An NLP-Based Analysis of Gendered Linguistic Styles in Writings Concerning Ecology, Nature, and Environmental Justice

Emma Virnelli **Caleb Deitch** **Harrison Chang**

Abstract

Writing styles have frequently differed across gender lines throughout history. Whether from societal stigmas, lack of opportunity for women, or a simple over-representation of men, differences in vocabulary, rhetoric, and style have existed since the inception of modern writing, particularly with regard to contemporary issues. Environmental justice is one such topic with a much stronger association with one gender (namely, women) than the other. Thus, differences in writing within the field are presumably quite pronounced and can serve as examples of broader gendered linguistic trends. It is therefore our intention to use machine learning methods to analyze writing samples to identify habits, trends, and other commonalities among samples from environmental authors of the same gender in order to quantify and visualize trends within the wider world of literature.

1 Introduction

The language we use is not neutral; our perceptions of the world are reflected in and shaped by our writing. This idea is central to ecofeminist theory. Ecofeminism highlights a connection between how society views women and how it views nature. Historically, both have been undervalued and marginalized within male-dominated cultures. This perspective is mirrored by how we frame environmental problems, which affects the solutions we prioritize and decisions we make. By limiting the subject matter to nature, we seek to explore how ecofeminist ideals translate into linguistic patterns. With this in mind, we will explore these patterns in syntax, grammar, and sentiment between men and women to see whether they can be used to predict the author of a text.

2 Literature Review

2.1 Detecting Online Sexism

Sexism is particularly difficult to detect due to its varied forms and the lack of a universal definition for categorization. To address the growing challenge of detecting sexism online, Belbachir et al. (2023) conducted a study on whether hybrid approaches (combining sentiment and sexism detection models) could outperform individual models. The results showed that combining models improved F1 scores by up to 1.63% over the best individual model, rejecting the null hypothesis of no improvement and demonstrating that model combination enhances detection. The study's methodology is relevant to our approach, as we similarly seek to demonstrate that multiple methods of detecting linguistic patterns will be able to predict the author's gender and interpret the underlying features that drive this prediction.

2.2 Identifying Linguistic Gender Bias in Electronic Medical Records

Prior research demonstrates that bias in institutional contexts can be identified through linguistic choices. Xu and Sun (2025) studied electronic medical records in China and identified four linguistic features in each record: judging, reporting, quoting, and fudging. They found that the language used in records showed different biases towards male and female patients, with language more often distancing the physician and casting doubt on the credibility of men (judging), and vague, incomplete, or unsubstantiated entries on women (fudging). Although situated in a different setting, this research provides a foundation for examining systemic bias through linguistic styles. These insights motivate the use of NLP methods to examine how linguistic styles in environmental texts may differ by gender.

2.3 Female Over-Contribution

As explored by Fontanarrosa et al. (2024), female authors contribute more to research papers published in the academic journal *Ecology* - which publishes papers on environmental research - than their representation would suggest. Using a mathematical formula to calculate contribution to a paper as a fraction - using an author's position in the author list as well as the total number of authors in the paper - a Women's Contribution Index (WCI) was calculated, representing the overall portion of work on an academic paper completed by women. From their analysis of papers from *Ecology* published in the 21st century, the authors found that, despite female authors comprising a far smaller share of the overall author pool, they contributed more to papers on average than their male counterparts (WCI > 0.5 per paper).

This paper suggests differences in overall interest or relevance in environmental topics along gender lines. Female over-representation in environmental papers might indicate that women take a greater interest in the subject, in turn suggesting differences in how female authors write about the environment compared to male authors. Thus, the over-contribution found in Fontanarrosa et al. (2024) evidences the linguistic, grammatical, and emotional differences in environmental texts between genders that we are looking for.

3 Dataset

The dataset consists of passages from books by 12 authors - 6 male and 6 female - sourced from Project Gutenberg or other online archives. Each passage consists of 5,000 words taken from the middle of a book with themes of ecology, environmental justice, ecofeminism, or conservation. The selected works date from 1854 to 2023, offering a wide historical range of environmental writing. Each passage was cleaned to remove headers, footnotes, image descriptions, and page numbers. To increase the number of observations, the passages were split into equal-sized chunks of 200 or 500 words, depending on the author, retaining information about the author and their gender.

