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

Enzo Doyen
LiLPa, University of Strasbourg
enzo.doyen@unistra.fr

Amalia Todirascu
LiLPa, University of Strasbourg
todiras@unistra.fr

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, an 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. While such systems have been developed in a number of languages (English, Arabic, Portuguese, German, French), automatic use of gender neutralization techniques (as opposed to inclusive or gender-switching techniques) has only been studied for English. This paper presents GeNRe, the very first French gender-neutral rewriting system using collective nouns, which are gender-fixed in French. We introduce a rule-based system (RBS) tailored for the French language alongside two fine-tuned language models trained on data generated by our RBS. We also explore the use of instruct-based 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, several psycholinguistic studies have shown how language influences thoughts (Berlin and Kay, 1969; Kay and McDaniel, 1978). Further studies examining gender in language proved 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 (MG), that is masculine words that are supposed to refer to mixed groups of men and women or to someone who gender is unknown (Braun et al., 2005; Gygax et al., 2008, 2019; 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 MG 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 bias in the training data (Lu et al., 2020; Stanczak and Augenstein, 2021; Kotek et al., 2023; Ducel et al., 2024). This problem is particularly relevant for machine translation systems, which are highly susceptible to gender bias when translating between languages with different grammatical gender systems (Savoldi et al., 2021; Wisniewski et al., 2021; Vanmassenhove, 2024). Data augmentation, which involves balancing the amount of data for all genders in a specific language, has been suggested as a potential solution to debias NLP systems (Zhao et al., 2018). This led to an NLP task known as "gender rewriting," whose goal is to automatically propose alternatives to gendered sentences.

As of yet, automatic gender neutralization techniques have only been developed in English (Vanmassenhove et al., 2021; Sun et al., 2021). In French, the only gender rewriting system created is, to our knowledge, that of Lerner and Grouin (2024), which converts MG to gender-fair forms highlighting the feminine suffix[1]. Thus, we develop a French gender-neutral rewriting system using human collective nouns (CN), defined by Lecolle (2019) as "nouns referring to entities comprised of groups of individuals."[2] CNs have been widely discussed in the literature, especially when

[1] For example, "danseurs" (male dancers or MG) and "danseuses" (female dancers) can be merged into "danseur·euses" with an interpunct (·) to refer to both men and women.

[2] In French: "nom désignant une entité composée d'un ensemble d'individus humains."

Findings of the Association for Computational Linguistics: ACL 2025, pages 7889–7909
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

it comes to French (Flaux, 1999; Lammert, 2010; Lammert and Lecolle, 2014; Lecolle, 2019). Since, in French, this type of noun has a gender which does not depend upon the referent's,[3] it is an effective way of achieving gender neutralization. By focusing on gender neutralization, our work targets a writing technique that, compared to visibilization (used by Lerner and Grouin (2024)), tends to be less contentious among native French speakers, as it does not alter the spelling of existing words nor does it introduce non-standard or new punctuation marks to separate the feminine suffix from the base word form (Burnett and Pozniak, 2021). Finally, gender neutralization challenges the binary male/female gender dichotomy and is better adapted for people whose gender falls outside of the traditional categories. Our gender-neutral rewriting system, GeNRe (**Ge**nder-**N**eutral **Re**writing System Using French Collective Nouns), is the very first gender-neutral rewriting system for French[4]. We present three different versions of our system: one relying on a rule-based system (RBS), one relying on fine-tuned language models and one relying on an instruct-based model.

2 Gender in French

In French, nouns (N) are classified as either masculine or feminine, and the gender of a noun influences the form of determiners (D), adjectives (A) and past participle verbs (V) that are syntactically related. Similarly, coreferent pronouns (P), that is pronouns referring to a previously mentioned entity, also feature the same gender. Examples 1 (masculine) and 2 (feminine) highlight the syntactic differences that arise when using either an inanimate masculine noun ("courrier", *mail*) and an inanimate feminine noun ("lettre", *letter*).

(1) **Le**[D] courrier[N] **recommandé**[A] a été **écrit**[V] récemment. **Il**[P] est **adressé**[V] à son mari.

(The registered [m.] mail [m.] was written [m.] recently. It [m.] is addressed [m.] to her husband.)

(2) **La**[D] lettre[N] **recommandée**[A] a été **écrite**[V] récemment. **Elle**[P] est **adressée**[V] à son mari.

(The registered [f.] letter [f.] was written [f.] recently. It [f.] is addressed [f.] to her husband.)

