# Gender Bias in Historical Word Embeddings

**Andy Lee**
**Department of Computer Science, University of Toronto, Toronto, Canada**
andy.lee@cs.utoronto.edu

## Abstract

English gender-neutral words are often imbued with gender connotations that were passed down through generations. These biases are often embedded into historical text, and although we are aware of its existence, it has been difficult to observe and quantify. Meanwhile, Word2Vec relies on corpora to train its models, a means to manipulate text as vectors in high dimensional space, has been an invaluable tool in the field of natural language processing. However, multiple studies have shown that biasness and stereotypes were also picked up by word embedding models on supposedly gender-neutral words from the historical text that *taught* them. In this study, we replicated a gender scoring method to quantify gender bias in word embeddings, then we repeated the examination for every decade to confirm the existence of gender bias in historical word embeddings.

## 1 Introduction

Word2Vec is a natural language processing technique that was invented in the past decade, but it already has profound implications on how we map, understand, and manipulate words in natural language processing. Large amounts of corpora are tokenized and fed into neural networks to generate a multi-dimensional representation of each word, resulting to word embeddings. Words of the same meaning tend to cluster together with its vectorized position forming some sort of relationship with its neighbours (Mikolov et al., 2013). However, many researchers identified through pairing certain words or analogies that the results returned often involve hints of biasness and stereotypes in them (Garg et al., 2018). Naturally, the biasness only existed because the corpora used by the model was written by humans. Nevertheless, some modern applications utilizing these word embeddings might be sensitive to this biasness and recognizing this phenomenon is the precursor to applying appropriate countermeasures. Few research has already suggested methods to debias words in word embeddings, which often involve correcting the vector's position to midway between two genders regions (Zhao et al., 2018). However, this study will not be debiasing words. Instead, we will focus on trying different methods to quantify gender biasness and track its trends in every decade using pretrained historical word embeddings of the last two centuries. This paper is broken into the following parts. In *Computation Methodology*, we will discuss the dataset used, the curated list of words to test, and how the gender scoring system works. In *Validation*, we will describe a simple method to validate the gender scoring system. Next, *Results* will present the findings and some observations. In *The Missing French*, we will discuss hurdles that were encountered. In *Limitations*, we will offer some perspectives of the study and potential countermeasures to various issues. At last, we will offer our conclusion to the study.

## 2 Related Work

The focus of this study is to identify gender bias in word embeddings (Mikolov et al., 2013). The inspiration was drawn from multiple prior studies about the existence of such biasness and offered various methods to address them (Bolukbasi et al., 2016; Garg et al., 2018; Zhao et al., 2018). In our study, we will attempt to explore a systematic method to observe the degree of biasness of words in different periods of time but will not be attempting any debiasing.

## 3 Computational Methodology

The methodology is straightforward, and we will cover each part in more details. In a high-level overview, the process includes identifying historical word embedding dataset, curating lists of words to observe, identify the gender specific regions, calculate a gender-bias score for each word, fit and find the regression line based on the scores, and observe the historical trends.

### 3.1 Word Embedding Dataset

The dataset chosen for this study was downloaded from Stanford University's NLP page on HistWords: Word Embeddings for Historical Text (Hamilton et al., n.d.). These word-embeddings have already been pre-trained using Google Ngram (Google Ngram Viewer, n.d.) and were ready to use.

There are various language packages to choose from, and for our study, we downloaded the English and French package. Each package consists of 40 files—two files (a pkl and an npy file) per decade, from 1800 to 1990. For each decade, the pkl file stores a list of vocabulary that is to be used in tandem with the npy file which stores the corresponding word vector. Since we need to repeatedly access the vectors for each word and each decade, a function was built to preprocess, transform, and combined the npy and pkl files into Gensim models—one model per decade, to improve processing performance.

### 3.2 Curating Lists of Observation Words

We needed to curate a list of words to observe the diachronic change in gender bias spanning two centuries. To limit the scope of this study, we chose words only from the following categories: occupational and professional nouns such as "doctor" and "nurse"; positive and negative adjectives such as "active" and "lazy", and appearance adjectives such as "aggressive" and "adorable". Because our collection of word embeddings spanned two centuries, we aimed to curate nouns that have existed as far as the $19^{th}$ century, and for that reason, we consulted the census archives of the United States between $19^{th}$ and $20^{th}$ century (United States Census Publications, n.d.). The census report appeared to be scanned from printed government publications at the time, hence the optical character recognition often failed to accurately detect all characters and must be manually typed out.

