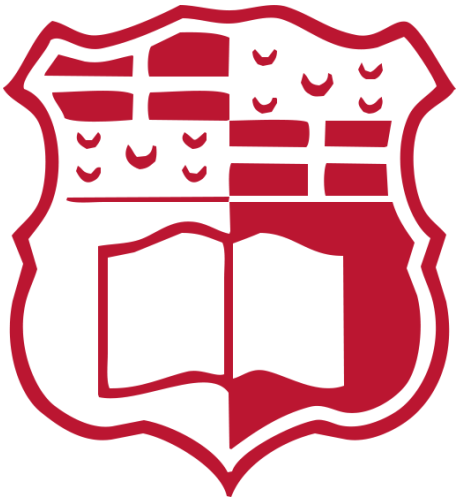


# LIN3012 Assignment

# Bias detection using distributional models within the Maltese language

Andrew Aquilina

andrew.aquilina.18@um.edu.mt

2020-2021

# Contents

# 1 Introduction

Word embeddings is a natural language processing technique used to encode words as vectors. The popularity of this technique has increased over the past decade, from information retrieval [1] to text classification [2]. Mikolov et al. [3] and Rubenstein and Goodenough [4] noted that these embeddings can represent simple semantic relationships between words, as shown below.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Unfortunately, the persuasive bias present in language has made itself into word embeddings. Bolukbasi et al. [5] show that certain words are more strongly associated with certain genders. For example, the authors pinpoint implicit sexism within word embeddings trained on Google News articles: *man is to computer programmer, as woman is to homemaker* is an example. There is a substantial amount of debate whether machine learning practitioners should aim to address bias in distributional semantic models [6]. Bolukbasi et al. claim that lack of care towards this field will only intensify the pertained biases, while Caliskan et al. [6] view "debiasing" word embeddings with scepticism.

## 1.1 Motivating Scenario

The majority of work concerning bias detection and reduction using distributional semantic models has been done using the English language. In this project, gender bias within the Maltese language will be explored. Maltese assigns grammatical gender not only to animate nouns (such as man, woman, father, mother) but also to inanimate and abstract nouns (e.g. health, strength, table). For example, the word 'health' (saħħa) is feminine in Maltese. This makes the exploration and minimisation of gender bias a challenging task. The work of Gonen et al. [7] note that standard debiasing methods fail to remove this effect in languages with gender-marked nouns, providing alternative methods.

### 1.2 Problem Definition

The purpose of this research is to explore, identify and address any gender bias present within a Maltese corpus. To address this problem, the following aims and objectives are considered.

1. Train and fit a number of distributional semantic models on a Maltese corpus, namely Word2Vec [8] and FastText [9]. The outcomes of these embeddings will also be compared with a traditional distributional semantic model: TF-IDF Vectorizer.

2. Use core 'seed' terms to explore the clustering of other terms and determine the level of gender bias present within the trained embeddings.

3. Explore methods of gender debiasing in the face of gender-marked language.

The structure of this report is as follows: Section 2 briefly reviews any surrounding literature concerning bias detection in word embeddings, Section 3 includes a full description of the corpora and methods used, and Section 4 reports the produced reports. Concluding remarks are made in Section 5.

## 2 Literature Review

The psychological status of societal groups can be strengthened or weakened by the bias pertaining their language [10]. Even in the cases where biases may appear to be positive, their effect remains to be harmful and discriminatory [11]. Bias in our language does not only affect how we behave and interpret each other. In the context of machine learning, Bolukbasi et al. [5] provide a hypothetical scenario where a given search engine ranks males higher than females, as a result of gender bias in trained word embeddings.

There have been a number of approaches to quantify and observe gender bias. Caliskan et al. [6] proposed the Word Embedding Association Test (WEAT), which can be used to measure the association between gender and a given concept word, such as an adjective

or occupation. Bolukbasi et al. [5] differentiate between direct and indirect gender biases, addressing them separately. By identifying the gender direction $\overrightarrow{g}$, and a set of words, $N$, which should be gender-neutral, the direct bias $B$ of a word embedding can be calculated as shown in Equation 1. The parameter $c$ determines measure strictness.

$$B_c = \frac{1}{|N|} \sum_{w \in N} |\cos(\overrightarrow{w}, \overrightarrow{g})|^c \tag{1}$$

To mitigate against this, Bolukbasi et al. proposes the following debiasing algorithms: mitigation through neutralisation followed by equalisation, or through softening. The former ensures that every word within the set $N$ is equidistant to all others words in some specified gender set. This removes certain semantic associations which may be considered valuable. The latter method attempts to balance between bias mitigation and maintaining semantic association.

