts from Wikipedia[1] and the Common Crawl Project[2]. The languages that are used in this research, are chosen based on the following criteria; (a) the language has a pre-trained word embedding from Grave et al. (2018) (b) the language should be machine translatable with the help of tools like DeepL[3] or Google Translate[4] and (c) the language is part of a different language family than the languages that were already chosen. This resulted in 26 languages from 16 different language families. An overview of these languages, their language families and the translations of *male* and *female* can be found in table 2 in the Appendix.

2.1 Explanation of WEAT

A popular way of measuring bias in word embeddings uses the Word Embedding Association Test (WEAT) method by Caliskan et al. (2017). WEAT has been successfully used in previous research (Gonen & Goldberg, 2019; Swinger, De-Arteaga, Heffernan IV, Leiserson, & Kalai, 2019). This method measures the association between a set of *attribute words* (e.g. *man*, *male*, *woman*, *female*) and a set of *target* words (e.g. *programmer*, *family*, *nurse*, *engineer*).

The WEAT method calculates the association between attribute words and target words by measuring the difference in cosine similarity. Formally, let X and Y be two equally-sized sets of target words and let A and B be two sets of attribute words. The test statistic is defined as

$$s(X,Y,A,B) = \sum_{\vec{x}\in X} s(\vec{x},A,B) - \sum_{\vec{y}\in Y} s(\vec{y},A,B) \quad \text{where} \quad (1)$$

$$s(\vec{w},A,B) = \text{mean}_{\vec{a}\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{\vec{b}\in B}\cos(\vec{w},\vec{b}) \quad (2)$$

where $\cos(\overrightarrow{a}, \overrightarrow{b})$ measures the cosine of the angle between $\vec{a}$ and $\vec{b}$ and where the mean is the sample mean of the values. Formula (2) can be described as a measurement of the association of the target word w with an attribute, and formula (1) can be described as a measurement of the differential association of the sets of target words with the attributes (Caliskan et al., 2017).

In WEAT, the effect size d_1 of the test statistic is defined as the normalized measure

[1] https://wikipedia.org/
[2] https://commoncrawl.org/
[3] https://deepl.com/
[4] https://translate.google.com/

of separation between two distributions of associations (Caliskan et al., 2017)

$$d_1 = \frac{\text{mean}_{\vec{x}\in X}\, s(\vec{x}, A, B) - \text{mean}_{\vec{y}\in Y}\, s(\vec{y}, A, B)}{\text{std-dev}_{\vec{w}\in X\cup Y}\, s(\vec{w}, A, B)} \tag{3}$$

The effect size d_1 can be interpreted as the "amount of bias" (Lauscher & Glava, 2019).

2.2 Weaknesses of WEAT for multilingual use

The WEAT method works for when measurements are being performed in a single language, but some problems could arise when the differences between different languages are compared.

First, not all words have a one-to-one translation into another language, because certain words might have a different meaning, depending on the context in which the word is used. It is therefore not possible to translate all the attribute and target words into a different language, because it might lead to an incorrect and unfair comparison. A great example of this would be the English word *man*, which can mean *male* or *mankind*, depending on the context.

Second, the inclusion or exclusion of words from the target set is a subjective decision (Nissim, van Noord, & van der Goot, 2020). Therefore, the measurement itself could be biased, depending on what words are or are not in the set of target words. This problem will only be amplified when WEAT is used for multiple languages, because this subjective decision has to be made for every single language.

2.3 Proposed multilingual WEAT

The first problem of translation can be addressed by limiting the amount of words that have to be translated. The method that is used in this research only uses the two attribute words *male* and *female*. The assumption is made that these words are universal and translatable in all languages.

It is proposed to solve the second problem of subjectivity with two different approaches. The goal is to remove the subjective decision and include words based on a language-independent metric that is not based on the meaning of words. This has been done by (1) looking at *all* the words in the word embedding and by (2) looking at the *most used* words in the word embedding.

