GeNRe: a French Gender-Neutral Rewriting System Using Collective Nouns

Anonymous ACL submission

Abstract

A significant portion of the textual data used in the field of Natural Language Processing (NLP) exhibits gender biases, particularly due to the use of masculine generics (masculine words that are supposed to refer to mixed groups of men and women), which can perpetuate and amplify stereotypes. Gender rewriting, a NLP task that involves automatically detecting and replacing gendered forms with neutral or opposite forms (e.g., from masculine to feminine), can be employed to mitigate these biases. Such systems are available for English, Arabic, Portuguese and German, but no French system is available. We create an original French gender-neutral rewriting system using collective nouns, which are gender-fixed in French. This paper presents GeNRe, the very first French gender-neutral rewriting system. We introduce a rule-based system (RBS) tailored for the French language alongside two fine-tuned large language models trained on data generated by our RBS. We also explore the use of instruction models to enhance the performance of our other systems and find that Claude 3 Opus combined with our dictionary achieves results close to our RBS. Through this contribution, we hope to promote the advancement of gender bias mitigation techniques in NLP for French.

1 Introduction

Since the 1970s, a number of psycholinguistics studies have focused on how language influences thoughts (Berlin and Kay, 1969; Kay and McDaniel, 1978). Further studies examining gender in language showed that it could lead to cognitive biases (Jacobson and Insko, 1985; Sczesny et al., 2016), particularly when it comes to the use of masculine generics (masculine words that are supposed to refer to mixed groups of men and women)[1] (Braun et al., 2005; Richy and Burnett, 2021). For example, Stahlberg et al. (2001) showed that when asked to name a celebrity in a certain field in German, respondents were more likely to give the name of a man when a masculine generic was used in the question.

Gender bias in natural language processing (NLP) models is a critical issue that can lead to biased predictions and the amplification of biases present in training data. This problem is particularly relevant for machine translation systems, which are highly susceptible to gender biases when translating between languages with different grammatical gender systems (Savoldi et al., 2021). Data augmentation, which involves balancing the amount of data for all genders in a specific language, has been proposed as a potential solution to debias NLP systems (Zhao et al., 2018). To achieve this goal, current research projects automatically propose alternatives to sentences containing masculine generics, contributing to an NLP task known as “gender rewriting”.

As of yet, gender neutralization techniques have not been developed for French, even though it is a heavily gendered language. Thus, we aim to create a French gender-neutral rewriting system using human collective nouns, defined by Lecolle (2019) as “nouns referring to entities comprised of groups of individuals”[2]. Collective nouns have been widely discussed in the literature, especially when it comes to French (Flaux, 1999; Lammert, 2010; Lammert and Lecolle, 2014; Lecolle, 2019). Since this type of noun has a fixed gender in French[3], it is an effective way of achieving gender neutralization. This gender-neutral rewriting system, GeNRe (**Ge**nder-**N**eutral **Re**writing System Using French Collective

[1] There is no strict equivalent in English, but an example could be the use of “policemen” to refer to both men and women pursuing that occupation, instead of using “police officers.”

[2] In French: « nom désignant une entité composée d’un ensemble d’individus humains. »

[3] For instance, “la police” (“police”) refers to both policemen and policewomen.

Nouns), is the very first gender-neutral rewriting system for French[4] and could foster the development of other types of gender rewriting systems for that language in the future.

2 The Task of Gender Rewriting

While Alhafni et al. (2022b) were the first to define this task as "gender rewriting," similar efforts had already been pursued for Arabic (Habash et al., 2019), German (Pomerenke, 2022), and English (Sun et al., 2021). Alhafni et al. (2022b) suggest the following definition for this task: "generating alternatives of a given Arabic sentence to match different target user gender contexts." (2). While this definition works well for Alhafni et al.'s work, as they focus specifically on Arabic and create a system to switch between the masculine gender and the feminine gender, it is not universally applicable. Indeed, among the aforementioned works, several approaches to gender rewriting have been explored: Habash et al. (2019) and Alhafni et al. (2022a) developed a system to transform Arabic sentences with masculine words into sentences with feminine equivalents, and vice versa. Pomerenke's (2022) system seeks to provide inclusive suggestions for input sentences in German; so does Veloso et al.'s (2023) system for Portuguese. Finally, Sun et al. (2021), Vanmassenhove et al. (2021) and He et al. (2021) created systems to neutralize gender in an English input sentence.

We suggest a new, universal definition for the task that works for all languages and all the transformation approaches when it comes to gender based on the latest works mentioned previously:

> The use of a gendered input sentence to generate one or several alternative sentences with different gender forms by neutralizing them, choosing inclusive forms or switching to another gender.