The gender of human role nouns reflects the sociological gender of the referent and is motivated (for instance, "danseuse" refers to a female dancer), while gender of nouns referring to unanimated beings is arbitrary (Watbled, 2012).

The masculine gender for human nouns is considered to be the "default" gender in French, and can be used in a non-specific context (in the singular form, as in Example 3[5]) or to refer to groups of people composed of both men and women (in the plural form, as in Example 4).

(3) **Un professeur** doit savoir faire preuve d'autorité.

(A professor [m.] has to know how to show authority.)

(4) **Les lecteurs assidus** financent le journal chaque mois.

(Avid readers [m.] provide financial support to the newspaper every month.)

However, the use of masculine as the default gender (masculine generics, abbreviated MG) in gender-marked languages promotes androcentric mental representations and participates in invisibilizing women (Jacobson and Insko, 1985; Stahlberg et al., 2001). A 2017 survey conducted by Harris Interactive (2017) among French speakers found that respondents tend to have male-centric representations when MG are used in the questions. Similarly, according to Gabriel et al. (2018), MG human nouns are more likely to be associated with male referents, and specifically highlighting the generic nature of MG does not have an effect on the biased perception of survey participants (Gygax et al., 2012; Rothermund and Strack, 2024). Consequently, two main types of writing techniques can be used to avoid the use of MG: 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: *professeur·e*, as in Example 5) or by affecting the feminine ending directly (using capital or bold letters). Neutralization techniques, on the other hand, mainly revolve around various types of words: epicene words, that is words whose form is the same for masculine and feminine, whether they may have a generic (e.g. "personne", *person*, as in Example 6) meaning or a specific (e.g. "spécial-

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

[4]Code and data are made publicly available on GitHub, under license CC BY-SA 4.0 `https://github.com/spidersouris/GeNRe`

[5]In this example, "professeur" is considered as a masculine generics insofar as it does not refer to one specific male individual, but to any individual serving as "professor".

iste”, *specialist*) one, or words that refer to groups of people, such as CNs (e.g. “lectorat”, *readership*, as in Example 7), which we use for this work. CNs have a fixed gender which is not associated with the genders of the people within that group.

(5) **Un·e professeur·e** doit savoir faire preuve d’autorité.
(A [m./f.] professor [m./f.] has to know how to show authority.)

(6) **Une personne professeure** doit savoir faire preuve d’autorité.
(A person teaching has to know how to show authority.)

(7) **Le lectorat assidu** finance le journal chaque mois.
(The avid readership provides financial support to the newspaper every month.)

Given the impact of inclusive formulations on mitigating MG-induced gender biases (Koeser et al., 2015; Kollmayer et al., 2018; Xiao et al., 2023), developing a system to automatically rewrite text and reduce the prevalence of MG could be a valuable tool for data augmentation. By focusing on gender neutralization, our work aims to fill this gap and explore the potential of CNs and epicene words in promoting more inclusive language. While gender rewriting works focusing on automatic gender neutralization do exist in English (He et al., 2021; Sun et al., 2021; Vanmassenhove et al., 2021), no such efforts have been pursued for French. Similarly, studies about gender neutralization and its application to NLP tools exist in Italian (Piergentili et al., 2023a) and German (Lardelli and Gromann, 2023), but gender neutralization in NLP for French has not yet been addressed.

3 The Task of Gender Rewriting

In order to mitigate gender bias in NLP textual data, various systems have been developed to rebalance or debias datasets by generating alternative formulations when it comes to the use of grammatical gender. 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) define this task as: “generating alternatives of a given Arabic sentence to match different target user gender contexts.” While this definition works well for the work pursued by Alhafni et al. (2022b), 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. The system created by Pomerenke (2022) provides inclusive suggestions for input sentences in German and has led to the publication of an online resource letting the user choose the type of inclusive transformation to apply. More recently, Veloso et al. (2023) also developed an inclusive gender-rewriting system for Portuguese, and Lerner and Grouin (2024) for French. Finally, Sun et al. (2021), Vanmassenhove et al. (2021) and He et al. (2021) created systems to neutralize gender in an English input sentence, but no such system exists for French. As part of this work and in order to accommodate a larger number of languages and transformation types, we reframe the initial task definition given by Alhafni et al. (2022b) as “generating one or more alternative sentences that either neutralize gender, adopt inclusive forms, or switch to a different gender”.