2.3.1 Method 1: Uniform weighting

The first method, the uniform weighting method, uses all the words in the word embedding as target words. This method uses an adjusted version of the WEAT method described in section 2.1. The adjustments made take into account that the compared languages have different target and attribute words. Formulas (1) and (2) are redefined as

$$s(X, Y, a, b) = \sum_{\vec{x}\in X} s(\vec{x}, \vec{a_X}, \vec{b_X}) - \sum_{\vec{y}\in Y} s(\vec{y}, \vec{a_Y}, \vec{b_Y}) \quad \text{where} \tag{4}$$

$$s(\vec{w}, \vec{a}, \vec{b}) = \cos(\vec{w}, \vec{a}) - \cos(\vec{w}, \vec{b}) \tag{5}$$

where $\cos(\vec{a}, \vec{b})$ measures the cosine distance between between $\vec{a}$ and $\vec{b}$. In formula (4), X and Y refer to the two languages that are compared, and a_X and b_X refer to the translations of the two attribute words a and b into language X respectively. The effect size d_1 is calculated the same way as WEAT, with formula (3).

The advantage of this approach is that all words are considered, which eliminates the inclusion or exclusion subjectiveness described by Nissim et al. (2020). This is because there is no metric on which a word could be excluded, because all words in a language are included by default. The disadvantage is that all words have the same weight, which might not be an accurate representation of the language, as not all words are used as often as others. This is addressed in an alternative method described in section 2.3.2.

2.3.2 Method 2: Frequency weighting

The disadvantage of method 1, the uniform weighting method, is addressed in method 2, the frequency weighting method. The second

method still uses all the words in the word embedding, but attaches a weight to each word, depending on its usage frequency. Effectively, this results in a measurement over the most commonly used words, because those words have the highest weight. The frequency weighting method uses an adjusted version of the uniform weighting method. The adjustments aim to include information about the usage frequency. Formulas (4) and (5) are redefined as

$$s(X, Y, a, b) = \sum_{\vec{x} \in X} f(\vec{x}) * s(\vec{x}, \vec{a_X}, \vec{b_X}) - \sum_{\vec{y} \in Y} f(\vec{y}) * s(\vec{y}, \vec{a_Y}, \vec{b_Y}) \quad \text{where} \tag{6}$$

$$s(\vec{w}, \vec{a}, \vec{b}) = \cos(\vec{w}, \vec{a}) - \cos(\vec{w}, \vec{b}) \tag{7}$$

where $\cos(\vec{a}, \vec{b})$ measures the cosine distance between between $\vec{a}$ and $\vec{b}$. In formula (6), X and Y refer to the two languages that are compared, and a_X and b_X refer to the translations of the two attribute words a and b into language X respectively. $f(\vec{x})$ refers to the probability that the word $\vec{x}$ occurs in a text written in language X.

The effect size d_2 of the test is defined as an adjusted version of formula (3) that takes the frequency of the words into account:

$$d_2 = \frac{\sum_{\vec{x} \in X} f(\vec{x}) * s(\vec{x}, A, B) - \sum_{\vec{y} \in Y} f(\vec{y}) * s(\vec{y}, A, B)}{\text{std-dev}_{\vec{w} \in X \cup Y} s(\vec{w}, A, B)} \tag{8}$$

where the standard deviation takes the frequencies defined by f into account.

3 Experimental setup

Pre-trained word embeddings created by Grave et al. (2018) were downloaded for all languages defined in table 2 in the Appendix. For each downloaded language, the words *male* and *female* have been translated beforehand. The vectors associated with the translated words *male* and *female* are extracted from the word embedding. Then, the cosine similarity is calculated according to the methodology. The code used to calculate the results can be found on GitHub[5].

3.1 Frequency information

The words contained in the pre-trained models provided by Grave et al. (2018) are sorted by frequency, but exact information about how often a word occurs in a regular text is omitted. It is approximated that the frequency with the help of Zipf's law (Gao, Zhou, Luo, & Huang, 2019; Zipf, 1935), defined as

$$\frac{1/k}{\sum_{n=1}^{N}(1/n)} \tag{9}$$

where k denotes the rank of a word and N denotes the amount of words in the word embedding.