The census was collected from persons ten years or older, who were at the time engaged in an occupation or profession. In the census, each row listed the main occupational and professional noun such as "Engineers", then sometimes further broken down into subcategories such as "Engineer (civil)", "Engineers mechanic", etc. From this list, we included occupation nouns that is only a single word and have at least 50,000 census count by either male, female, or both combined. Also, we handpicked a few extra nouns that did not satisfy the previous threshold due to historically a lower census count but rose to prominent as we know today in our modern society, such as "actor" and "driver".

Some occupation and profession nouns evolve over time, such as "typist" nowadays used to be commonly called a "typewriter". For this study, we did not align two semantically equivalent nouns but considered them independently—more about this in the discussion section.

As for positive and negative adjectives, we obtained those lists from various online sources (Grammarhere, 2019; Gunner, n.d.), and manually removed any adjectives that might have conflicting senses, such as "hot" and "light". We suspect multi-senses, especially spanning multiple parts of speech, will affect our results therefore pre-emptively omitted them from the list.

### 3.3 Identify the gender specific regions

To calculate whether a word falls on either gender regions, we need to first be able to distinguish the two gender regions or find where the border between the two gender-region lies. Various works were consulted, and methods were attempted, which includes but not limited to attempts using principal component analysis to reduce the word embeddings into its bare gender dimensionality (Monkey, 2020), simple cosine-similarity and compare each word against the gender vector (Banerjee, 2018; Berhane, n.d.), and *k*-means to identify the gender clustering (owygs156, 2017). While some methods showed some results, the main challenge was to accurately isolate the gender feature within the multidimension. Ultimately, we followed a method inspired by Chanin (2021) which produced a definitive gender score suitable for our study.

### 3.4 Scoring each word

First, we need to identify a small list of gendered pair-words which will help us calculate where the gender vector lies. Ideally, the gender vector should be composed of words that are gender definitive with low ambiguity in their respective gender regions. From *Language, Data, and Knowledge* (Bond et al., 2017, p. 361), we handpicked five gender-pairs that have high word frequency based on Fig. (1a), and low similarity score based on Fig. (1b). The five-pairs chosen were: wife-husband, her-him, herself-himself, mother-father, and queen-king. Equation (1) is the formula for calculating the gender vector. For each pair, $\vec{f}$ is the word embedding vector for the feminine word, $\vec{m}$ is the word embedding vector for the masculine word, $n$ is the total number of gender-pairs which is five pairs for English, and $i$ is the index for each gender-pair. This formula calculates the mean difference of each gender pair denoted as $\overrightarrow{G_d}$, where $d$ is the decade index—since each decade produces its own gender vector.

$$\overrightarrow{G_d} = \frac{1}{n}\sum_{i=1}^{n} \vec{f_\iota} - \vec{m_\iota} \tag{1}$$

Next, we need to find various projection values. Equation (2) shows the formula for calculating the mean projection of the feminine word for each gender word-pair on $\overrightarrow{G_d}$, and denoted as $P_f$. Equation (3) is a similar function but for the masculine word of the gender word-pair denoted as $P_m$. Equation (4) calculates the mean projection of $P_f$ and $P_m$.

$$P_f = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{f_\iota} \cdot \overrightarrow{G_d}\right) \tag{2}$$

$$P_m = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{m_\iota} \cdot \overrightarrow{G_d}\right) \tag{3}$$

$$P_\mu = \frac{1}{2}\left(P_f + P_m\right) \tag{4}$$

To get a score for each word in the curated list, we calculate the projection against the gender vector, as depicted in Eq. (5). Finally, we have all the pieces to formulate the score equation in Eq. (6), where the numerator calculates projection against the mean gender projection, and the denominator normalizes the value scaled to the difference of the feminine and masculine projections. The result is $S_{w_i}$ where $S$ is the gender score of word $w_i$, positive value means the word is in the feminine region, negative value means the word is in the masculine region, and values $\pm 1$ is one unit of feminine and masculine projection difference away from $P_\mu$.

$$P_{w_i} = \vec{w_\iota} \cdot \overrightarrow{G_d} \tag{5}$$

$$S_{w_i} = \frac{2\left(P_{w_i} - P_\mu\right)}{P_f - P_m} \tag{6}$$

Now that we have defined the scoring system, we can scale up by processing each word for each decade. Algorithm (1) shows the pseudocode of the process utilizing the gender scoring system.