3 Gender in French

In French, nouns are classified as either masculine or feminine, and the gender of a noun influences the form of adjectives, pronouns, and verbs that accompany it. The gender of human nouns reflects the sociological gender of the referent (for instance, "danseuse" refers to a female dancer), while gender of nouns referring to unanimated beings is arbitrary (Watbled, 2012).

The masculine gender is considered to be the "default" gender in French, and can be used in a non-specific context or to refer to groups of people composed of both men and women. The use of masculine as the default gender can however lead to both gender biases and invisibilizing women. As a result, two main writing techniques have been developed to avoid its use: visibilization techniques and neutralization techniques.

Visibilization techniques seek to highlight the feminine ending of words by separating the masculine ending from the feminine one through the use of specific symbols (asterisk, interpunct: *acteur.ice*) or by affecting the feminine ending directly (using capital or bold letters). Neutralization techniques, on the other end, aim to use epicene words, that is words whose form is the same for masculine and feminine (e.g. "spécialiste", *specialist*), or words that refer to groups of people, such as collective nouns (e.g. "lectorat", *readership*), these having a fixed gender which is not associated with the genders of the people within that group.

We chose to focus on gender neutralization due to it being a less explored issue in research comparatively to visibilization techniques. Moreover, while collective nouns are a great asset for gender neutralization, their usage is still restricted to a few words and their full potential has not yet been explored.

4 Methodology

We propose three different approaches for the task of gender-rewriting: a rule-based approach, a neural model fine-tuning approach and an instruction model approach. To build the resources used for these systems, we first create a dictionary of French collective nouns and their member noun counterparts, which we describe in Section 4.1. In Section 4.2, we then give details about the datasets that we extracted sentences from for the development of our rule-based system, large language model (LLM) fine-tuning and evaluation. Finally, in Section 4.3, we delve into the specifics of our experimental design with the aforementioned model types.

4.1 Dictionary

First, we manually created a dictionary with French collective nouns and their member noun counter-

[4]Code and data are publicly available on GitHub: https://github.com/REDACTED

parts. Two approaches were used to fill this dictionary: literature review, consisting of retrieving collective nouns mentioned in the French literature, and manual collecting, consisting of collecting occurrences on the Internet and in newspaper articles, as well as scraping French Wiktionary pages containing lists of such nouns. We respectively retrieved 210 and 105 nouns using these methods (315 in total). Table 1 contains a few examples of entries in our dictionary.

Collective noun	Member noun (masc. plural)
académie (*academy*)	académiciens (*academicians*)
armée (*army*)	soldats (*soldiers*)
milice (*militia*)	miliciens (*militiamen/women*)
artillerie (*artillery*)	artilleurs (*artillerists*)
auditoire (*listenership*)	auditeurs (*listeners*)
ballet (*ballet*)	danseurs (*dancers*)
police (*police*)	policiers (*police officers*)

Table 1: Collective noun-member noun dictionary overview

4.2 Datasets

Using our dictionary, we searched for occurrences of masculine plural member nouns in a French Wikipedia dataset with 1.58 million texts (graelo, 2023)[5]. We extracted 292,076 sentences containing such nouns. In addition, we also extracted French sentences from the Europarl EN-FR corpus (Koehn, 2005), a corpus created from the proceedings of the European Parliament and available in 21 languages, including English and French. This corpus was filtered to include French sentences only, and 106,878 additional sentences were extracted for neural model fine-tuning and evaluation (total 398,954). Both of these corpora are made available for research purposes.

For the rule-based system specifically, tags were automatically added at the beginning and at the end of each member phrase in the extracted sentences, with the ID of the entry in the dictionary. This was done because member nouns may have several collective noun counterparts, leading to several different sentences being generated in addition to the main one. For instance, the member noun "soldats" (*soldiers*) could well be replaced with collective nouns "armée" (*army*) "bataillon" (*battalion*), "infanterie" (*infantry*) or "régiment" (*regiment*). As we used data generated by our rule-based system for neural model fine-tuning (see Section 4.3.2), this was especially useful to generate all the possible variations of the input sentence, and thus increase the number of examples the models were trained on. Moreover, the use of tags also helps guarantee the member nouns to be replaced in the input sentence, as only those that are between tags will be taken into account.Example 1 shows how these tags are used.

(1) a. Un historique permet de lister <n-126>les auteurs</n> et de consulter les modifications successives de l'article par <n-68>ses rédacteurs</n>.
(A history allows one to list <n-126>the authors</n> and view successive modifications to the article by <n-68>its editors</n>.)

Finally, we created a corpus-specific evaluation dataset comprised of 250 sentences from each corpus (total 500), and we manually gender-neutralized each sentence to have gold sentences.

4.3 Models

In this section, we present three different model types for gender-neutral rewriting: a rule-based model, two neural models, and an instruction model. Each model takes a different approach to the task, allowing us to compare their performance.