3.2 Dummy language

A "dummy language" has been created in order to test whether and to what extent a language has gender bias. This language is defined to have no gender bias and is therefore used as a baseline to which the other languages are compared. It can be constructed by creating a language where $s(\vec{w}, \vec{a}, \vec{b}) = 0$ for all values of $\vec{w}$. All languages in table 2 in the Appendix are compared to this dummy language. In all formulas, the word embedding X is an actual embedding from the languages in table 2, and Y is the dummy language that has the same size as X. This has the effect that all terms containing Y are reduced to 0, except in formulas where X and Y are used in the same term such as in formulas (3), (8) and (10).

3.3 Hypothesis test

The null hypothesis is defined as "The word embedding of the tested language is not biased towards gender". This null hypothesis will be tested for all 26 languages and for both the

[5]Git repository can be found on https://github.com/ThijsRay/cse3000

uniform and the frequency weighting method. Each language is compared against the dummy language with formula (4) and (6). The significance of each comparison is measured with an $\alpha = 1 - \sqrt[26]{1 - 0.05} \approx 0.001$ and a two-sided approximated permutation test

$$p = \begin{cases} \Pr_i[s(X_i, Y_{,i}, a, b) > z] & \text{if } z > 0 \\ \Pr_i[s(X_i, Y_{,i}, a, b) < z] & \text{if } z < 0 \end{cases}$$

$$\text{where } z = s(X, Y, a, b) \tag{10}$$

where X_i and Y_i stand for all equally-sized partitions of $X \cup Y$ (Caliskan et al., 2017). The permutation test has been performed with $N = 1000$ random permutations.

4 Results

4.1 Uniform weighting method

The permutation test has been performed on the uniform weighting method and it showed that all 26 languages have a p value $< \alpha$, and are therefore significant, with respect to the dummy language described in section 3.2. Therefore, the null hypothesis, defined in section 3.3, is rejected for all of the tested languages when the uniform weighting method is used.

In figure 1a the effect size d_1 is shown for each language for the uniform weighting method. Finnish has the highest positive d_1 of all the tested languages, indicating that, on average, Finnish words have the highest association with *male* of all the tested languages. On the other hand, Basque has the highest negative d_1 of all the tested languages, showing that Basque words, on average, have the strongest relation with *female* of all the tested languages. 16 out of 26 of the tested languages have a positive effect size d_1, which demonstrates that most of the languages have a stronger link with *male* than *female* when this is calculated with the uniform weighting method.

Figure 2a presents the mean of the cosine distance and it shows a similar result to figure 1a. Again, 16 out 26 of the languages have a positive value, which leads to the same observation as in figure 1a. In figure 2a, Hungarian has the most association with *male* and Hindi has the most association with *female*.

4.2 Frequency weighting method

The permutation test has also been performed on the frequency weighting method, which showed that 16 out of 26 languages have a p value $< \alpha$ with respect to the dummy language described in section 3.2. This is deemed significant and the null hypothesis can be rejected for these languages.

The 10 languages that have a p value $> \alpha$ and their respective p values are shown in table 1. The null hypothesis can *not* be rejected for these languages.

Table 1

Calculated p *values from values of the frequency weighting method*

Language	p value
Portuguese	0.173
Russian	0.424
Japanese	0.008
Turkish	0.033
Korean	0.003
French	0.146
Polish	0.448
Hungarian	0.456
Thai	0.494
Javanese	0.017

Note: All languages with p value $< \alpha$ are not shown in this table.

Figure 1b shows the effect size d_2 of each language for the frequency weighting method. Burmese has highest positive d_2 of all the tested languages, indicating that the most used words in Burmese have the highest link with *male* of all the tested languages. Greek has the highest negative effect size d_2 of all the tested languages, indicating that the most used words in Greek have the strongest relation with *female* of all the tested languages.

Figure 2b displays the weighted mean of the cosine distance for the frequency weighting method. 15 out of 16 of the significant languages have a positive value, which leads to the

same observation made for figure 1b.

The overall result is similar for both methods. A notable difference is that there are more languages that have a positive effect size d_2 in figure 1b than d_1 in figure 1a. This means that the most used words of the tested languages are more associated with word *male* than all of the words in the language are.