To make it easier to visualize diachronic change, unobstructed, we added a function depicted in Algorithm (1) to purge any words that do not appear in each word embedding models for a minimum consecutive number of decades. This constraint helped simplify building the trend regression line, and it helped produce a smoother and uninterrupted time-series graph. For this study, we set our minimum consecutive decade thresholds to be 7, 14, and 20. Seven decades set us into a period before and after World War 2, 14 decades into machine and technology revolution, and 20 decades into the industrial revolution. Each era arguably held significant influence on the distribution of occupation and profession, and those effects should extend to our gender score.

The source code is available at Github: https://github.com/danieljai/CSC2611H/blob/main/final.ipynb

## 4 Validation

To validate the robustness of the scoring system in Eq. (6), we curated a small list of gender specific nouns from online sources (Gender: Masculine and Feminine Nouns, n.d.) to validate whether the scores adhere to our expectations—that feminine nouns result to a positive value, and masculine nouns result to a negative value. Ten words were enlisted to conduct this validation, and the results in Table 1) shows that the scoring system did adhere to our expectations. However, we noticed the slope does not always trend outwards deeper

**Algorithm 1:** Get score from curated list

```
test_word_list = curated list of words to be tested
filtered_word_list = [empty list]
tally_score = [empty list]
consecutive_threshold = [7, 14, 20] decades
for each word in test_word_list
    if max(consecutive_decade_appearance(word)) > consecutive_threshold
        append word to filtered_word_list
for each decade_model in models_1800_1990
    for each word in filtered_word_list
        score = get_score(word, decade_model)
        append score to tally_score
return tally_score
```

into the gender region as seen in Fig (1). For example, some words such as “mistress” and “duke” have been trending inwards towards gender neutral. We were unable to offer any explanation.

## 5 Results

As explained in detail earlier, we curated three sets of lists: Occupation and Profession nouns, Positive and Negative Adjectives, and Appearance Adjectives, to conduct analysis in three different thresholds. To qualify, each word needed to appear 7, 14, and 20 decades consecutively in all word embedding models.

Of the 183 Occupation and Profession nouns curated, 58, 47, and 28 words met the minimum 7-, 14-, and 20-decades criteria respectively; 30, 20, and 12 of 36 Positive and Negative Adjectives met the respective decade criteria; and 68, 63, and 48 of 91 Appearance Adjectives met their respective decade criteria. As the number of consecutive decade requirement increased, the number of words that failed to meet the criteria also increased, which resulted to decrease in qualifying words. Table 2) shows a complete overview of the steepest regression slope. Each curated list is categorized into three major columns, then further split into sub-columns of word, slope value, and mean gender score. The mean gender score is the mean of scores from qualifying decades. Each row is a word, grouped into one of the three consecutive decade criterion. The list should be read from top to bottom as the consecutive decade criterion increases. Any rows that are highlighted in a blue shade means the word is seen for the first time, and the opposite means that word has already appeared previously. For example, in the column occupation and profession, “printers” appeared in all three criteria, therefore only the entry in “minimum 7 consecutive decades” was shaded blue.

| word | slope | mean_score |
|---|---|---|
| grandmother | 0.00217 | 0.69969 |
| grandfather | -0.00021 | -0.38267 |
| god | 0.00114 | -1.11670 |
| goddess | 0.00130 | 0.27071 |
| master | 0.00017 | -0.75256 |
| mistress | -0.00173 | 0.47220 |
| nephew | 0.00128 | -0.47777 |
| niece | -0.00036 | 0.47701 |
| duke | 0.00172 | -0.46748 |
| duchess | 0.00197 | 0.61440 |

Table 1: Validation with gendered words

We can quickly deduce that when the slope and mean score values are in the opposite sign (positive and negative), that means the word is trending towards gender neutral. Contrarily, when the slope and mean score values share the same sign, that means the word is trending deeper into the sign’s gender region. With this in mind, we can draw a few observations. In Table 2), there are 22 unique occupation and profession nouns listed, and all their mean score were negative, which means it resided mostly in the masculine region, and most words are trending towards gender neutral, except for “merchants”, “officer”, and “physician”. Figure (2) shows the top occupation and profession nouns trending towards the neutral line. This suggests a strong bias shift towards gender neutralizing or balancing on occupations and professions that were historically and predominantly held by males.