The work discussed so far have focused on quantifying and mitigating gender bias within the English language. Bolukbasi et al. [5] conclude that it is unclear whether the findings reported in their work carry over to other languages, especially to gender-marked languages. Furthermore, Zhou et al. [13] note that such languages may contain more than two gender classes. For the scope of this project, we follow the literature which address binary gender. Recent work by Gonen et al. [7] has demonstrated the failure of standard debiasing methods for German and Italian, which are both gender-marked languages. Gonen et al. proposed lemmatisation and gender change methods. The former lemmatises context words within the dataset, omitting gender signals completely, while the latter neutralises gender signals by changing context words to the same gender. The authors found that lemmatisation worked better for German, while the latter received better results for Italian. Zhou et al. [13] worked on debiasing methods for other gender-marked languages, namely Spanish and French. Zhou et al. extended the WEAT definition discussed above to quantify individual words in gendered languages, introducing the Modified Word Embedding Association Test.

The authors carried on to report that by aligning Spanish embeddings to English, a semantic discrepancy between the genders was found; whereby the distance between the Spanish female-denoting words, and their English counterparts, was far greater than the distance between the Spanish male-denoting words. One of the methods the authors proposed in mitigating such effect is by shifting the two forms of the same word. Gonen and Goldberg [14] show that mitigating gender bias through such methods are mostly superficial. However, Zhou et al. point out that Gonen and Goldberg's work was not applied directly for gender-marked languages. Their reported results indicate that their methods successfully reduce gender biases, while maintaining the original embedding's quality [13].

# 3 Data and Methodology

## 3.1 Data

The Maltese Language Resource Server (MLRS) v3 corpus[1] was obtained, containing around 250 million tokens in a variety of categories, mainly from parliamentary speeches. The corpus' files are in vertical XML format, where every word is on a separate line, containing a Part of Speech tag, lemma and morphological root. Pre-processing steps included: formatting each file as a standard text file, lowercasing, omitting punctuation, and converting Maltese letters into their English equivalents (for example, "Baħar iċ-Ċagħaq" → "bahar ic caghaq"). Additionally, this procedure was repeated for an alternative version of the dataset containing lemmatised words, obtained from the MLRS corpus itself. We will refer to these datasets as *original* and *neutral* respectively. The motivation behind this is explained in Section 3.2.4. Splitting the dataset into training and testing was deemed to not be applicable given the discussed aims. Furthermore, we adapt the work of Bolukbasi et al. for the Maltese language, gathering the following sets for gender bias mitigation.

[1] `https://mlrs.research.um.edu.mt/`

$S_0$: 7 definitional pairs to identify gender subspace $\overrightarrow{g}$ (for example, "raġel/mara").

$S_1$: 73 occupational pairs[2] (for example, "tabib/tabiba").

$S_2$: 48 adjective pairs[3] (for example, "sabiħ/sabiħa").

$G$: a list of 70 core "seed" terms which are naturally gendered.

$E$: 19 equality sets, as defined by Bolukbasi et al.

(It may be unclear to note the significance of $G$, given that every Maltese word is inherently gendered. The motivation behind obtaining each set is outlined in the following section.)

## 3.2 Methodology

The methods discussed in this section are packaged within a singular Jupyter notebook, found within the deliverables of this assignment. Open-source code provided by Bolukbasi et al. was also obtained[4] and tailored accordingly.

### 3.2.1 Building distributional semantic models

To align our work with previous research [5, 13], we have trained the following state of the art embedding models, Word2Vec [8] and FastText [9], on the MLRS corpus. The Python library `gensim` was utilised for training. Both models were trained on an AMD Ryzen 5 3550H 4-core 2.93 GHz CPU using the same set of training parameters: 100-vector dimensions, 10-word context windows, a minimum word count of 2, and 4 worker threads (one for each CPU core). They were trained for 10 epochs using the *original* and *neutral* datasets, resulting in a total of 4 trained models. Figure 1 plots the Word2Vec[5] loss values

[2]The English occupational terms provided by Bolukbasi et al. were translated to Maltese using the Python library `google-trans-new`, followed up by careful curation.