4.3.1 Rule-based model

We developed a rule-based system (RBS) to automatically apply the correct syntactic rules when converting a member noun into a collective noun, which leads to number and gender changes in the sentence.

The RBS consists of two main components: a syntactic dependency detection component and a generation component.

The dependency detection component primarily relies on spaCy (Honnibal et al., 2020) with the fr_core_news_sm pipeline as well as a set of rules to detect the words that are syntactically related to the member noun that needs to be replaced.

The generation component replaces each member noun in the sentence with its collective noun

[5]Dataset is available here: https://huggingface.co/datasets/graelo/wikipedia. License: CC-BY-SA-3.0

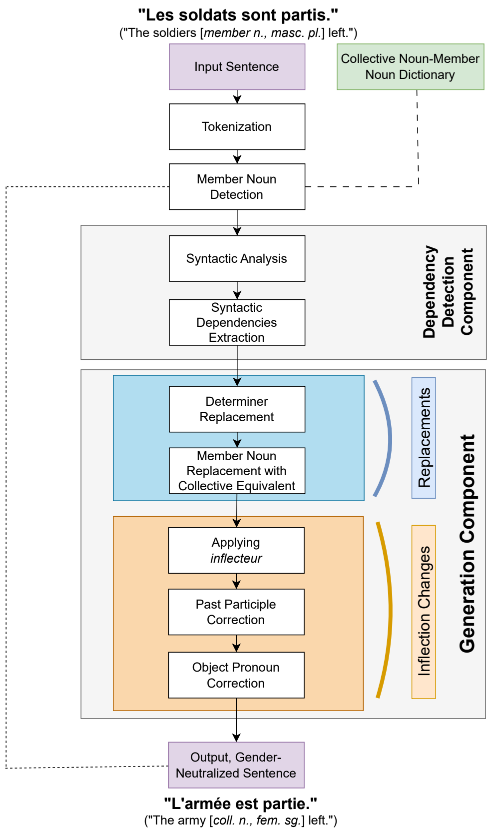

Figure 1: Rule-based model replacement pipeline overview

counterpart found in the dictionary, adjusting the determiner, handling elision, and reinflecting the detected dependencies using *inflecteur* (Chuttarsing, 2021), a Python module leveraging the Delaf French morphological dictionary[6] and *french-camembert-postag-model*[7], a CamemBERT-based (Martin et al., 2020) part of speech (POS) tagging model for French. Our RBS also makes additional replacements for past participles and object pronouns as these are not always being well handled by the *inflecteur* Python module. If no member nouns are detected in the sentence, the original sentence will be returned instead as it is already considered gender-neutral. Figure 1 shows an overview of the rule-based model pipeline.

4.3.2 Neural models

Recent research on gender rewriting has focused on training neural models as well as fine-tuning large language models using data generated by RBS to improve task-specific performance. While some studies (Sun et al., 2021; Veloso et al., 2023) showed a decrease in performance compared to RBS, Vanmassenhove et al. (2021) found a notable improvement of 0.27 in WER. We aim to investigate whether fine-tuning large language models can significantly improve the results of RBS, hypothesizing that the linguistic knowledge acquired by these models during training on large text corpora will help resolve errors in the training corpus and enhance results.

Two Seq2seq large language models (LLMs), t5-small (Raffel et al., 2020) and m2m100_418M (Fan et al., 2020), were selected for the experiments, and were fine-tuned using our two RBS-generated corpora (Wikipedia and Europarl) containing gender-neutralized and non-gender-neutralized sentence pairs. The training dataset for each model consisted of 60,000 sentence pairs per corpus, and the validation dataset had 6,000 (10%). Hyperparameters used for training are available in Appendix A.

4.3.3 Instruction model

The rapid development of LLMs and advances in NLP have demonstrated the ability to manipulate language models' behavior to predict text continuations and perform specific tasks without explicit training, leading to "instruction models" such as InstructGPT (Ouyang et al., 2022) or, more recently, Mixtral 8x7B Instruct (Jiang et al., 2024). This is primarily achieved through the use of "prompts" or instructions given to the language model (Liu et al., 2021). While some studies have briefly mentioned the potential of instruction models to reduce gender biases in automatically generated texts, and have occasionally experimented with such models[8], no gender rewriting study has yet conducted a comprehensive analysis of their capabilities for this specific task. As a result, we aimed to leverage this kind of model in order to evaluate its performance for this task. We chose Claude 3 Opus, which is, at the time of writing, considered to be the best model for textual generation according to specific benchmark (Anthropic, 2024).