5 Discussion

The objective of this study was to identify to what extent word embeddings of different languages are biased towards gender. It was hypothesised that word embeddings of the tested languages were not biased towards gender.

5.1 *P* values

It can concluded from the significance test that all of the tested languages are gender biased, when gender bias is measured across *all the words* in the word embedding (uniform weighting method). It can also be concluded that a majority of the tested languages is gender biased, when this is calculated across *the most used words* in the word embedding (frequency weighting method). This conclusion builds upon the assumption that there is a relation between word association and bias.

It is interesting to see that there is a difference in p values for both methods. A probable explanation for this is that most of the values in the frequency weighting method are extremely close to zero due to the introduction of frequency weights in formula (6). Therefore, the values are not significantly different from the zero values of the dummy language. Unless the most used words of a language show bias, the permutation test is essentially comparing negligible values with each other.

5.2 Observations

A few observations can be made from figure 1a. 16 out of 26 languages have a significant effect size d_1 above 0, indicating that most of the reviewed languages are more associated with *male*. Assuming that there is a relation between the association of a word and bias, then this could mean that most of the tested languages have a bias towards *male*.

Another observation from figure 1a and 2a is that some languages that are from the same language family have a similar rank. For example, both Finnish and Hungarian are Uralic languages and have a relatively high effect size d_1 compared to the other languages. A similar pattern can be seen with Portuguese and Spanish, both Iberian languages. However, this might be incidental because some closely related languages seem to have opposite effects, like German and Dutch. Further research could show whether this effect is indeed incidental or if there is a underlying reason behind it.

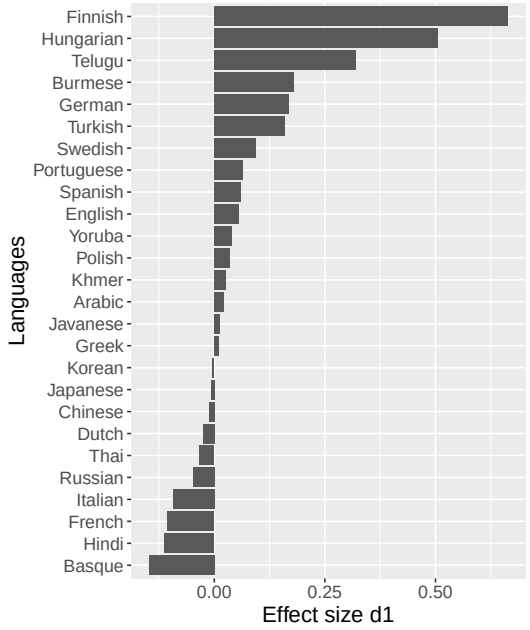
Comparable observations can be made from figure 1b and 2b. 15 out of 16 significant languages have an value above 0, indicating that *most used words* of the reviewed languages are more associated with *male* than with *female*. This is interesting because the most used words of a language say something about how a language is actually used by people. From this result it can be concluded that either people are more inclined to use *male*-related words or the most used words happen to be more related to *male*.

Some of the languages seem to be closer associated with *female*-related words, like Basque, Hindi and French in figure 1a and 2a, and Greek in figure 1b and 2b. Several questions remain unanswered at present, such as why these languages have a stronger relation with *female*-related words and why the majority of the languages have a stronger link with *male*-related words. Another important issue for future research is the possible link between gender bias in languages and gender equality. Future studies on these topics are therefore recommended.