[3]The majority of which were obtained from `http://mylanguages.org/maltese_adjectives.php`.

[4]`https://github.com/tolga-b/debiaswe/`

[5]At the time of writing, the `gensim` library does not provide loss-epoch values for the FastText model.

for every epoch. It is to be noted that the *neutral* dataset achieved a better loss due to a smaller vocabulary.

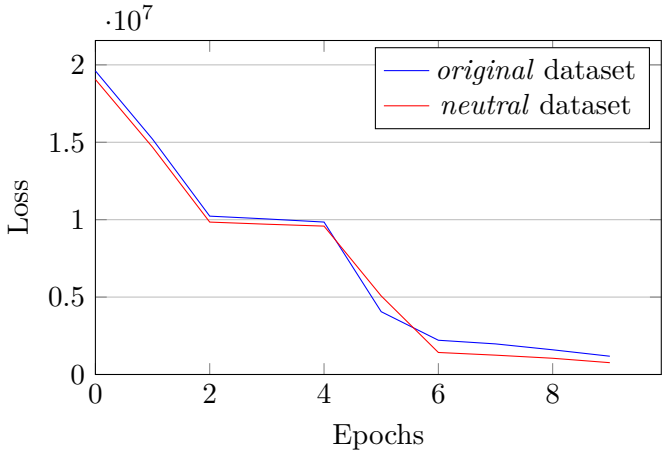



Figure 1: Word2Vec loss by epochs graph.

We have also trained a TF-IDF Vectorizer model to compare it with the aforementioned state of the art embeddings. In general it was unable to properly capture semantic information. As a result, the model was not considered for gender bias observation and mitigation. More on this in Section 4.1.1. However, a prototype for the word analogy problem was developed. Given the word analogy $a : x :: b : y$, the prototype utilises the TF-IDF model to calculate the cosine similarity for $\overrightarrow{w_y}$ and $(\overrightarrow{w_x} + \overrightarrow{w_b} - \overrightarrow{w_a})$.

### 3.2.2 Gender direction identification

Before gender bias can be observed within the trained embeddings, the gender direction $\overrightarrow{g}$ is to be identified. This was done by taking seven gender pair difference vectors (from $S_0$), computing their principal components (PC), and taking the top PC [5].

### 3.2.3 Gender bias observation

The following methods were utilised to quantify and observe gender bias in the face of gender-marked languages. An apparent disadvantage of both methods is that they do not account for words which associate themselves with only one gender. Examples include most abstract and inanimate nouns, such as “saħħa” (female), “imħabba” (female), “bieb” (male) and “siġġu” (male).

- $O_1$**: Aggregating biases $b_m$ and $b_f$ for a given word $w$.**

Given a word gender pair $w = (w_m, w_f)$, its bias $b_w$ is simply defined as the aggregation of $b_m$ and $b_f$, where $b_m = \overrightarrow{w_m} \cdot \overrightarrow{g}$ and $b_f = \overrightarrow{w_f} \cdot \overrightarrow{g}$. Values on either side (+ for male, − for female) indicate the direction and magnitude of the word's bias.

- $O_2$**: Using the Modified Word Embedding Association Test (MWEAT).**

As discussed in Section 2, the MWEAT measures the difference in the association strength of two target sets (male and female) with two attribute sets (such as $S_1$ and $S_2$). Given a word gender pair $w = (w_m, w_f)$, its bias $b_w$ can be calculated as shown in Equation 2, where $M$ and $F$ are gender definitional sets, and $s(\overrightarrow{w}, M, F)$ is the difference between average cosine similarities of the given attributes.

$$b_w = ||s(\overrightarrow{w_m}, M, F) - |s(\overrightarrow{w_f}, M, F)|| \tag{2}$$

### 3.2.4 Gender bias mitigation

The following methods were utilised to mitigate the observed gender bias within the trained embeddings.

- $M_1$**: Bolukbasi's Hard Debias.**

We experiment with Bolukbasi's hard debias method to report its performance on the Maltese language. The method requires a gender direction $\overrightarrow{g}$, a set of definitional gender

pairs (equality sets in $E$), and a list of "seed" terms which are naturally gendered ($G$). In their work, Zhou et al. note that while the method does successfully manage to debias gender-marked embeddings, it loses core semantic relationships in the process. We expect similar results.