To comprehensively evaluate the performance of Claude 3 Opus, we designed three distinct types of instructions to test its ability to generate gender-neutral texts. Corresponding prompts are available

[6] https://uclouvain.be/fr/instituts-recherche/ilc/cental/delaf-2-0.html

[7] https://huggingface.co/gilf/french-camembert-postag-model

[8] For instance, Veloso et al. (2023) tried to make use of ChatGPT to generate gender-inclusive sentences in Portuguese, and suggested that the use of instruction models could prove useful to automatically create gender-inclusive datasets.

in Appendix A.

- The "BASE" instruction provides a basic task description, asking the model to make the sentence inclusive by replacing masculine generics with their collective noun equivalents, without explicitly specifying the replacement word.

- The "DICT" instruction leverages our collective noun dictionary and asks the model to replace masculine generics with their corresponding collective nouns, those being explictly mentioned. There are two different versions for the "DICT" instruction: "DICT-SG", used when only one generic masculine noun with a matching collective noun was found in the sentence, and "DICT-PL", used when several generic masculine nouns with matching collective nouns were found.

- The "CORR" instruction takes sentences generated by our RBS as input and tasks the model with correcting potential errors, such as mismatches between verb and adjective numbers and genders.

5 Results

To evaluate the performance of our different rewriting models, we leverage two evaluation metrics commonly used for the task of gender rewriting: Word Error Rate (WER) and BLEU (Papineni et al., 2002). JiWER 3.0.3[9] and bleu 0.3[10] Python packages were used with default parameters.

Average results of each model on the two corpora are available in Table 2.

Type	WER (↓)	BLEU (↑)
Baseline (unchanged)	13.35%	80.55
GeNRe-RBS	**3.40%**	93.43
GeNRe-T5	5.11%	90.68
GeNRe-M2M-100	5.40%	90.17
Claude 3 Opus-BASE	12.16%	82.98
Claude 3 Opus-DICT	3.75%	**93.64**
Claude 3 Opus-CORR	10.17%	85.13

Table 2: Results by model type. Bold indicates the best results overall.

The RBS and Claude 3 Opus-DICT achieved the best results in our experiments. While the RBS model achieved the best WER score, Claude 3 Opus-DICT achieved the highest BLEU score. These results can be explained by the fact that WER and BLEU scores capture distinct aspects of text generation. Due to its reliance on predefined rules, the RBS easily preserves original words and word order, likely leading to a lower WER. On the contrary, instruction models are known to be more prompt to slightly deviate from the original formulation of sentences, which may increase the WER without significantly affecting the BLEU score due to the order of words not being taken into account.

The neural models also showed mostly promising results. Comparing the two of them, they achieved similar results, with the T5 model slightly outperforming M2M-100. However, both models showed a minor decrease in performance compared to the RBS. As a result, similarly to Veloso et al. (2023), we do not find a significant improvement compared to our RBS following fine-tuning.

Moreover, we also provide the distribution of errors made by GeNRe-RBS, GeNRe-T5 and GeNRe-M2M-100 in Figure 2. Error types can be divided into three main categories: POS (ADJ, DET, DET_COREF, PRON_COREF, VERB), text generation (CASE, GEN_FAILURE, SPECIAL_CHAR) and other (ELISION, MISID_NOUN, PUNCT, SEM, UNREPLACED).

Text generation errors, labeled with (N) in Figure 2, are strictly specific to neural models. CASE refers to capitalization errors (missing/extra uppercase or lowercase); GEN_FAILURE refers to token-specific generation errors (for instance, incorrectly replacing a proper name with a non-existent name); SPECIAL_CHAR refers to errors related to special characters (for instance, accents).

When it comes to other errors, ELISION is used when there was an issue with how one or multiple words in the generated sentence were elided[11]. MISID_NOUN occurs when a word in the automatically annotated corpus was mistaken as a noun. PUNCT refers to errors related to punctuation or typography (double spaces, for example). SEM is used to label automatically generated sentences which cannot be considered semantically correct due to the replacement of the member noun with its collective noun counterpart[12] Finally, UNRE-

[9] https://pypi.org/project/jiwer/

[10] https://pypi.org/project/bleu/

[11] For instance, in French, the masculine determiner "le" and the feminine determiner "la" (*the*) should be elided and written as "l'" when the word that follows begin with a vowel or a mute "h".

[12] As discussed by Lecolle (2019), collective nouns in

PLACED occurs when a member noun that is in our dictionary was not replaced.

