

# Detecting and Mitigating Sexist Language in Greek Using Large Language Models

Eleni Karadimou

National and Kapodistrian University of Athens  
Athens, Greece  
eleni97k@gmail.com

Katerina Pastra

Institute for Language & Speech Processing, Athena  
Research Center  
Athens, Greece  
kpastra@athenarc.gr

## Abstract

Human communication relies heavily on language, which is shaped by social conditions and cultural norms. As a result, both written and spoken discourse may embed gender biases and non-inclusive expressions, and even propagate stereotypes even when the speaker has no intention of doing so. Our work focuses on the development of a recommendation system for detection and mitigation of sexist language, with a special focus on Greek. To this end, we have developed a training dataset with examples of sexist and non-sexist language use, along with a lexicon and selected linguistic mitigation strategies; this material is then fed into a large language model for identifying biased expressions and generating more inclusive alternatives. The proposed framework emphasizes the critical role of data design in guiding model behavior, supporting both the detection and the reformulation of sexist language.

## Keywords

Sexist Language Detection, Non-Inclusive Language, Greek, Large Language Model, NLP, Gender Bias

## 1 Introduction

Language reflects social structures and norms. As such, discourse often contributes to the expression and reproduction of stereotypes, including gender-based and other types of discrimination. Variability in verbal expression challenges detection of such phenomena. In inflectional languages such as Greek, where morphology is particularly rich, the manifestation of verbal bias presents additional challenges. Words can appear in many different forms depending on grammatical gender, number, and case, increasing lexical variability and making it harder for computational systems to consistently identify biased expressions. The prominence of grammatical gender in Greek further facilitates the intentional or unintentional reproduction of stereotypes, giving rise to the phenomenon of sexist language.

In recent years, non-inclusive language in general has attracted increasing attention, particularly in relation to computational systems. Advances in Artificial Intelligence and Natural Language Processing have enabled systems to process and generate language, raising concerns about how they handle non-inclusive language and linguistic bias. Although this issue has been widely recognized, computational approaches for detecting and mitigating non-inclusive language remain limited and focus mainly on high-resource languages such as English, leaving languages like Greek relatively underexplored. On the other hand, resources for non-inclusive language, such as glossaries and guidelines are being developed at

national levels or at the level of international organisations, however, these endeavours are limited in scope, and fail to scale up. This work addresses sexist language in Greek and investigates how it can be detected and mitigated using Artificial Intelligence. The proposed approach emphasizes the importance of selecting and constructing appropriate datasets for the effective use of state-of-the-art large language models.

## 2 State of the Art

Interest in the detection and mitigation of non-inclusive language, including sexist language, has increased significantly in recent years. However, most available tools primarily target corporate environments, particularly the job market. For instance, Allybot[1] is an inclusive language tool for Slack that detects non-inclusive expressions and suggests alternative formulations, aiming both to prevent such patterns in workplace communication and to raise awareness among employees regarding inclusive language practices. At the same time, considerable research and practical efforts have focused on reducing gender bias in language within the field of human resources (HR). Gender Decoder[2], based on a study[7] to demonstrate the presence of gendered language in job advertisements, has been developed to identify linguistic patterns that may reinforce gender inequality. Similarly, Textio[5] supports the use of inclusive language throughout the recruitment process, from drafting job postings to communicating with candidates. Beyond application-oriented tools, recent work has also explored computational approaches to bias detection. Tang et al. (2017)[9] developed two algorithms, inspired by existing approaches, to quantify gender bias in the job market. More recent work by Raza et al. (2023)[8] proposes a multi-layered framework for bias identification in textual data, integrating techniques from Named Entity Recognition to detect biased words and phrases. Despite these advances, existing approaches remain limited in scope, often focusing on specific domains or high-resource languages, while available resources do not easily support scalable solutions. This work addresses these limitations by focusing on Greek and by combining the development of targeted datasets with a large language model for the detection and reformulation of sexist language.

## 3 Dataset, Lexicon and Linguistic Mitigation Strategies

Modeling non-inclusive language detection and systematic mitigation of appropriate alternatives may rely on a number of resources, each one contributing in different ways and to a different extent, depending on the modeling method. In what follows, we present the resources we developed specifically for Greek, comprising (a)

pairs of sentences capturing representative examples and their mitigation, (b) a seed repository of words/phrases and their suggested alternatives, and (c) systematic observations of the grammatical and syntactico-semantic patterns in which non-inclusive expressions are embedded, along with conversion patterns for facilitating the adoption of inclusive language. These resources (i.e., dataset, lexicon and verbal strategies) are unique for Greek; the lexicon and the verbal strategies go also beyond the type of resources currently employed in state of the art research on inclusive language generation [6].