Almost a similar pattern can be observed in positive and negative adjectives. In Table (2), there are 18 unique adjectives listed, and only two words’ mean score reside in the feminine region. While most words are trending towards gender neutral, “lazy” and “selfish” are trending deeper into the masculine region. Another observation we

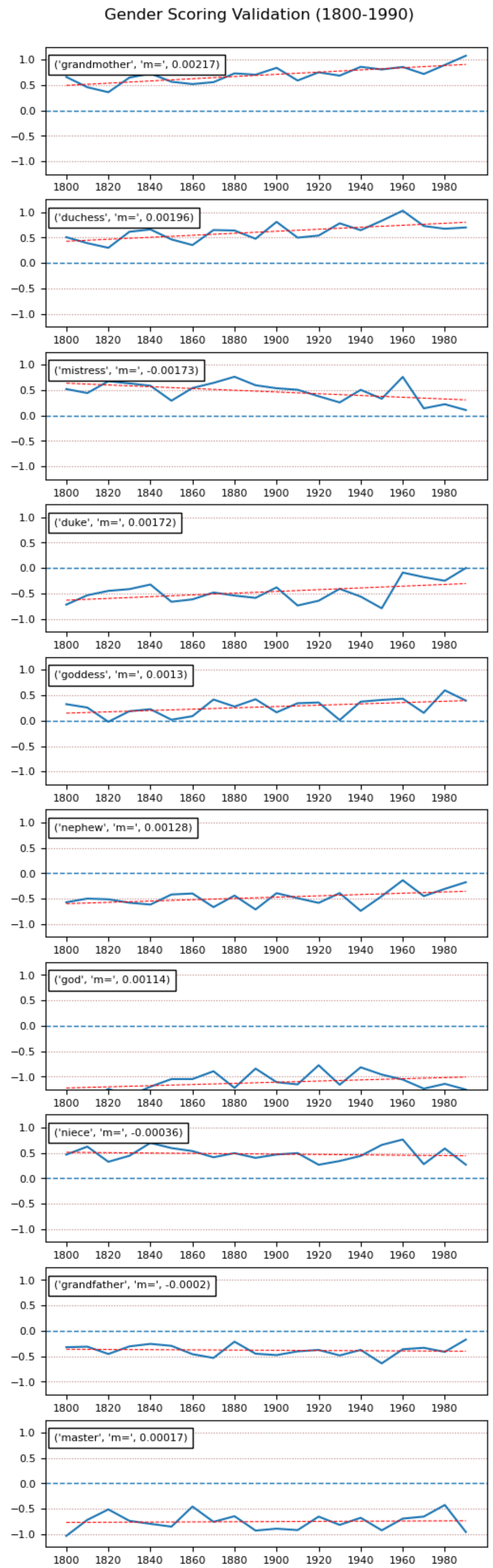


Figure 1: Graph for validation with gendered *words*

can draw from is that the slope values in this category is much lower than the other two. This suggests the trend to gender neutral happens at a much slower pace for behavioural adjectives.

Appearance Adjective in Table (2) exhibited a different pattern compared to the other two categories. Of the 15 unique words listed, only six were in the masculine region. Also, only six words are trending towards gender neutral, while the rest are trending deeper into their respective gender regions. This suggests a stronger reluctance in debiasing on gender-neutral words that describe common physical appearance. We can further examine this phenomenon by referring to Fig (3). Words like “gleaming” and “slender” started near the gender-neutral line, however in the past two centuries, it trended more deeply into the feminine region. The update-to-date sense of either word remained gender-neutral, but the provided dictionary example reads “her slender neck” which suggest gender connotations in adjectives can extend beyond word sense and lie at how the word is commonly used in a sentence.