4 Methodology

To build our automatic gender neutralization system, we propose three different approaches: a rule-based approach, a model fine-tuning approach, and an instruct-based model approach. We first create a dictionary of French CNs and their member noun counterparts, which will be used to make the necessary gender transformations. We describe this dictionary in Section 4.1. In Section 4.2, we then give details about the datasets that we extracted sentences from for language model (LM) fine-tuning from RBS data, and evaluation. Finally, in Section 4.3, we explain our experimental design with the aforementioned model types. While our work focuses specifically on French, the methodology presented below is applicable to any language which can use collective nouns as a gender-neutralizing technique (e.g., Spanish) given a dictionary of human-member nouns being already available or being created following our methodology. When it comes to syntactic changes, especially considering gender and number, those would be very similar in languages with similar inflection, such as Spanish, Italian or Portuguese. As a result, the amount of work needed to adapt our methodology to these languages specifically would be much lower compared languages encoding number or gender differently.

4.1 Dictionary

First, we manually created a dictionary with French CNs and their member noun counterparts. Three approaches were used to fill this dictionary: literature review, manual collecting and semi-automatic collecting.

Literature review. French CNs have been extensively studied in the linguistic literature. We drew on the list of 138 CNs by Lecolle (2019), the most exhaustive list of French CNs to our knowledge, which provided a comprehensive starting point for our dictionary. Some nouns were excluded from our dictionary due to their polysemy or restrictive semantics. For example, the CN "troupe" has multiple meanings (*troop*, *troupe*), and its use would require specifying the associated subdomain or group members to avoid confusion. Similarly, nouns like "duo" or "trio" have too restrictive a semantics because they only apply to groups of exactly two or three people, respectively. After careful selection, we retained 105 entries from Lecolle's list.

Manual collecting. We empirically collected CNs from media and Internet sources over an extended period, and used Sketch Engine (Kilgarriff et al., 2014) for corpus search. This approach allowed us to identify nouns not presented in the literature on CNs, providing a complementary perspective to the literature review. With this approach, we added 46 entries to our dictionary.

Semi-automatic collecting. We scraped the French version of Wiktionary[6] to retrieve CNs with the suffix "-phonie", which refer to the speakers of a language (e.g. "anglophonie", *English-speaking world*). We developed a Python script to generate equivalent CNs by replacing the suffix "-phonie" with "-phone" (e.g. "anglophone"). This approach enabled us to efficiently collect a set of nouns that follow a specific pattern, adding 164 entries, manually checked.

[6]https://fr.wiktionary.org/wiki/

In total, our dictionary thus contains 315 entries. 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 two different datasets made available for research purposes: a French Wikipedia dataset with 1.58 million texts (graelo, 2023)[7] and the Europarl corpus (Koehn, 2005), which was created from the proceedings of the European Parliament and available in 21 languages, including English and French. We extracted 292,076 sentences containing member nouns from the Wikipedia dataset. The Europarl corpus was filtered to include French sentences only, and 106,878 additional sentences were extracted for model fine-tuning and evaluation (total 398,954). We publish both filtered corpora.

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 CN 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 CNs "armée" (*army*) "bataillon" (*battalion*), "infanterie" (*infantry*) or "régiment" (*regiment*). As we used data generated by our rule-based system for model

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

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 ensure the correct member nouns will be replaced in the input sentence, as only those that are between tags will be taken into account. Example 8 shows how these tags are used.

(8) 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 fine-tuned language models, and an instruct-based language model. Each model takes a different approach to the task, allowing us to compare their performance.

4.3.1 Rule-based model

We developed an RBS to automatically apply the correct syntactic rules when converting a member noun into a CN, which leads to number and gender changes in the sentence (compare Examples 4 and 7). The RBS consists of two main components: a syntactic dependency detection component and a generation component. Figure 1 shows an overview of the RBS pipeline.

The dependency detection component primarily relies on spaCy (Montani et al., 2024) 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. We evaluated the performance of our rule-based dependency detection component in finding the correct member noun's dependencies to change compared to default spaCy-detected member noun's dependencies (excluding punctuation), which is our baseline. We computed precision, recall and F1. Results are shown in Table 2.

The generation component replaces each member noun in the sentence with its CN counterpart found in the dictionary, adjusting the determiner, handling elision, and re-inflecting the detected dependencies using *inflecteur* (Chuttarsing, 2021), a Python module leveraging the Delaf French morphological dictionary[8] and *french-camembert-postag-model*[9], a CamemBERT-based (Martin et al., 2020) part of speech (POS) tagging model for French. Our RBS also makes additional changes 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.