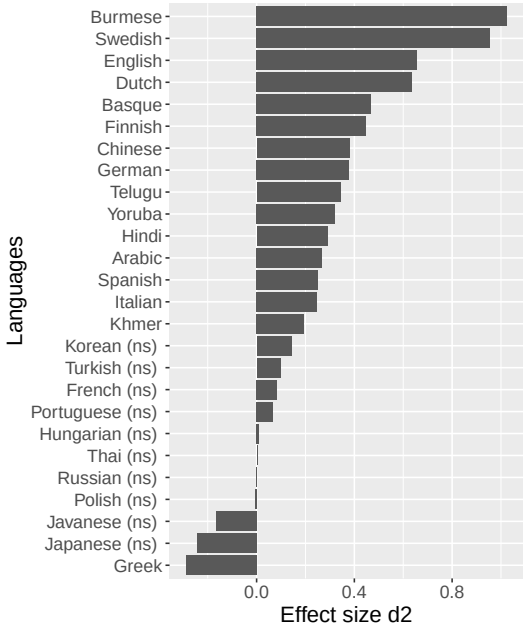
These results show that the words of the majority of the tested languages have a stronger association with *male* than with *female*. If there is a link between this result and gender-equality, this could be seen as empirical evidence in favour of the linguistic relativity hypothesis, that states that the language we speak

Figure 1

Figures 1a and 1b show the measured effect size d_1 and d_2 per language for both the uniform and the frequency weighting method respectively. Because these languages were compared against the dummy language described in section 3.2, the effect size is a metric of whether words are, on average, more associated with male *or more associated with* female*.*



(a) Uniform weighting method. The effect size d_1 has been calculated with formula (3). A value greater than zero indicates that words are on average more associated with *male*, while a value below zero indicates that words are on average more associated with *female*.



(b) Frequency weighting method. The effect size d_2 has been calculated with formula (8). A value greater than zero indicates that the most used words are more associated with *male*, while a value below zero indicates that words are more associated with *female*. Languages appended with (ns) were found to have a p value $> \alpha$ and therefore not significant.

---

influences the way we think (Lucy, 1997). However, one can also argue that it is the other way around, i.e. that the language itself is influenced by the way that we think.

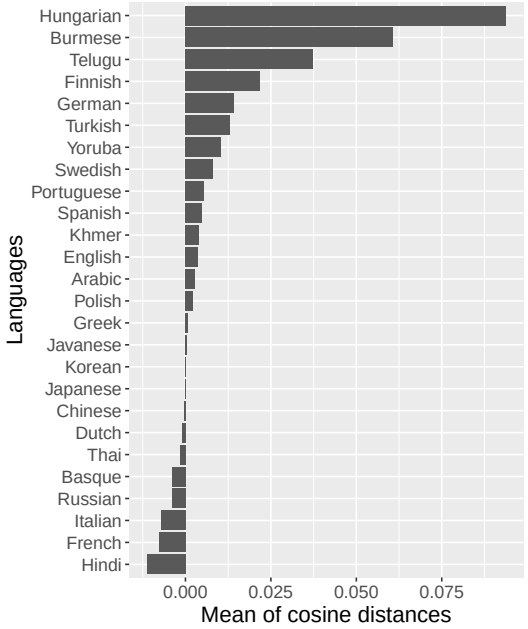
Thus, from these results and observations it is concluded that the null hypothesis can be fully rejected for the uniform weighting method, and partly rejected for the frequency weighting method. This is because only 62% of the tested languages had a significant result for the frequency weighting method. Based on the results from the null hypothesis and on the observations that were made, it can be stated that word embeddings of different languages are biased towards gender, but that the amount of bias differs per language. These results are in line with those of previous research (Caliskan et al., 2017).

5.3 Limitations

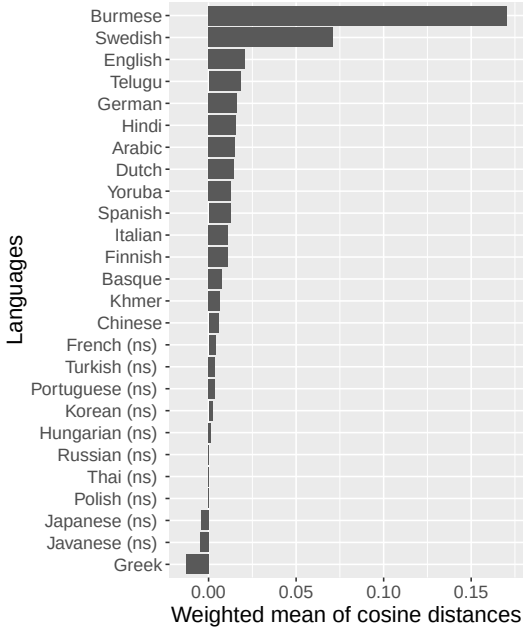
This study has potential limitations, mainly concerning the assumptions the research is based on, the critique the used methodology has previously received, and the generalisability of the results. The first of these limitations is the assumption is that the attribute words *male* and *female* are universal and that their translations have identical meaning in all of the tested languages. This is difficult to verify