### 3.1 Greek Gender Inclusive Language Sentence Pairs Dataset

We have developed a dataset of 300 sentence pairs, each comprising one sentence exhibiting language sexism and its corresponding non-sexist alternative. Two sources were defined to select sentences that reflect sexist language. The first source was a specific set of administrative documents of the Athena Research Center, while the second source was the Hellenic National Corpus of Greek Language[3] (HNC). While the administrative documents were scanned for potential use of non inclusive language in the framework of a corresponding reform endeavour of the Center, the corpus was queried with known non inclusive lemmas in order to retrieve relevant instances of potentially biased language. Non-sexist alternatives were then formulated. The goal was to achieve the optimal yet minimal modification of the original sentence, allowing the changes to be easily described in a subsequent stage. Several revisions were required before the gender-inclusive sentences were finalized.

### 3.2 Gender-Inclusive Language Lexicon for Greek

Additionally, more than 1,000 potentially sexist words and phrases were collected for the construction of a lexicon. The material was compiled through a combination of complementary methods. First, existing glossaries and guidelines for non-inclusive language—both in Greek and other languages—were examined to identify commonly used terms and their inclusive alternatives, providing an initial reference framework. Secondly, a Python script was developed to extract nouns and adjectives from a set of 27 official administrative documents of the Athena R. C., comprising a total of 155,665 tokens, enriching the dataset with terms from real-world official administrative language use. Finally, additional material was collected using the Simple Browser tool[4], through which lexical entries were retrieved along with their semantic categories. These categories were then used to annotate the remaining entries, enabling the distinction between different meanings of the same lexical form. This was particularly important for disambiguating terms with multiple interpretations (e.g., animate vs. inanimate), ensuring that the lexicon focuses on human referents. Non-sexist alternatives were determined in parallel with their collection. From the outset, a key objective was to develop a gender-inclusive lexicon capable of representing *all members of society*. To this end, the Greek non-sexist language lexicon records available masculine and feminine forms and, where applicable, proposes a **gender-neutral alternative** for going beyond binary distinctions, as well as for cases where gender is unknown or irrelevant. We consider such

approach important for broadening one’s linguistic options and for a fully gender inclusive language use, beyond binary dichotomies (as in the case of dual gender marking and exclusive feminine form use or plural form use). The proposed gender neutral alternatives rely on natural language mechanisms and constructs rather than the use of "made up" strategies to express neutrality, as in the case of using the symbol as inflectional ending in person denoting nouns or person qualifying adjectives. Beyond the "artificial" nature of the latter, its use in formal settings is prohibitive.

### 3.3 Gender-neutral Linguistic Mitigation Strategies

In parallel with the creation of the dataset and lexicon described above, the groundwork was laid for the development of a set of strategies for addressing sexist language. These strategies allow one to go beyond direct replacement of non-inclusive term(s) with inclusive one(s) as found in the lexicon, and focus on ways natural language (Greek in our case) allows for gender-neutral meaning expression. These include:

- (1) Omission of non-inclusive term(s) when inferable, including non-essential relative clauses, and prepositional and nominal modifiers
- (2) Use of metonymy
- (3) Use of impersonal constructions
- (4) Transformation of nominal into verbal construction and vice-versa
- (5) Use of second person plural
- (6) Use of active voice present participle(s)
- (7) Transformation of non-agreeing nominal modifiers into agreeing adjectival modifiers.

## 4 Computational Experiments

Following dataset construction, the second stage of this work focuses on the implementation of a system for detecting and reformulating sexist language in Greek. In this context, a series of experiments were conducted using a large language model (ChatGPT 5 Thinking) guided through prompt engineering:

(a) In the first experiment, the model was provided with minimal input. A subset of 50 gender-biased sentences from the dataset (test set) was used, and the model was prompted to identify instances of sexist language and propose alternative formulations. No examples or lexical resources were included, aiming to assess whether minimal guidance is sufficient for achieving satisfactory results.

(b) In the second experiment, the same test set was used, but the prompt was enriched with examples of transforming gender-biased sentences into gender-neutral ones, along with the strategies applied in each case.

(c) In the third experiment, the prompt was further enriched by including both transformation examples and the Greek non-sexist language lexicon, providing more comprehensive guidance to the model.

The evaluation of the results did not meet initial expectations. It was observed that the complexity of the task, combined with the volume of the provided information, led to model confusion and inconsistency in the generated outputs. This highlighted the need for a more structured methodology, consisting of clearly defined

and sequential steps, to facilitate both supervision and evaluation of the system’s performance.

To address these challenges, the process was restructured into two distinct stages: (a) the detection of sexist language, and (b) the generation of alternative gender-neutral formulations.

#### 4.1 Detection of Gender-Biased Terms

The first stage focuses exclusively on the identification of gender-biased terms. A series of experiments was conducted to evaluate performance under different prompting conditions.