- $M_2$**: Using the *neutral* dataset.**

Differentiating between grammatical and semantic gender is instrumental for bias mitigation in gender-marked languages. Gonen et al. note that doing so is not straight-forward within a post-processing phase, proposing word lemmatisation before training. This removes gender inflection from the context, but also decreases the corpus' vocabulary size. We expect an improvement from the observed gender bias on the *neutral* dataset.

- $M_*$**: Hybrid method.**

We consider a hybrid method that integrates the aforementioned approaches, i.e. hard debiaising models which were trained on the *neutral* dataset.

# 4 Results

In this section, we report the performance of the word embeddings before debiasing, provide visualisations of the observed gender bias, and tabulate results from the aforementioned methods.

## 4.1 Word Embeddings Evaluation

Embeddings are evaluated by how well they perform on generating analogies. To the best of our knowledge, there are no relevant baselines reported for the MLRS corpus. Figure 2 plots the number of generated word analogies for each model, where $\epsilon$ is the confidence threshold of how good a given analogy might be. Naturally as the $\epsilon$-value increases, the

number of generated word analogies decreases. Contrary to the work of Bolukbasi et al., the differentiation between stereotypical and appropriate word analogies was not considered.

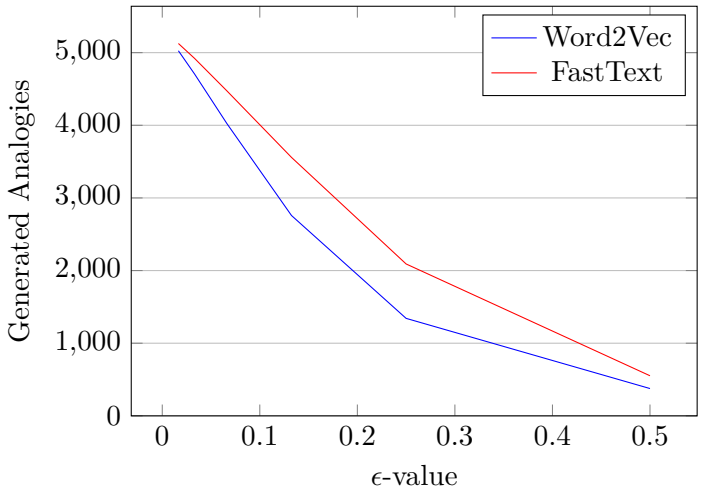



Figure 2: Number of generated analogies before debiasing.

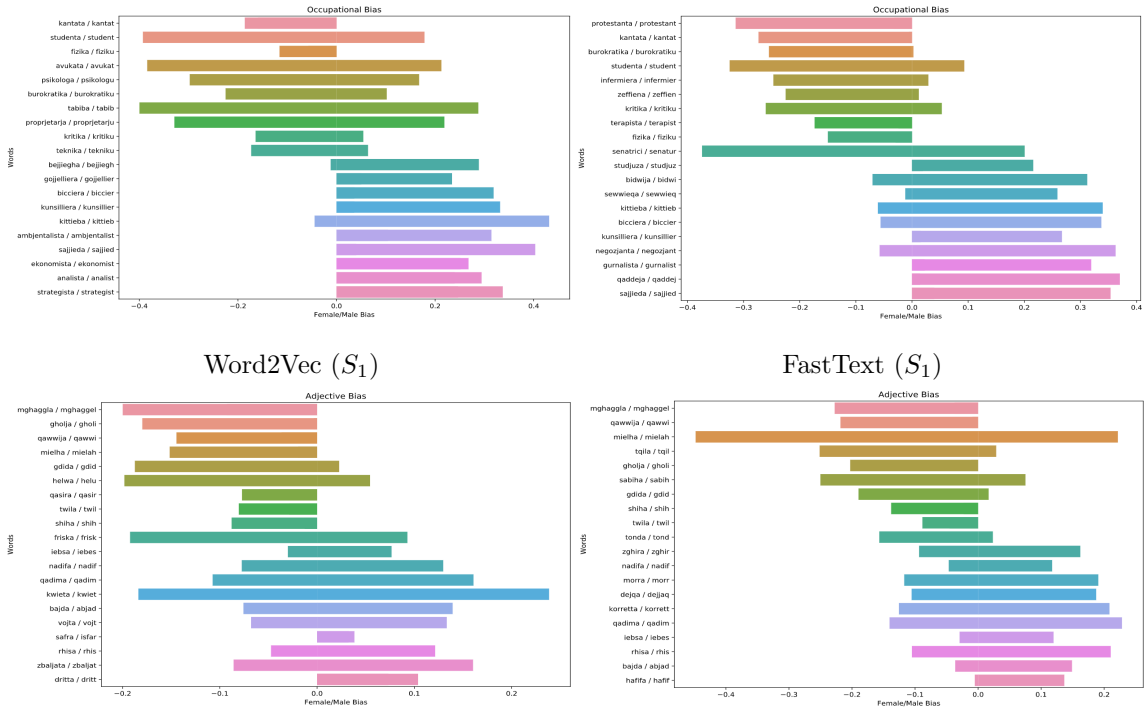
#### 4.1.1 A small note on the TF-IDF Vectorizer model