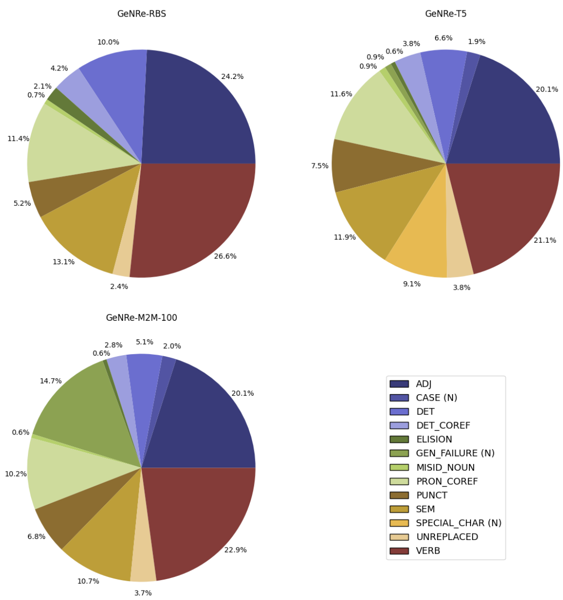

Figure 2: Distribution of errors across GeNRe-RBS, GeNRe-T5 and GeNRe-M2M-100

Across all three models, the most prominent error types are related to verbs and noun cases. Verbs account for 26,6% of errors for GeNRe-RBS, 21,1% for GeNRe-T5, and 22,9% for GeNRe-M2M-100. On the other hand, adjectives account for 24,2% of errors for GeNRe-RBS, and 20,1% for both GeNRe-T5 and GeNRe-M2M-100.

The M2M-100 model is highly prone to making token-specific generation errors (14,7%), this type of error being strictly specific to this model. Similarly, we find that the T5 model also makes specific errors related to the handling of special characters. We discuss these issues more in detail in Section 6.

6 Discussion

A qualitative analysis of the generated sentences revealed that the RBS was making most of its errors when modifying adjectives and verbs. This is not surprising given that these two part-of-speech categories are the ones which require the most complex changes when transitioning from a member noun to a collective noun. Indeed, in French, adjectives undergo a certain number of changes when changing number or gender. Verbs can also have these same changes when used as past participles; otherwise, only number change will affect them. For instance, in Example 2, the verb "seront" (pl., *will be*) should have been changed to "sera" (sg.) to match with the new collective noun "citoyenneté" (*citizenry*).

(2) a. Cette démarche fera progresser les droits **des citoyens**, car, par l'intermédiaire du Parlement, **les citoyens seront** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [original sent.]
(This approach will increase **citizens' [masc.]** rights, because, through the Parliament, **citizens will [pl.]** have a direct line to the Commission thereby generating considerable legitimacy.)

b. Cette démarche fera progresser les droits **de la citoyenneté**, car, par l'intermédiaire du Parlement, **la citoyenneté seront** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [GeNRe-RBS]
(This approach will increase the rights of **the citizenry**, because, through the Parliament, **the citizenry will [pl.]** have a direct line to the Commission thereby generating considerable legitimacy.)

c. Cette démarche fera progresser les droits **de la citoyenneté**, car, par l'intermédiaire du Parlement, **la citoyenneté sera** en contact direct avec la Commission, ce qui lui confèrera une légitimité considérable. [manual sent.]
(This approach will increase the rights of **the citizenry**, because, through the Parliament, **the citizenry will [sg.]** have a direct line to the Commission thereby generating considerable legitimacy.)

Similarly, in Example 3, the adjective "chargés" (pl., *in charge of*) should match the new singular collective noun "parlement" (*parliament*) and be changed to "chargé".

(3) a. Je vous invite à informer **les députés**

French, and more specifically human collective nouns, feature specific semantic characteristics due to how they are used to group human beings under a common denomination, based for example on their profession (« le professorat » [*professorate*]), their social status (« l'aristocratie » [*the aristocracy*]), or their political leaning (« la gauche » [*the left*]). Combining human collective nouns with specific verbs or contexts may thus not be considered semantically correct, and may occur when transforming a sentence. We labeled such transformed sentences with this error.

européens chargés des dossiers agricoles de l'avancement des négociations. [original sent.]
(I urge you to inform **the Members of European Parliament [masc] in charge of [pl.]** the agricultural issues about the progress of negotiations.)

b. Je vous invite à informer **le parlement européen chargés** des dossiers agricoles de l'avancement des négociations. [GeNRe-RBS]
(I urge you to inform **the European parliament in charge of [pl.]** the agricultural issues about the progress of negotiations.)

c. Je vous invite à informer **le parlement européen chargé** des dossiers agricoles de l'avancement des négociations. [manual sent.]
(I urge you to inform **the European parliament in charge of [sg.]** the agricultural issues about the progress of negotiations.)

When it comes to the neural models (T5 and M2M-100), analysis shows that they were able to generalize linguistic rules and correct dependencies that were not properly modified by the RBS, especially verbs and adjectives, slightly reducing the number of errors for these POS. As a result, in spite of their lower results compared to the RBS, fine-tuned models may still prove useful in certain scenarios where the RBS struggles to apply linguistic rules correctly, such as in sentences with complex dependencies or nuanced contextual relationships. Example 4 shows a case where the verb "vouloir" (*want*) is correctly inflected by the neural model.