4 Methods

4.1 Method 1: Part-of-Speech Tagging

To analyze differences in sentence structure using part-of-speech (POS) tagging, we employed three methods: logistic regression, a convolutional neural network (CNN), and DistilBERT on raw text. For our logistic regression and CNN, each chunk of text was converted to POS tags, where words were tokenized and each token was replaced with its corresponding part-of-speech label (e.g., NN, VBZ, PRP, JJ). Our goal was to identify patterns in POS tokens useful for determining the author's gender. For each method, authors were split into a training set (80%) and a test set (20%) to ensure no overlap between the training and test sets.

Logistic regression was performed on 200 and 500 token length sequences of POS tags that were vectorized using a term frequency–inverse document frequency (TF–IDF) vectorizer. The model was first run with 500 word chunks and included n-grams with window sizes 1-2.

To improve accuracy, we increased the number of observations by using 200-word chunks, resulting in 300 total observations. The TF-IDF vectorizer was tuned to use at most 30 features and to consider n-grams with window sizes 2-5. An L2 penalty was added to the logistic regression to regularize and reduce overfitting. We then fit the regularized logistic regression classifier using the SAGA solver. The model was then tested on both a validation set containing passages from the same authors and a validation set containing passages from unseen authors.

We then implemented a 1D convolutional neural network (CNN), also trained on sequences of POS tokens to learn deeper grammar patterns. To reduce overfitting, we used dropout to regularize the model and trained it with cross-entropy loss using the Adam optimizer.

To get a benchmark for prediction accuracy without information loss, we evaluate the raw text with a transformer-based classifier using DistilBERT, a lightweight pre-trained language model. Vocabulary and word features such as length are not present in earlier models but may have some predictive power. Training was conducted on chunked text sequences, which were tokenized using the model's pretrained tokenizer and padded or truncated to a fixed maximum length to ensure consistent input lengths.

4.2 Method 2: Probabilistic Context-Free Grammar Tagging

To analyze the writing beyond vocabulary, Probabilistic Context-Free Grammar Parsing (PCFG) was used to examine the authors' syntax. PCFG parsing branches off the foundation of Context-Free Parsing (CFG), which assigns grammar rules to define valid sentence structures. What makes PCFG parsing unique is that it weights each grammar rule based on how commonly it occurs, moving beyond a simple yes-or-no answer about whether it's correct (Fernando, 2025).

To facilitate gender-based comparison in the parsing analysis, the dataset was split by gender: all texts written by male authors were placed in one list, and those written by female authors in another. Each list contained all the pieces written by that gender group.

The method involves defining the grammar rules by manually entering rules and analyzing the data to tailor the vocabulary to our work. We found the top 20 most frequent words in each of the 5 categories: nouns, verbs, adjectives, adverbs, and determiners (top 9 for determiners). The high-frequency words become the vocabulary for the grammar rules due to their commonality and the likelihood that these words will appear in the sentences being parsed. The sentence structures (S = NP + VP) and function words ("in," "on," "at," "and," "or," "but") were manually selected because they primarily structure sentences rather than convey specific meaning. Therefore, the approach combines both methods: the vocabulary is derived from the actual text data, while the grammar structure is predefined. A total of 11 grammar rules were created for this method. The creation of these grammar rules provides a baseline for our comparison, enabling us to identify deviations from the norm by gender. The probabilities for these rules were informed by reference materials, particularly PCFG lecture notes from San Diego State University (Gawron), which provide established frequency distributions for English grammatical constructions. Each rule has an assigned probability. For example, among determiners, 'these' appears with probability 0.28 and 'some' with probability 0.27, while among verbs, 'being' has probability 0.14 and 'found' has probability 0.13.

After clarifying the grammar rules, each sentence was parsed. Male texts parsed through 1,622 trees with 35,812 production rules applied and female text parsed through 1,016 trees with 35,379 production rules applied. After parsing, the rules are then analyzed to extract the production rules, tree depth (to understand syntactic complexity), and the number of leaves (to understand sentence length). After that, we compared the rule frequencies between male and female text, along with tree metrics, and normalized the percentages for fair comparison.

Structural Rules:

```
S → NP VP
NP → Det N | N | Det Adj N
VP → V | V NP | V NP PP
PP → P NP
```

Lexical Rules (Top 2 Shown With Probabilities):

```
N → time [0.12], many [0.00]
V → black [0.00], being [0.14]
Adj → many [0.14], like [0.00]
Adv → out [0.00], then [0.10]
Det → these [0.28], some [0.27]
P → in [0.15], on [0.10]
Conj → and [0.70], or [0.20]
```

Figure 1: 11 grammar rules (4 sentence structure, 5 vocabulary, 2 function words) used for this analysis. Sentences (S) consist of noun phrases (NP), verb phrases (VP), and prepositional phrases (PP). The lexical rules show the top two most frequent words for each part of speech category, nouns (N), verbs (V), adjectives (Adj), adverbs (Adv), determiners (Det), prepositions (P), and conjunctions (Conj), extracted from the text corpus, with probabilities in brackets indicating the relative frequency of each word within its category.