Figure 2

Figures 2a and 2b show the mean and weighted mean of the cosine distance per language between each word and male *and each word and* female *for the uniform and frequency weighting methods respectively.*



(a) Uniform weighting method. The mean of the cosine distances is calculated with formula 5. A value greater than zero indicates that a words are on average more associated with *male*, while a value below zero indicates that words are on average more associated with *female*.



(b) Frequency weighting method. The weighted mean of the cosine distances is calculated with formula 7 and $f(\vec{x})$ as defined in section 2.3.2. A value greater than zero indicates that a words are more associated with *male*, while a value below zero indicates that words are more associated with *female*. Languages appended with (ns) were found to have a p value $> \alpha$ and therefore not significant.

without having a good understanding of all the languages that are involved. This issue could be alleviated if language experts would confirm that the translations are accurate. This limitation can also be alleviated by using more than one attribute word for each gender, as originally done by Caliskan et al. (2017). This reduces the probability that a word is mistranslated.

Another assumption is that the most commonly used words, like *and* and *the* are gender neutral. This might not necessarily the case, since words like *he* and *she* are also very common. Attaching a weight based on the frequency might do nothing more than measure the frequency of gender specific words, because they have a more pronounced cosine distance. If the word *he* occurs much more often than the word *she*, then the word embedding as a whole will appear to have a strong bias towards *male*. Based on the reasoning behind the linguistic relativity hypothesis (Lucy, 1997), one could argue that the fact that masculine words have a higher frequency is also a form of bias. If people are more probable to use masculine words, then it is likely that people will also, on average, think and act more in favor of males.

It is also difficult to draw any conclusions based on the relation between word embeddings and the language itself, because a word embedding is not a perfect representation of a lan-

guage. The used models have been trained on text that was posted on the internet, on websites such as Wikipedia (Grave et al., 2018). Research has found that less than 15% of the contributors of Wikipedia were female (Collier & Bear, 2012). It is reasonable to assume that a word embedding would be a better representation of a language if it was trained on text that was written for 50% by males and 50% by females.

Besides that, it should be takes into account that the WEAT method by Caliskan et al. (2017) has received critique from Ethayarajh, Duvenaud, and Hirst (2019), who argue that WEAT has theoretical flaws that cause it to systematically over-estimate bias. Since our methodology is an adjusted form of WEAT, it might be the case that these theoretical flaws are present in our methodology.

Finally, the results are heavily reliant upon the generated dummy language from chapter 3.2. If the parameters used for generating the dummy language are changed, the entire result changes. The results can therefore only be considered in the context of the dummy language. This reduces the generalisability of the results.

6 Conclusion

This study set out to assess to what extent word embeddings of different languages are biased towards gender. The results of this investigation show that word embeddings of different languages have a bias towards gender and that they are generally more biased towards *male* than towards *female*. The insight gained from this contribution may be of assistance to future debiasing efforts and possibly affirms the linguistic relativity hypothesis (Lucy, 1997). In spite of the limitations of this study, the study certainly adds to our understanding of gender bias in word embeddings of different languages.

Future research is needed to fully understand why certain languages seem to be more biased towards *male* or *female*. Further work needs to be done to establish why the majority of the tested languages have a bias towards *male*. Finally, it is recommended that this research is replicated for different forms of bias, such as racial or sexuality bias.

7 Acknowledgements

I would like to thank my supervisors for the help and expertise they have provided during the course of this project. I would also like to thank my peers who were able to give me feedback during the process.