(4) a. Un deuxième élément concerne le soutien apporté à la Commission à **l'actorat local** qui **veulent** participer à ces programmes afin d'avoir accès aux sources de financement correspondantes. [GeNRe-RBS]
(A second factor is the Commission's support for **local actors [coll. sg.]** who **want [pl.]** to take part in these programmes, so that they can access the corresponding funding mechanisms.)

b. Un deuxième élément concerne le soutien apporté à la Commission à **l'actorat local** qui **veut** participer à ces programmes afin d'avoir accès aux sources de financement correspondantes. [GeNRe-FT-M2M-100]
(A second factor is the Commission's support for **local actors [coll. sg.]** who **want [sg.]** to take part in these programmes, so that they can access the corresponding funding mechanisms.)

Additionally, the fine-tuned models were capable of utilizing different collective noun equivalences from the dictionary (some collective nouns being associated to the same member noun).

Errors observed in the fine-tuned models and different from the RBS included token generation failures (M2M-100, Example 5, where "Nebski" was generated instead of "Zemski"), and incorrect generation of special characters (T5, as in Example 6 where "main-d'uvre" was generated instead of "main-d'œuvre" [*labour*]). The first error might come from the multilingual aspect of the model, as it may generate words or mix tokens from other languages, while the second error is probably due to the model being mostly trained on English data. For both models, we also found cases where words were not uppercased correctly, as in Example 7.

(5) a. Juin, Russie : le **Nebski** sobor prend des décisions importantes. [GeNRe-FT-M2M-100]
(June, Russia: the **Nebski** Sobor makes important decisions.)

(6) a. Il est allé à Cologne, où il est devenu président de l'association de la **main-d'uvre** et a aidé à propager les idées marxistes parmi ses membres. [GeNRe-FT-T5]
(He went to Cologne, where he became president of the **labour** organization and helped spread Marxist ideas among its members.)

(7) a. **l'**armée arriva avec une lance à eau pour disperser les détenus. [GeNRe-FT-T5]
(**the** army arrived with a water hose to disperse the prisoners.)

As far as the instruction model is concerned, Claude 3 Opus-BASE and CORR were found to be highly prone to altering the formulation of sentences, as shown in Example 8. Claude 3 Opus-DICT was found to have a similar effect, but to a much lesser extent, likely due to the increased precision of the prompt.

(8) a. Dans une lettre à la **famille** datée du 13 juin 1861, Zeng Guofan a ordonné à **ses propres navires** de surveiller les navires commerciaux britanniques après avoir remarqué que des marchands étrangers déchargeaient du riz **à la rébellion** à Anqing. [GeNRe-RBS]
(In a letter addressed to the **family** and dated June 13, 1861, Zeng Guofan ordered **his own vessels** to monitor British commercial vessels after noticing that foreign sellers were giving rice to **the rebellion** in Anqing.)

b. Dans une lettre à la **parenté** datée du 13 juin 1861, Zeng Guofan a ordonné à **sa propre flotte** de surveiller les navires commerciaux britanniques après avoir remarqué que des marchands étrangers déchargeaient du riz **aux rebelles** à Anqing. [Claude 3 Opus-BASE]
(In a letter addressed to the **kinfolk** and dated June 13, 1861, Zeng Guofan ordered **his own fleet** to monitor British commercial vessels after noticing that foreign sellers were giving rice to **rebels** in Anqing.)

Notably, the DICT prompt was observed to generate sentences with correct verbs and adjectives, indicating its ability to effectively leverage the collective noun dictionary to produce grammatically accurate sentences. We give such an example in Example 9.

(9) a. Mais l'armée protestante, toujours agressive, **restaient** à la charge des habitants et **constituaient** une lourde charge. [GeNRe-RBS]
(But the Protestant army, still aggressive, **remained [pl.]** in the care of the local people and **constituted [pl.]** a heavy burden.)

b. Mais l'armée protestante, toujours agressive, **restait** à la charge des habitants et **constituait** une lourde charge. [Claude 3 Opus-DICT]
(But the Protestant army, still aggressive, **remained [sg.]** in the care of the local people and **constituted [sg.]** a heavy burden.)

Nonetheless, among the errors made by Claude 3 Opus-DICT, we identified instances of unreplaced nouns, where the model failed to substitute the masculine generics with their corresponding collective noun equivalents, such as in Example 10.