In the initial experiment, no additional guidance was provided to the model. Specifically, the model was given a set of 50 sentences from the test dataset and instructed to identify gender-biased words or phrases within each sentence. The primary objective of this experiment was to establish a baseline, enabling comparison with subsequent experiments that incorporate our resources. Notably, the model demonstrated strong performance even under these minimal conditions, achieving robust F1 of 88.5% and thereby establishing a strong baseline for future improvements.

The second phase of detection experiments consisted of a sequence of prompt variations incorporating the Non-Sexist Greek Lexicon. The key difference among these experiments lies in the progressive enrichment of a base prompt with targeted instructions aimed at optimizing performance. Furthermore, in line with the revised two-stage architecture of the system, a simplified version of the Non-Sexist Greek Lexicon was introduced, tailored specifically to the requirements of the detection task. In this form, the lexicon functions as a curated list of terms that are, or may be considered, gender-biased. Gender-neutral alternatives were intentionally excluded at this stage, as they are not required for detection and could potentially introduce noise or negatively affect performance. Near-perfect detection of gender-biased terms was achieved with the 50 sentences test set, with an F1 score reaching 97.6%, signaling the importance of the lexicon and the need for large-scale evaluation employing the whole dataset. The outputs of this stage serve as the input for the subsequent generation phase.

#### 4.2 Generation of Gender-Neutral Sentences

The second stage of the system, which focuses on the generation of alternative gender-neutral formulations, is currently ongoing. In this phase, the model receives as input the output of the first stage, the full version of the Non-Sexist Greek Lexicon (including both gender-biased terms and their corresponding gender-neutral alternatives), as well as a set of transformation strategies for converting gender-biased sentences into gender-neutral ones.

Similarly to the detection phase, this stage consists of a series of experiments in which the prompt is iteratively refined to guide the model toward producing optimal gender-neutral reformulations. Preliminary results indicate that the model achieves a high level of performance, generating satisfactory outputs in most cases. However, there remains room for further improvement, and ongoing experimentation aims to enhance both the consistency and the linguistic quality of the generated sentences.

## 5 Future Work

The present work does not conclude with the completion of the system’s second stage. A primary direction for future research involves the further expansion and continuous updating of the Non-Sexist Greek Lexicon, with the aim of improving its coverage, accuracy, and adaptability to evolving linguistic usage.

Beyond lexicon enhancement, a key objective is the extension of the system’s scope. Rather than focusing exclusively on the detection and reformulation of sexist language, future iterations will aim to support inclusive language more broadly within the Greek linguistic context. This includes addressing additional dimensions of bias and exclusion, such as those related to gender identity, ethnicity, disability, and other sociolinguistic factors.

Ultimately, the goal is to develop a comprehensive framework for inclusive language processing in Greek, capable of both identifying potentially exclusionary expressions and generating contextually appropriate, inclusive alternatives. Such a system could contribute to a wide range of applications, including educational tools, content moderation systems, and assistive writing technologies.

## 6 Conclusion

We presented an approach for the detection and reformulation of gender-biased language in Greek, with particular emphasis on the development of dedicated linguistic resources and their integration with large language models. To address task complexity, the proposed system was structured as a two-stage framework, separating detection from generation. The results demonstrate strong performance in identifying gender-biased terms, while the generation stage is ongoing, with promising preliminary outcomes and potential for further improvement.

## References

- [1] 2026. Allybot. <https://allybot.io/>. Accessed: 2026.
- [2] 2026. Gender Decoder. <https://gender-decoder.katmatfield.com/>. Accessed: 2026.
- [3] 2026. Hellenic National Corpus. <https://hnc.ilsp.gr/>. Accessed: 2026.
- [4] 2026. PAROLE-SIMPLE-CLIPS Lexicon. <https://www.cnr.it/en/institutes-databases/database/445/parole-simple-clips-lexicon>. Accessed: 2026.
- [5] 2026. Textio. <https://textio.com/>. Accessed: 2026.
- [6] Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Meghann L. Drury-Grogan, Miguel Ángel García Cumbreiras, Salud María Jiménez Zafra, Thomas Mandl, Sylvia Jaki, Rahul Ponnusamy, Anand Kumar M, Dhanalakshmi V, Bharathi B, Premjith B, Senthil Kumar B, and Sathiyaraj T. 2026. Insights from Multilingual Gender Inclusive Language Generation Shared Task. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.
- [7] Danielle Gaucher, Justin Friesen, and Aaron Kay. 2011. Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology* (2011).
- [8] Shaina Raza et al. 2024. Nbias: A Natural Language Processing Framework for Bias Identification in Text. *Expert Systems with Applications* 237 (2024), 121542. doi:10.1016/j.eswa.2023.121542
- [9] Shiliang Tang et al. 2017. Gender Bias in the Job Market: A Longitudinal Analysis. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19. doi:10.1145/3134734