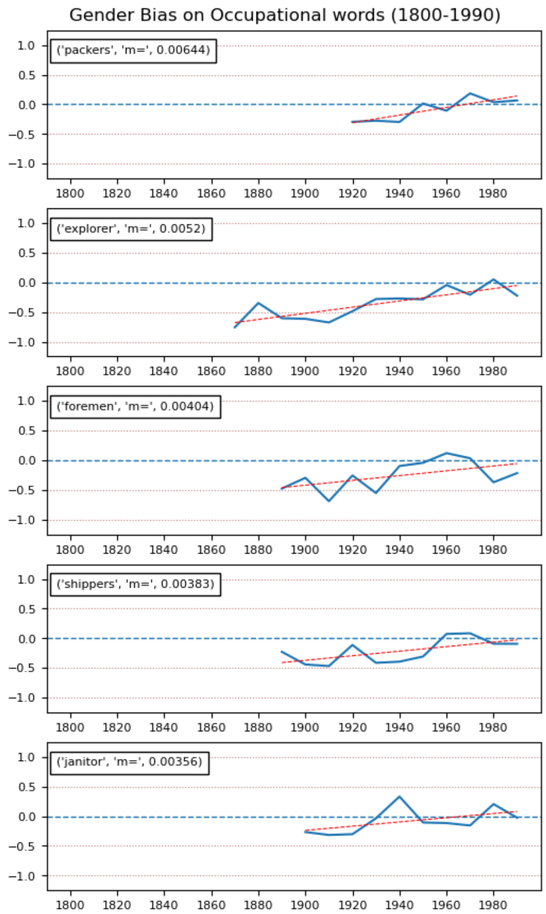


Figure 2: Graph for occupation and profession nouns set to 7 consecutive decade threshold

| | Occupation and Profession Nouns | | | Positive and Negative Adjectives | | | Appearance Adjectives | | |
|---|---|---|---|---|---|---|---|---|---|
| | word | slope | mean_score (pos=fem, neg=mas) | word | slope | mean_score (pos=fem, neg=mas) | word | slope | mean_score (pos=fem, neg=mas) |
| Min. 7 consecutive decades | packers | 0.0064 | -0.0835 | resourceful | 0.0048 | -0.2782 | cute | 0.0062 | -0.0062 |
| | explorer | 0.0052 | -0.3659 | pessimistic | 0.0043 | -0.1529 | muscular | 0.0040 | 0.0812 |
| | foremen | 0.0040 | -0.2579 | argumentative | 0.0026 | -0.2863 | gleaming | 0.0034 | 0.2219 |
| | shippers | 0.0038 | -0.2178 | gregarious | 0.0025 | -0.0836 | elegant | 0.0031 | 0.0509 |
| | janitor | 0.0036 | -0.0786 | inconsiderate | 0.0023 | -0.2697 | slender | 0.0029 | 0.3203 |
| | helpers | 0.0032 | -0.3614 | hysterical | -0.0016 | 0.3142 | aggressive | 0.0028 | -0.3537 |
| | bakers | 0.0029 | -0.0768 | childish | 0.0015 | 0.1189 | alert | 0.0028 | -0.2297 |
| | printers | 0.0029 | -0.1376 | optimistic | 0.0014 | -0.1809 | precious | 0.0027 | -0.0612 |
| | miners | 0.0027 | -0.1900 | active | 0.0012 | -0.3428 | bright | 0.0027 | 0.1967 |
| | teacher | 0.0024 | -0.4342 | weak | 0.0011 | -0.1666 | drab | 0.0027 | 0.1212 |
| Min. 14 consecutive decades | helpers | 0.0032 | -0.3614 | inconsiderate | 0.0023 | -0.2697 | muscular | 0.0040 | 0.0812 |
| | bakers | 0.0029 | -0.0768 | hysterical | -0.0016 | 0.3142 | gleaming | 0.0034 | 0.2219 |
| | printers | 0.0029 | -0.1376 | childish | 0.0015 | 0.1189 | elegant | 0.0031 | 0.0509 |
| | miners | 0.0027 | -0.1900 | active | 0.0012 | -0.3428 | slender | 0.0029 | 0.3203 |
| | teacher | 0.0024 | -0.4342 | weak | 0.0011 | -0.1666 | aggressive | 0.0028 | -0.3537 |
| | geologist | 0.0024 | -0.3580 | enthusiastic | 0.0010 | -0.3383 | alert | 0.0028 | -0.2297 |
| | shoemakers | 0.0020 | -0.0687 | sincere | 0.0010 | -0.4129 | precious | 0.0027 | -0.0612 |
| | merchants | -0.0020 | -0.4577 | cheerful | 0.0010 | -0.0451 | bright | 0.0027 | 0.1967 |
| | butchers | 0.0018 | -0.2085 | lazy | -0.0009 | -0.1931 | excited | 0.0026 | -0.2888 |
| | innkeeper | 0.0018 | -0.2931 | moody | 0.0007 | -0.1236 | feeble | 0.0025 | 0.0517 |
| Appeared in all decades | printers | 0.0029 | -0.1376 | childish | 0.0015 | 0.1189 | muscular | 0.0040 | 0.0812 |
| | teacher | 0.0024 | -0.4342 | active | 0.0012 | -0.3428 | elegant | 0.0031 | 0.0509 |
| | merchants | -0.0020 | -0.4577 | weak | 0.0011 | -0.1666 | slender | 0.0029 | 0.3203 |
| | editor | 0.0017 | -0.4675 | enthusiastic | 0.0010 | -0.3383 | precious | 0.0027 | -0.0612 |
| | messengers | 0.0016 | -0.6854 | sincere | 0.0010 | -0.4129 | bright | 0.0027 | 0.1967 |
| | musician | 0.0015 | -0.3126 | cheerful | 0.0010 | -0.0451 | excited | 0.0026 | -0.2888 |
| | officer | -0.0011 | -0.8692 | lazy | -0.0009 | -0.1931 | feeble | 0.0025 | 0.0517 |
| | physician | -0.0010 | -0.4057 | affable | 0.0006 | -0.2344 | dark | 0.0024 | 0.0550 |
| | overseers | -0.0010 | -0.4838 | adventurous | 0.0006 | -0.4110 | distinct | 0.0023 | -0.4437 |
| | engineers | 0.0010 | -0.4807 | selfish | -0.0005 | -0.3407 | sparkling | 0.0022 | 0.3868 |