(10) a. Paradoxalement, cette progression en voix s'accompagne d'un recul en nombre d'élus, du fait de la poussée des candidats indépendants (pour la plupart de la **représentation** de la communauté kurde) et du CHP. [GeNRe-RBS]
(Paradoxically, this increase in votes paralleled a decrease in the number of elected representatives due to better results for the independent candidates (most of them **coming from the representation** of the Kurdish community) and CHP.

b. Paradoxalement, cette progression en voix s'accompagne d'un recul en nombre d'élus, du fait de la poussée des candidats indépendants (pour la plupart des **représentants** de la communauté kurde) et du CHP. [Claude 3 Opus-DICT]
(Paradoxically, this increase in votes paralleled a decrease in the number of elected representatives due to better results for the independent candidates (most of them **being representatives** of the Kurdish community) and CHP.

7 Conclusion

Our work represents a step towards addressing gender-biased textual data in French. We make three key contributions to the task of gender rewriting in NLP: 1) a dictionary of French collective nouns and their corresponding member nouns, which serves as a resource for future research in this area; 2) a dataset of gender-neutralized and non-gender-neutralized sentences; and 3) a rule-based system that effectively gender-neutralizes French sentences using collective nouns, laying groundwork for further advancements for this task in that language. Our experiment combining our manually created dictionary with the Claude 3 Opus instruction model also shows promise for the use of such models for the task of gender rewriting. We strongly believe that future research further exploring the capabilities of these models for that task could lead to the development of effective solutions for mitigating gender bias in many languages.

Ethics Statement

We did not filter the datasets that were used for the development of the RBS and for fine-tuning models for harmful, hateful, inappropriate or personal content. Considering the sources used to constitute these datasets (Wikipedia and Europarl), we believe it very unlikely for those to display such type of content. Similarly, when it comes to output sentences generated by the fine-tuned models, since those were trained on replacing specific words or part of speech in sentences, the generation of such content seems unlikely. As discussed in the paper, instruction models are more prone to reformulating input sentences: while we did not find any inappropriate content in the Claude 3 Opus-generated sentences we evaluated, this kind of models may be trained on such data, which might lead to the generation of harmful or hateful content.

Limitations

French collective nouns adhere to specific semantic rules, which means that their usage may not be universally applicable to all sentences, sometimes resulting in constructions that appear asemantic. This limitation is further compounded by the fact that only a small subset of these nouns is actively employed in everyday language by native speakers, which restricts their versatility and adaptability in various linguistic contexts. We however believe that they are good candidates for gender neutralization, and the development of our system may help promote a broader use of such nouns. In addition, combining our system with a contextual or semantic analysis framework could help address these issues by ensuring that the collective noun equivalents are both contextually relevant and semantically appropriate.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022a. User-Centric Gender Rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, Houda Bouamor, Ossama Obeid, Sultan Alrowili, Daliyah Alzeer, Khawlah M. Alshanqiti, Ahmed ElBakry, Muhammad ElNokrashy, Mohamed Gabr, Abderrahmane Issam, Abdelrahim Qaddoumi, K. Vijay-Shanker, and Mahmoud Zyate. 2022b. The Shared Task on Gender Rewriting. *Preprint*, arxiv:2210.12410.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press.

Friederike Braun, Sabine Sczesny, and Dagmar Stahlberg. 2005. Cognitive Effects of Masculine Generics in German: An Overview of Empirical Findings. *Communications*, 30(1):1–21.

Adrien Chuttarsing. 2021. Inflecteur.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. *Preprint*, arxiv:2010.11125.

Nelly Flaux. 1999. À propos des noms collectifs. *Revue de linguistique romane*, (63):471–502.

graelo. 2023. Graelo/wikipedia dataset. https://huggingface.co/datasets/graelo/wikipedia.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. *Preprint*, arxiv:2109.11708.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Marsha B. Jacobson and William R. Insko. 1985. Use of nonsexist pronouns as a function of one's feminist orientation. *Sex Roles*, 13(1-2):1–7.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *Preprint*, arxiv:2401.04088.

Paul Kay and Chad K. McDaniel. 1978. The Linguistic Significance of the Meanings of Basic Color Terms. *Language*, 54(3):610–646.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation.

Marie Lammert. 2010. *Sémantique et cognition : les noms collectifs*. Droz, Genève.

Marie Lammert and Michelle Lecolle. 2014. Les noms collectifs en français, une vue d'ensemble. *Cahiers de lexicologie*, (105):203–222.

Michelle Lecolle. 2019. *Les noms collectifs humains en français. Enjeux sémantiques, lexicaux et discursifs*. Lambert-Lucas, Université de Lorraine.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Preprint*, arxiv:2107.13586.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arxiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania. Association for Computational Linguistics.

David Pomerenke. 2022. INCLUSIFY: A benchmark and a model for gender-inclusive German. *Preprint*, arxiv:2212.02564.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Preprint*, arxiv:1910.10683.

Célia Richy and Heather Burnett. 2021. Démêler les effets des stéréotypes et le genre grammatical dans le biais masculin : une approche expérimentale. *GLAD!*, (10).

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7.