We evaluate the accuracy of adjustments made by *inflecteur* (as no official evaluation has been conducted) and of the additional changes made by our RBS for past participles and object pronouns. Two annotators annotated the correct inflection of syntactic dependencies found in the 500 sentences of our evaluation dataset introduced in Section 4.2 (129 dependencies for Wikipedia; 115 for Europarl: 264 in total), where $\kappa = 0.947$. *inflecteur* without additional RBS changes achieves 73.01% accuracy. With our RBS, it achieves 75.35% accuracy (+2.34 improvement).

4.3.2 Fine-tuned models

Previous research on gender rewriting has focused on training neural models as well as fine-tuning LMs 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 LMs 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 LMs, t5-small (Raffel et al., 2020) and m2m100_418M (Fan et al., 2020), were selected for the experiments. t5-small is a 60-million-parameter model trained on English, French, Romanian, and German. M2M100 is a multilingual model with 418 million parameters, trained on 100 languages including French. These models

[8] https://uclouvain.be/fr/instituts-recherche/ilc/cental/delaf-2-0.html
[9] https://huggingface.co/gilf/french-camembert-postag-model

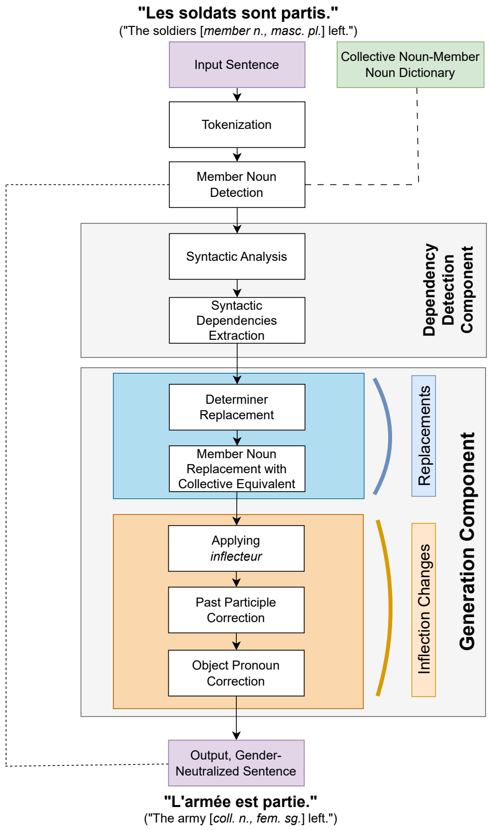

Figure 1: Rule-based model replacement pipeline overview

were chosen for their great text-to-text performance and their relatively small size, making the training process easier. Furthermore, as M2M100 had already been used by Veloso et al. (2023), we want to compare the results we can get for our specific task. Both models 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 Instruct-based model

The rapid development of large language models (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 instruct-based models such as InstructGPT (Ouyang et al., 2022), or, more recently, Llama 3 (Grattafiori et al., 2024) or DeepSeek-V3 (DeepSeek-AI et al., 2024). This is primarily achieved through the use of "prompts" or instructions given to the LLM (Liu et al., 2021). Several studies (Nunziatini and Diego, 2024; Bartl and Leavy, 2024; Veloso et al., 2023) have shown how instruct-based models may be used to automatically generate non-gendered-biased texts, including for translation tasks (see Jourdan et al. (2025) for English to French). However, no study has yet analyzed their capabilities for the specific task of gender rewriting in French. We chose Claude 3 Opus `claude-3-opus-20240229` due to its best text generation performance at the time of the experiments (Anthropic, 2024) and its API being free to use during the period the experiments were conducted.[10]

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 in Appendix B.

- The "BASE" instruction provides a basic task description, asking the model to make the sentence inclusive by replacing MG with their CN equivalents, without explicitly specifying the replacement word.
- The "DICT" instruction leverages our collective noun dictionary and asks the model to replace MG with their corresponding CNs, those being explicitly mentioned. There are two different versions for the "DICT" instruction: "DICT-SG", used when only one generic masculine noun with a matching CN was found in the sentence, and "DICT-PL", used when several generic masculine nouns with matching CNs 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 use the evaluation dataset presented in Section 4.2 and we leverage three evaluation metrics: Word Error Rate (WER), BLEU (Papineni et al., 2002) and cosine similarity. WER and BLEU have been commonly used in previous gender rewriting works. WER measures differences

[10]For the announcement, see https://www.anthropic.com/news/claude-3-family.