4.3 Method 3: Emotional Flow Evaluation

One aspect of environmental writing in which we expected large differences based on author gender was the emotional sentiment in the writing. Differences in emotional understanding, usage, and acceptability are common knowledge both within literature as well as the wider world, so we expected a certain level of difference with regards to emotional usage and frequency within our text samples. We also wanted to quantify the level of difference such that it was able to be visualized and thus easily understood. Therefore, we decided to use a method of analyzing not just sentiment of writing but also the specific emotions used by an author in their writing per chunk of text, such that the frequency of individual emotions as well as their trends throughout the sample could be clearly recorded.

The model used for this method was a modified version of the original RoBERTa model created by Meta AI and modified by Sam Lowe. RoBERTa is itself a modified BERT model, trained to predict masked words from natural language sentences.

Sam Lowe's version was trained on a dataset titled *go_emotions*, a set of tens of thousands of Reddit comments all scored for emotions based on a wheel of 28 default emotions (neutrality, joy, anger, sadness, frustration, etc.). The model itself is thus able to score new text samples for emotions based on the 28 defaults. All of our text chunks were run through this method and scored as such. We then took those scores, aggregated them according to author gender, and graphed them according to sentiment per chunk of text chronologically as well as averaged across all chunks.

4.4 Method 4: LLM

We gave ChatGPT 5.2 the full labeled training data, including 120 rows and 10 authors. Then we asked it to predict the gender of the validation set, a file that contained only sample text taken from different parts of the same book. We then merged the file with the predicted labels back to the full dataset on the "sample" column to ensure labels would be aligned.Accuracy was 54.2% which is very poor compared to our models.

5 Results

5.1 POS Tagging

Across models, the highest test accuracy obtained was approximately 75%, and no single model produced strong evidence of gendered differences in linguistic structure. After tuning the parameters, logistic regression achieved only 60–70% test accuracy, which is around 20-40% better than random chance. The results also did not generalize to passages outside the original train/test data, where the model performed the same or worse than random chance.

The performance of the CNN was approximately the same as that of logistic regression, but it was also highly variable. The typical test accuracy ranged from 60% to 72%. The model performed well on validation data that contained distinct passages from the same books used in training.The accuracy dropped when it was used to predict the author's gender on unseen author. This suggests that the natural stylistic differences between authors overshadow any gendered differences in our dataset.

The DistilBERT model achieved the highest predictive performance at 80-90% test accuracy on unseen data and authors. Despite this improvement, much of the predictive power appears to come from topic or author-specific factors. The model overfits individual authors and context, which limits its ability to isolate gendered linguistic style. As a result, while DistilBERT was able to predict the authors' genders accurately, it does not provide evidence that it is possible to do so using only grammatical or stylistic structure.

5.2 PCFG Parsing

Looking at the grammar results, out of 71,191 analyzed production rule instances, only 64 showed meaningful gender differences (0.09%). The top differences were female authors using commas 2.33% more frequently, male authors using periods 1.53% more frequently, and female authors using the definite article 'the' 1.09% more frequently. While females wrote sentences that were 60% longer on average (34 vs 21 words), their actual grammatical choices were strikingly similar to male authors'. The similarity in syntactic tree depth was both averaging around 3, meaning that despite differences in length men and woman have similar sentence structure, or it is too challenging to identify.

5.3 Emotional Flow

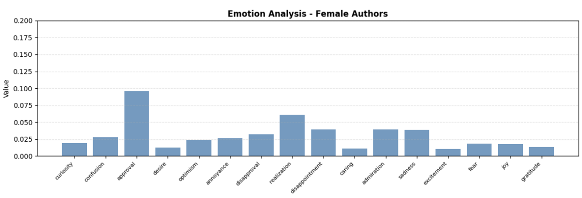

Figure 2: Frequency of emotional sentiments in texts by female authors as a percentage.