Table 2: 10 steepest gender bias shifts in three categories

In three categories, we observed that lower consecutive-decades criteria produce steeper slope values. The average absolute slope value for Occupation and Profession Nouns for the 7-, 14-, and 20-consecutive decades were 0.0037, 0.0024, and 0.0016; Positive and Negative Adjective were 0.0023, 0.0012, and 0.0009; and Appearance Adjectives were 0.0033, 0.0030, and 0.0027. Although the criterion was defined as *minimum* instead of *absolute*, words with shorter consecutive-decade span account for most of the steepest slopes as illustrated in Fig. (2). This, although inconclusively, suggests that gender bias change happens quickly and more aggressively in recent decades.

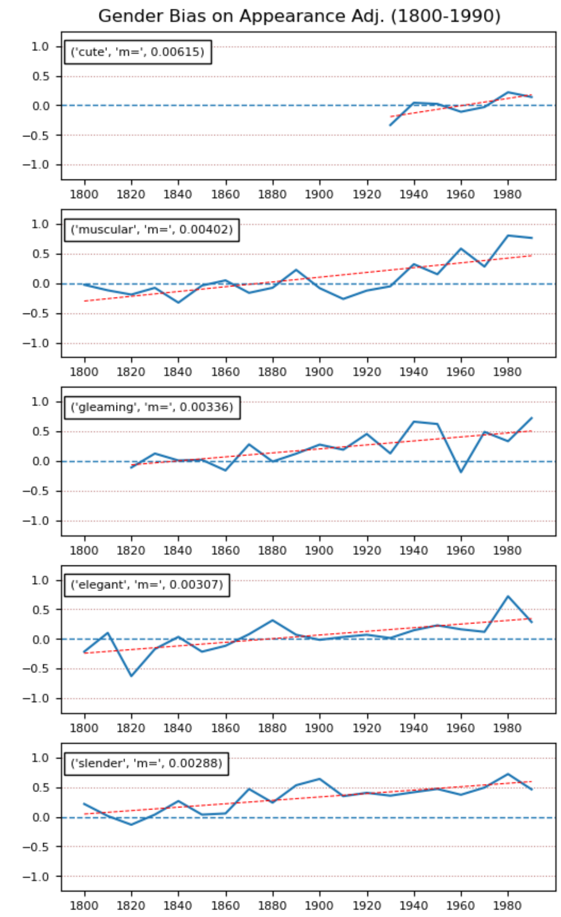


Figure 3: Graph for appearance adjective set to 20 consecutive decade threshold

## 6 The Missing French

After gathering a list of top gender-biased words from the three categories, we were interested in whether the same set of words exhibit any gender bias effects in French. We took the top five words from each of the three categories and manually translated the words to French using DeepL (DeepL Translate, n.d.). Immediately, we realized each English word often result to more than one French translations—different forms of the word such as feminine, masculine, feminine plural, and masculine plural. We were unable to confirm the translations are semantically equivalent to the English counterpart, but nevertheless, we picked one of each from both genders, and compiled a list of potential French words for our test.

Just like our English model, we performed the same pre-cleaning and converted the HistWords files into Gensim format. However, before we run the words in our gender scoring algorithm, we experienced an issue. Most translated French words did not exist inside any or all our word embedding models. We took a few more alternative translations taken from DeepL without much success, as we were unable to get any meaningful consecutive-decade count. This finding is surprising, as we did not expect a common word such as "elegant" which yielded three translations that none appeared in any of our French models.

We were unable to proceed with the analysis as planned given this limitation, therefore, we changed directions to perform gender scoring on curated occupation and profession nouns that are gendered in French instead. A list of 57 feminine and masculine nouns were sourced online (lkl, 2018), then the consecutive-decade criterion filtered the list down to 22 nouns, and the gender scoring function was applied to the remaining list. The top 10 steepest slope was extracted and translated back to English using DeepL.

Table 3) shows the list of slopes and mean scores for the list of French occupation and profession nouns. Like the English results, majority, if not all, of the nouns had a negative mean score, suggesting a bias towards masculinity. However, this can be attributed to more masculine nouns than feminine in the French language, as some words lack a feminine counterpart.