References

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and Reducing Stereotypes in Word Embeddings. *arXiv e-prints*, arXiv:1606.06121. arXiv: 1606 . 06121 [cs.CL]

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

Collier, B., & Bear, J. (2012). Conflict, criticism, or confidence: An empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 383–392). doi:10.1145/2145204.2145265

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1696–1705). doi:10.18653/v1/P19-1166

Gao, L., Zhou, G., Luo, J., & Huang, Y. (2019). Word embedding with zipfs context. *IEEE Access*, *PP*, 1–1. doi:10.1109/ACCESS.2019.2954691

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 609–614). doi:10.18653/v1/N19-1061

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*.

Grbovic, M., & Cheng, H. (2018). Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 311–320). doi:10.1145/3219819.3219885

Hoang, P., Javed, F., Mahoney, T., & McNair, M. (2017). Large-scale occupational skills normalization for online recruitment.

Lauscher, A., & Glava, G. (2019). Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the eighth joint conference on lexical and computational semantics (*SEM 2019)* (pp. 85–91). doi:10.18653/v1/S19-1010

Lucy, J. A. (1997). Linguistic relativity. *Annual Review of Anthropology*, *26*(1), 291–312. doi:10.1146/annurev.anthro.26.1.291. eprint: https://doi.org/10.1146/annurev.anthro.26.1.291

Nasser, S., Sreejith, C., & Irshad, M. (2018). Convolutional neural network with word embedding based approach for resume classification. In *2018 international conference on emerging trends and innovations in engineering and technological research (icetietr)* (pp. 1–6). doi:10.1109/ICETIETR.2018.8529097

Nissim, M., van Noord, R., & van der Goot, R. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, *0*(ja), 1–17. doi:10.1162/COLI_a_00379. eprint: https://doi.org/10.1162/COLI_a_00379

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (pp. 305–311). doi:10.1145/3306618.3314270

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1555–1565). doi:10.3115/v1/P14-1146

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1671–1682). doi:10.18653/v1/P16-1158

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *arXiv e-prints*, arXiv:1501.06307. arXiv: 1501.06307 [cs.CY]

Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 1006–1011). doi:10.3115/v1/N15-1104

Zipf, G. (1935). *The psychobiology of language: An introduction to dynamic philology*. Cambridge, Mass.: M.I.T. Press.

8 Appendix

Table 2

Chosen languages and their translations

Language	Language family	Branch	Translation *male*	Translation *female*
Arabic	Afroasiatic	Semitic	ذكر	أنثى
Basque	Isolate	Basque	gizonezkoa	emakumezkoak
Burmese	Sino-Tibetan	Lolo-Burmese	အထီး	အမျိုးသမီး
Chinese	Sino-Tibetan	Sinitic	男	女
Dutch	Indo-European	Germanic	mannelijk	vrouwlijk
English	Indo-European	Germanic	male	female
Finnish	Uralic	Finnic	uros	nainen
French	Indo-European	Romance	mâle	femelle
German	Indo-European	Germanic	männlich	weiblich
Greek	Indo-European	Hellenic	αρσενικός	θηλυκός
Hindi	Indo-European	Indo-Aryan	पुरुष	महिला
Hungarian	Uralic	Ugric	férfi	nő
Italian	Indo-European	Romance	maschio	femmina
Japanese	Japonic	Japanese	男性	女性
Javanese	Austronesian	Malayo-Polynesian	lanang	wadon
Khmer	Austroasiatic	Khmer	ប្រុស	ស្ត្រី
Korean	Koreanic	Koreanic	남성	여성
Polish	Indo-European	Balto-Slavic	mężczyzna	kobieta
Portuguese	Indo-European	Romance	masculino	feminino
Russian	Indo-European	Balto-Slavic	мужчина	женщина
Spanish	Indo-European	Romance	hombre	mujer
Swedish	Indo-European	Germanic	manlig	kvinnlig
Telugu	Dravidian	South-Central	పురుషుడు	స్త్రీ
Thai	Kra-Dai	Tai	ชาย	หญิง
Turkish	Turkic	Oghuz	erkek	kadın
Yoruba	Niger-Congo	Volta-Niger	akọ	abo

Note: An overview of all the languages that were used in this research with their respective translations of *male* and *female*.