Depicted in Figure 2 is the frequency of specific emotions in texts authored by women, averaged across all samples. Arbitrary emotions such as "neutrality" or very infrequently-occurring emotions were omitted for visual purposes. As can be seen in the figure, female authors tend to communicate a wide variety of emotions throughout their texts; most notably "approval", occurring in nearly 10% of all text. Although average emotional frequency trends low, the wide range of notable emotions stands in start contract to usage by male authors.

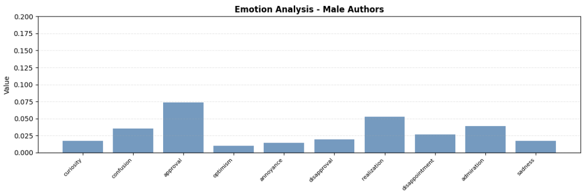

Figure 3: Frequency of emotional sentiments in texts by male authors as a percentage.

Depicted in Figure 3 is the frequency of specific emotions in texts authored by men (again averaged and cleaned). The smaller range of emotions is clearly visible in the smaller number of bars present (16 for women; 10 for men), and the emotions present on average occur less frequently in male works as opposed to female works (high value of 10% for women; 7.5% for men). Additionally, the ratio of notable negative emotions in male works (50%) is higher than the ratio in female works (37.5%). Complex emotions such as "desire", "caring", and "gratitude" that were frequent in female works were also notably absent from male works, demonstrating a gap in emotional range along gendered lines.

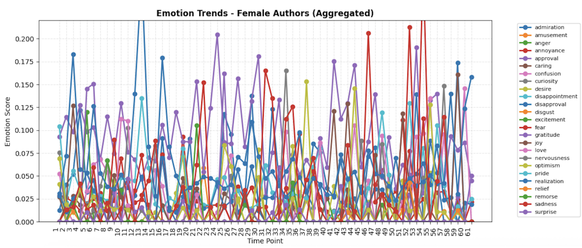

Figure 4: Emotional score per text chunk, averaged between female authors.

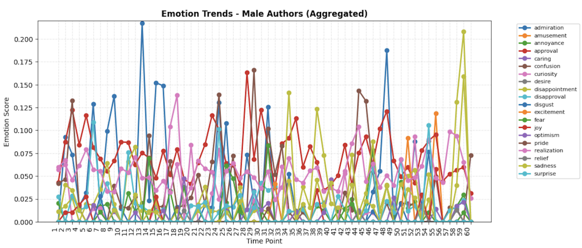

Figure 5: Emotional score per text chunk, averaged between male authors.

Figures 4 and 5 depict the aggregate emotional score per chunk of a text ordered chronologically for women and men respectively. Unfortunately, the results could not be cleaned due to notable occurrences of nearly every emotion being observed when using smaller text samples, but a larger trend is still visible here: female authors averaged a larger number of high-scoring chunks compared to men. Only two emotions ever exceeded a score of 20% in any chunk authored by a man, while five emotions exceeded that threshold among chunks authored by women. Additionally, emotional scores per chunk averaged purely higher for women than men.

6 Analysis

Our results suggest that gendered differences in environmental writing are more strongly reflected in emotional expression than in syntactic or grammatical structure. Sequences of POS tags and PCFG derived grammatical patterns did not consistently find systemic differences corresponding to the author's gender. Methods involving POS tags suffered as a result, as both logistic regression and CNN models trained on POS sequences overfitted the data, achieving higher accuracy than random chance but not generalizing across authors. Furthermore, any patterns that do exist may be masked by individual writing style, historical context, or genre conventions. The PCFG based analysis focused on syntactic probabilities, but did not find patterns.

In contrast, emotional flow analysis revealed consistent and interpretable differences. Texts authored by women exhibited a broader emotional range, higher average emotional intensity, and greater frequency of complex emotions such as approval, caring, gratitude, and desire. Texts written by men displayed a narrower range of emotions and had a higher proportion of negative emotions. These findings are relevant to ecofeminist theory, which emphasizes care and emotional engagement with nature, emotions that are featured more prominently in texts written by women.

DistilBERT's classification accuracy highlights an important methodological distinction. The training data for both DistilBert and the LLM model consisted of raw text, including information lost when focusing only on linguistic structure. It is also unlikely that it isolated any gendered linguistic style; most likely, it picked up differences corresponding to authors instead. The poor performance of the LLM classifier when asked to generalize a prediction model further illustrates the complexity of the literature. These results caution against interpreting high classification accuracy as evidence of inherent stylistic gender differences, especially when model interpretability is limited.