The two words "enseignant" and "prof" gets translated to the same word "teacher" is a testimony to the difficulties we experienced trying

| | English word | French word | slope | mean_score |
|---|---|---|---|---|
| Min. 7 consecutive decades | carpenter | charpentier | 0.0020 | -0.1286 |
| | civil servant | fonctionnaire | 0.0018 | -0.3092 |
| | teacher | enseignant | 0.0016 | -0.0885 |
| | lawyer | avocat | 0.0016 | -0.1889 |
| | employer | employeur | 0.0011 | -0.1113 |
| | insurer | assureur | -0.0010 | -0.2259 |
| | teacher | prof | -0.0010 | -0.2039 |
| | manager | chef | -0.0009 | -0.4271 |
| | scientist | scientifique | 0.0009 | -0.3435 |
| | baker | boulanger | -0.0007 | -0.0223 |
| Min. 14 | govemor | gouverneur | -0.0005 | -0.2121 |
| | actor | acteur | 0.0005 | -0.3387 |
| | painter | peintre | 0.0005 | -0.3391 |
| All | author | auteur | -0.0004 | -0.2432 |

Table 3: Top 10 steepest gender bias shifts in three categories in French

to get French equivalents from English words that many words shared the same sense.

However, an observation that can be drawn from Table 3), is that assuming the English translation is accurate, these nouns is what we would determine as gender neutral. Based on the slope value, the change in gender is much less drastic compared to what was observed in English occupation and profession nouns. Unfortunately, our understand of the French language is extremely limited, therefore we were unable to draw any more inferences.

## 7 Discussions and Limitations

There are several limitations to this study that we foresaw but were unable to tackle due to various reasons.

### 7.1 Aligning Semantic Equivalence

Occupation and profession nouns evolve over time, and ideally, semantically equivalent nouns should be aggregated. For example, historically one who operates a typewriter machine is also called a “typewriter”, but that noun has evolved into “typist”. This would require lots of resource into researching historical context which we were unable to devote, therefore we treated those nouns, including singularly and plurality, separately.

### 7.2 Validation by staggering five decades

In our proposal, we planned to validate the robustness of the gender scoring system, by staggering the word embeddings to start five years later. Which means the initial study will use models from 1800-1810, 1810-1820…1980-1990, and the validation will use models from 1805-1815, 1815-1825…1985-1995. Unfortunately, HistWord dataset did not have such capabilities to stagger the dataset. To validate as planned, we would have to rebuild word embeddings from scratch using Google Ngram, which was not feasible. Therefore, we changed the validation to using scoring definitive gendered words instead to make sure the score produces expected results.