We were unable to use the TF-IDF Vectorizer model to generate word analogies. In general, the model does not attempt to capture semantic information about individual words, since it is purely a frequency-based model. In a TF-IDF matrix, each element in a word vector represents a function of its frequency in a particular document. The assumptions that i) the relative frequency of words (such as “rażel”, “mara”, “re”, u “regina”) are strongly correlated, and that ii) these words occur within each individual document, are quite a big ask for a model that simply counts word frequencies. For a wider context on why vectors created by neural models are more suitable for distributional similarity tasks, we refer the reader to [15] and the references therein.

## 4.2 Gender Bias Observation

### 4.2.1 Using $O_1$ to visualise the magnitude and direction of gender biases

Figure 3 shows the extreme gender projections of $S_1$ and $S_2$ attribute words. For example, a gender projection of $-0.4$ and 0.2 for the "studenta" / "student" word pair indicates that the female Maltese word for student is used more than its male counterpart. Interestingly, the Word2Vec ($S_1$) plot shows that nearly all occupational words extreme on the male dimension have a value of 0 on the female dimension. Given that this is not the case vice-versa, this might indicate that there is indeed a level of semantic gender bias within the embeddings. The opposite is true for the Word2Vec ($S_2$) plot.




Word2Vec ($S_1$) FastText ($S_1$)

Word2Vec ($S_2$) FastText ($S_2$)

Figure 3: Gender projections on $S_1$ and $S_2$ attribute words.

#### 4.2.2 Using $O_2$ to tabulate average bias

Table 1 features the average gender bias (as determined by the MWEAT) for the $S_1$ and $S_2$ attribute words.

| | Word2Vec | FastText |
|---|---|---|
| $S_1$ | 0.590 | 0.567 |
| $S_2$ | 0.242 | 0.379 |

Table 1: Average gender bias before debiasing.

### 4.3 Gender Bias Mitigation

Table 2 outlines the average bias (as determined by the MWEAT) after debiasing the models using the mentioned methods in Section 3.2.

| | Word2Vec ($S_1$) | Word2Vec ($S_2$) | FastText($S_1$) | FastText($S_2$) | Avg. bias reduction |
|---|---|---|---|---|---|
| $M_1$ | 0.049 | 0.084 | 0.054 | 0.090 | 82.8% |
| $M_2$ | 0.305 | 0.239 | 0.169 | 0.156 | 48.4% |
| $M_*$ | 0.059 | 0.071 | 0.047 | 0.057 | 85.5% |

Table 2: Average gender bias after debiasing.

To determine whether the word embeddings maintained core semantic relationships after debiasing, the evaluation in Section 4.1 was repeated. The performance of the $M_1$ and $M_*$ debiasing methods was very poor, both generating approximately 30 analogies using both models with an $\epsilon$-value of 0.01. Figure 4 plots the number of generated word analogies for the Word2Vec and FastText models, after debiasing using the $M_2$ method.