Limitations

Several limitations should be acknowledged. The dataset is relatively small, with a limited number of authors, which increases the risk of overfitting. Moreover, our data was not balanced in other respects; many of the passages written by men were older and sourced from Project Gutenberg, while most of those written by women were more modern and sourced independently. It is possible that this chronological imbalance caused some margin of error in our methods.

7 Conclusion

This study used NLP techniques spanning grammatical structure, probabilistic syntax, emotional sentiment, and large language models to investigate gendered linguistic patterns. While grammatical and syntactic features alone proved insufficient for gender prediction, emotional expression showed a significant difference between men and women. While women showed greater emotional diversity in their writing, men demonstrated a narrower range of emotions in their texts. This suggests that differences are due to how experiences, values, and relationships with the natural world are emotionally framed, rather than how sentences are constructed. While we have verified that there are emotional, grammatical, and linguistic differences between environmental texts authored by women compared to those authored by men, it seems that it is still difficult for a model to accurately predict an author's gender considering these trends alone.

Acknowledgements

The authors wish to acknowledge the contributions of each team member to this research. Emma Virnelli contributed to finding Project Gutenberg texts, initial logistic regression, and sections on PCFG. Caleb Deitch was responsible for finding texts outside of Project Gutenberg, cleaning the data, and sections on POS tagging. Harrison Chang contributed to finding Project Gutenberg texts and the emotional flow evaluation sections. All authors worked on the literature review, discussion, and results, provided critical feedback, and approved the final version of the manuscript. We would also like to thank Sam Lowe for providing the modified RoBERTa model used in emotional flow evaluation and ChatGPT for formatting the references.

References

Mary Austin. 1903. *The Land of Little Rain*. Houghton Mifflin, Boston.

Faiza Belbachir, Thomas Roustan, and Assia Soukane. 2023. Detecting online sexism: Integrating sentiment analysis with contextual language models. *Computing Research Repository*. Preprint.

Rachel Carson. 1962. *Silent Spring*. Houghton Mifflin, Boston.

Sandani Fernando. 2025. How probabilistic context-free grammars (pcfgs) and dependency grammar teach machines to read smarter. Medium.

Gabriela Fontanarrosa, Lucía Zarbá, Valeria Aschero, Daniel Andrés Dos Santos, María Gabriela Nuñez Montellano, Maia C. Plaza Behr, Natalia Schroeder, Silvia Beatriz Lomáscolo, María Elisa Fanjul, A. Carolina Monmany Garzia, Marisa Alvarez, Agustina Novillo, María José Lorenzo Pisarello, Romina Elisa D'Almeida, Mariana Valoy, Andrés Felipe Ramírez-Mejía, Daniela Rodríguez, Celina Reynaga, María Leonor Sandoval Salinas, Verónica Chillo, and María Piquer-Rodríguez. 2024. Over twenty years of publications in ecology: Overcontribution of women reveals a new dimension of gender bias. *PLOS ONE*.

Jean Mark Gawron. Probabilistic cfgs. San Diego State University, Computational Linguistics. Lecture notes; no date available.

Anne Therese Gennari. 2023. *The Climate Optimist Handbook*. Routledge, London.

Amy Godine. 2014. *The Black Woods*. Green Writers Press, Brattleboro, VT.

Jane Goodall. 1988. *My Life with the Chimpanzees*. Pocket Books, New York.

Dave Goulson. 2021. *Silent Earth: Averting the Insect Apocalypse*. Vintage, London.

Kate Greenaway. 1885. *Marigold Garden*. George Routledge and Sons, London.

Charles Frederick Holder. 1913. *The Sea and Its Living Wonders*. G. P. Putnam's Sons, New York.

William T. Hornaday. 1908. *The American Natural History*. Charles Scribner's Sons, New York.

Robin Wall Kimmerer. 2013. *Braiding Sweetgrass*. Milkweed Editions, Minneapolis, MN.

Val Plumwood. 2012. *Eye of the Crocodile*. Australian National University Press, Canberra.

Frederick Courteney Selous. 1900. *African Nature Notes and Reminiscences*. Macmillan, London.

Henry David Thoreau. 1854. *Walden; or, Life in the Woods*. Ticknor and Fields, Boston.

Greta Thunberg. 2019. *No One Is Too Small to Make a Difference*. Penguin, London.

Alfred Russel Wallace. 1911. *The World of Life*. Chapman and Hall, London.

Site Xu and Mu Sun. 2025. Natural language processing (nlp): Identifying linguistic gender bias in electronic medical records (emrs). *PubMed Central*. Accessed via PMC; no publication date available.