Dagmar Stahlberg, Sabine Sczesny, and Friederike Braun. 2001. Name Your Favorite Musician: Effects of Masculine Generics and of their Alternatives in German. *Journal of Language and Social Psychology*, 20(4):464–469.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *Preprint*, arxiv:2102.06788.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender-Neutral Alternatives. *Preprint*, arxiv:2109.06105.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A Rewriting Approach for Gender Inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.

Jean-Philippe Watbled. 2012. Linguistique du genre. *L'Harmattan*, pages 167–179.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Fine-Tuning Details

Models were trained on a single NVIDIA RTX 4090 GPU. Training time took approximately 3 hours for each model.

A.1.1 GeNRe-T5

```
BATCH_SIZE = 48
NUM_PROCS = 16
EPOCHS = 5
LEARNING_RATE = 0.0005
WEIGHT_DECAY = 0.02
```

A.1.2 GeNRe-M2M-100

```
BATCH_SIZE = 8
NUM_PROCS = 16
EPOCHS = 5
LEARNING_RATE = 0.0005
WEIGHT_DECAY = 0.02
```

A.2 Instruction Model Hyperparameters

```
model="claude-3-opus-20240229",
temperature=0,
messages=[
    {"role": "user",
    "content": f"{message}"},
    {"role": "assistant",
    "content": "Here is the
    output sentence:"}
]
```

A.3 Instruction Details

Table 3 contains the different types of instructions given to Claude 3 Opus as well as their respective content.

"EXAMPLES" refers to the few-shot sentences given to the instruction model. See Tables 4 and 5 for more information.

"ORIGINAL SENTENCE" is replaced with the sentence containing one or several masculine generic nouns that we want to replace with their collective counterparts. It is part of the prompt in a similar way to the example sentences so that the instruction model is guided towards generating the final, gender-neutralized sentence.

Instruction Type	Content
BASE	Make this French sentence inclusive by replacing generic masculine nouns with their French collective noun equivalents. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
DICT-SG	Make this French sentence inclusive by replacing generic masculine noun {NM} with its respective French collective noun equivalent {NCOLL}. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
DICT-PL	Make this French sentence inclusive by replacing generic masculine nouns {NM1, NM2, . . . } with their respective French collective noun equivalents {NCOLL1, NCOLL2, . . . }. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →
CORR	Correct grammar in this French sentence. Generate the final sentence only without any comments nor notes. {EXAMPLES} {ORIGINAL SENTENCE} →

Table 3: Content of instructions per type given to Claude 3 Opus

A.4 Few-shot sentences given to Claude 3 Opus

Tables 4 and 5 contain the few-shot sentences used respectively for the "BASE" and "DICT" instructions, and the "CORR" instruction. They were formatted as such in the prompt:
[Sentence with masculine generic] → [Gender-neutralized sentence].

Sentence with masculine generic	Gender-neutralized sentence
Le président de la FIFA Sepp Blatter rejette les accusations **des manifestants** en les accusant d'opportunisme. (FIFA President Sepp Blatter dismisses **the protesters**' accusatations as opportunism.)	Le président de la FIFA Sepp Blatter rejette les accusations de **la manifestation** en l'accusant d'opportunisme. (FIFA President Sepp Blatter dismisses **the protest**'s accusatations as opportunism.)
Les auteurs et **les spectateurs** ont été satisfaits des réponses des représentants. (**Authors** and **spectators** were pleased with **the representatives**' responses.)	**L'autorat** et **le public** ont été satisfaits des réponses de **la représentation**. (**The authorship** and **the audience** were pleased with the **representation**'s responses.)
Le vicaire général proposa de disperser **les religieux** dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments. (The vicar general suggested to disperse **religious people** to other houses of the order to repair the buildings.)	Le vicaire général proposa de disperser **le couvent** dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments. (The vicar general suggested to disperse **the convent** to other houses of the order to repair the buildings.)

Table 4: Few-shot sentences for "BASE" and "DICT" instructions. Bold indicates the differences between sentences with masculine generics and gender-neutralized sentences.

RBS-generated sentence with errors	Manual sentence
Le président de la FIFA Sepp Blatter rejette les accusations de la manifestation en **les** accusant d'opportunisme.	Le président de la FIFA Sepp Blatter rejette les accusations de la manifestation en **l'**accusant d'opportunisme.
L'autorat et le public **a** été satisfaits des réponses des la représentation.	L'autorat et le public **ont** été satisfaits des réponses de la représentation.
Le vicaire **générale** proposa de disperser le couvent dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments.	Le vicaire **général** proposa de disperser le couvent dans d'autres maisons de l'ordre et de procéder à la réfection des bâtiments.

Table 5: Few-shot sentences for "CORR" instruction. Bold indicates the differences between the RBS-generated ssentences with error and the manual, correct sentences.