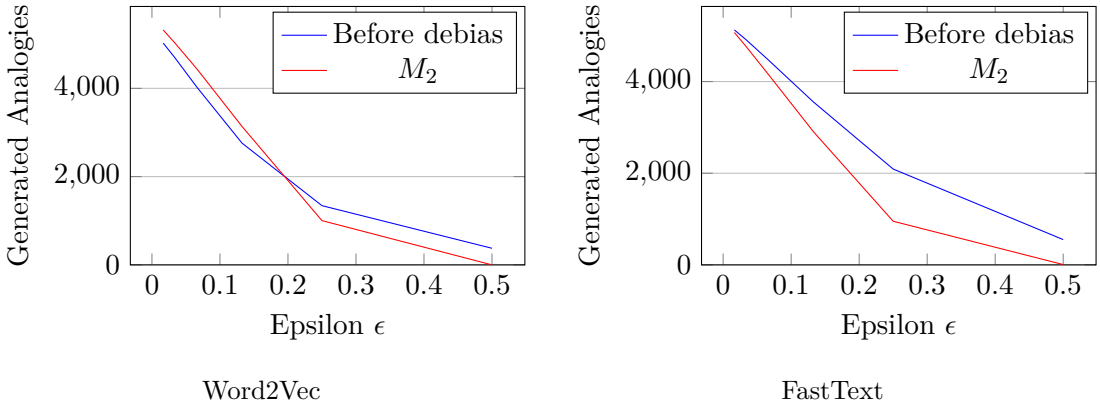



Figure 4: Number of generated analogies after debiasing using $M_2$.

## 5 Conclusion

The discussed aims and objectives have been satisfied. Word2Vec and FastText models were trained on the MLRS corpus. Experimentation was also conducted using a traditional model, based on counts instead of predictions. It was noted that the FastText model generated more analogies than the Word2Vec model, using the same set of training parameters. Despite Maltese being a gender-marked language, we comprehensively utilise two observation methods to quantify and visualise the magnitude of gender bias within the embeddings. Given a set of occupational word pairs, we found that the words "kantanta", "studenta" and "fizika" were strongly associated with the female dimension, while the words "strategist", "analist", and "ekonomist" were strongly associated with the male dimension (using the Word2Vec model). To mitigate against this gender bias, we adapt the Hard Debias method from Bolukbasi et al. and experiment with using a lemmatised dataset-equivalent. Our findings indicate that although the former method had an average bias reduction of 82.8%, valuable semantic relationships were lost. In fact, the word embeddings had failed to generate more than 30 word analogies. Therefore, the best gender mitigation approach was to utilise a lemmatised dataset instead. This had an average gender bias reduction of nearly 50%, generating around 1,000 word analogies at an $\epsilon$-value of 0.25. Gender bias

mitigation on gender-marked languages comes with several challenges, namely stemming from the requirement of differentiating between grammatical and semantic gender. However, gender-marked languages also provide insight that gender-neutral languages, such as English, do not. As seen from Section 4.2.1, the bias direction of each word pair can be easily recognised. A limitation of this study is that it does not explore gender-marked words which associate themselves with only one gender (such as "saħħa" and "xogħol"). The code used in this study is available at `https://github.com/wendru18/gender-debias`.

# 6 References


[1] Debasis Ganguly et al. "Word embedding based generalized language model for information retrieval". In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 2015, pp. 795–798.

[2] Lihao Ge and Teng-Sheng Moh. "Improving text classification with word embedding". In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE. 2017, pp. 1796–1805.

[3] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013, pp. 746–751.

[4] Herbert Rubenstein and John B Goodenough. "Contextual correlates of synonymy". In: *Communications of the ACM* 8.10 (1965), pp. 627–633.

[5] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *arXiv preprint arXiv:1607.06520* (2016).

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[7] Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. "How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?" In: *arXiv preprint arXiv:1910.14161* (2019).

[8] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[9] Armand Joulin et al. "Fasttext. zip: Compressing text classification models". In: *arXiv preprint arXiv:1612.03651* (2016).

[10] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. "Harvesting implicit group attitudes and beliefs from a demonstration web site." In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), p. 101.

[11] Peter Glick and Susan T Fiske. "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism." In: *Journal of personality and social psychology* 70.3 (1996), p. 491.

[12] Ben Schmidt. "Rejecting the gender binary: a vector-space operation". In: *Ben's Bookworm Blog* (2015).

[13] Pei Zhou et al. "Examining gender bias in languages with grammatical gender". In: *arXiv preprint arXiv:1909.02224* (2019).

[14] Hila Gonen and Yoav Goldberg. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them". In: *arXiv preprint arXiv:1903.03862* (2019).

[15] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 238–247.