### 7.3 Gender-Pair Selection

Gender-pairs played an extremely important role in the scoring system, as part of the gender scoring equation is to have words projected onto the mean gender vector. Based on our validation test, more gender pairs did not always lead to better results, but on the contrary, it added more noise to the gender vector. In French, we were only able to find three concrete gender-pairs, because many gender-pair words did not appear in every decade of word embeddings, and secondly, gender pairs that were translated from English were producing too many synonyms in French. Without ground knowledge of the French language, it was not possible for us to choose the best suitable gender-pairs.

## 8 Conclusion

The gender scoring formula provided us a means of quantifying gender bias in words from historical word embeddings spanning two decades. From data and visuals results, we were able observe and make several inferences into gender bias shifts, often slowly neutralizing or rebalancing, but in some cases further into the gender regions. Nevertheless, our data shows that gender biasness existed historically especially in occupation and profession nouns where there appears to be rapid correction in recent decades. There were hurdles throughout the project, and many methods often did not produce the result we expected. Should the study continue, certain procedures that were improvised can be further refined. Since we already establish how to identify gender bias in words, the next step would be how to properly and, more importantly, when is it appropriate to debias gender-biased words.

## 9 Reference


Banerjee, P. (2018, November 30). Word Embedding Biases—Basic Neutralization Equalization. https://github.com/panchambanerjee/Word_embedding_biases/blob/d89929464e840f687a801b78da0569d9e6d027b9/Basic_neutralization_equalization.ipynb (Original work published 2018)

Berhane, F. (n.d.). Operations on word vectors. Retrieved December 6, 2022, from https://datascience-enthusiast.com/DL/Operations_on_word_vectors.html

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (arXiv:1607.06520). arXiv. http://arxiv.org/abs/1607.06520

Bond, F., Buitelaar, P., Chiarcos, C., Gracia, J., Hellmann, S., & McCrae, J. P. (Eds.). (2017). Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings (1st ed. 2017). Springer

International Publishing: Imprint: Springer. https://doi.org/10.1007/978-3-319-59888-8

David, C. (2021, June 10). Exploring Gender Bias in Word2Vec. Chanind.Github.Io. https://chanind.github.io/nlp/2021/06/10/word2vec-gender-bias.html

DeepL Translate. (n.d.). Retrieved December 7, 2022, from https://www.DeepL.com/translator

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

Gender: Masculine and Feminine Nouns. (n.d.). Myenglishpages.Com. Retrieved December 6, 2022, from https://www.myenglishpages.com/english/grammar-lesson-masculine-feminine.php

Google Ngram Viewer. (n.d.). Google Books. Retrieved December 9, 2022, from http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Grammarhere. (2019, September 14). Appearance Adjectives List in English. English Grammar Here. https://englishgrammarhere.com/adjectives/appearance-adjectives-list-in-english/

Gunner, J. (n.d.). 50 Adjectives To Describe Personality: Find the Most Fitting Word. Retrieved December 6, 2022, from https://grammar.yourdictionary.com/parts-of-speech/adjectives/personal-adjective.html

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (n.d.). HistWords: Word Embeddings for Historical Text. https://nlp.stanford.edu/projects/histwords/

lkl. (2018, September 20). French Professions—Lawless French Vocabulary—Métiers. Lawless French. https://www.lawlessfrench.com/vocabulary/professions/

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. https://doi.org/10.48550/ARXIV.1310.4546

Monkey, C. (2020, October 3). PCA Dimensionality Reduction and Word Vectors. Neurotic Networking. https://necromuralist.github.io/Neurotic-Networking/posts/nlp/pca-dimensionality-reduction-and-word-vectors/

owygs156. (2017, December 7). K Means Clustering Example with Word2Vec in Data Mining or Machine Learning. Text Analytics Techniques. https://ai.intelligentonlinetools.com/ml/k-means-clustering-example-word2vec/

United States Census Publications. (n.d.). Retrieved December 6, 2022, from https://www2.census.gov/library/publications/decennial/1900/occupations/

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods (arXiv:1804.06876). arXiv. http://arxiv.org/abs/1804.06876