	Wikipedia			Europarl			Avg.		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Baseline	0.096	0.723	0.169	0.115	0.689	0.197	0.1055	0.706	0.183
GeNRe-RBS	0.773	0.855	0.812	0.758	0.813	0.785	0.7655	0.834	0.7985

Table 2: Results of the RBS dependency detection component per corpus and on average

	Wikipedia			Europarl			Avg.		
	WER (↓)	BLEU (↑)	Cosine (↑)	WER (↓)	BLEU (↑)	Cosine (↑)	WER (↓)	BLEU (↑)	Cosine (↑)
Baseline	12.611%	80.688	97.436	12.446%	82.871	97.008	12.529%	81.779	97.222
GeNRe-RBS	**4.03%**	92.096	**98.88**	**3.814%**	93.707	99.22	**3.81%**	92.887	**99.05**
GeNRe-T5	6.726%	87.358	98.508	4.259%	93.111	99.1	5.492%	90.234	98.804
GeNRe-M2M100	6.566%	88.186	97.232	4.247%	93.197	98.992	5.406%	90.692	98.112
Claude-BASE	13.87%	80.205	96.532	10.713%	85.313	97.128	12.291%	82.759	96.83
Claude-DICT	4.702%	**92.79**	98.812	4.197%	**94.247**	**99.264**	4.45%	**93.519**	99.038
Claude-CORR	11.282%	84.954	98.092	8.992%	85.257	98.056	10.137%	85.25	98.074

Table 3: Results by model type and corpus. Bold indicates the best results overall.

between the reference sentences (manually gender-rewritten) and the transformed sentences by taking into account word insertions, deletions and substitutions, and was previously used by Sun et al. (2021) and Vanmassenhove et al. (2021). BLEU, on the other hand, measures the n-gram overlap between the reference and the transformed sentences ($n = 4$), and was previously used by Sun et al. (2021), Veloso et al. (2023) and Lerner and Grouin (2024). Using these metrics can help compare results from previous works (even though languages may be different) and gives an overview of the similarity between the references and the transformations (for a more precise analysis, see error labelling in Section 6). To compute WER and BLEU, we respectively use JiWER 3.0.3[11] and sacrebleu 2.4.2[12] (Post, 2018) Python packages with default parameters. In addition, we also use cosine similarity, which has not been used in previous works. Cosine similarity evaluates the semantic closeness between sentence embeddings of the reference and transformed sentences. Recent work by Piergentili et al. (2023b) argues that semantic similarity metrics may be ill-suited for evaluating gender-neutral rewritings as they are robust to form differences. We acknowledge this limitation and use cosine similarity not as a primary signal of neutralization quality, but as a complementary measure of meaning preservation, alongside form-sensitive metrics like BLEU and WER. For computation, we use SBERT (Reimers and Gurevych, 2019) and model `Sentence-CamemBERT-Large`, based on CamemBERT (Martin et al., 2020), for its best performances for French. The baseline corresponds to the unchanged sentence in line with previous works. Results of each model per corpus and on average are available in Table 3.

The RBS and Claude 3 Opus-DICT achieved the best results in our experiments, with the RBS achieving 3.81% WER and 99.05 cosine similarity, and Claude 3 Opus-DICT achieving 93.519 BLEU. The fine-tuned models also showed mostly promising results, even though lower than the RBS and Claude 3 Opus DICT (5.492% WER, 90.234 BLEU and 98.112 cosine similarity for T5; 5.406% WER, 90.692 BLEU and 98.112 cosine similarity for M2M100). Comparing the two fine-tuned models, they achieved similar results, with the T5 model slightly outperforming M2M100. However, both models showed a minor decrease in performance compared to the RBS. As a result, similarly to Veloso et al. (2023) and in contrast with the findings of Vanmassenhove et al. (2021), we do not find a significant improvement compared to our RBS following fine-tuning. In comparison with previous automatic gender neutralization works in English (Sun et al., 2019; Vanmassenhove et al., 2021), we achieve similar (although lower for WER) difference scores between the baseline and the RBS (-11.77 WER and +9.63 BLEU for Sun et al. (2019); -10.947 WER for Vanmassenhove et al. (2021); -8.449 WER and +11.10 BLEU for GeNRe-RBS).

6 Discussion

We conduct a quantitative and qualitative analysis of errors made by each model type. This analysis is done manually for the RBS and fine-tuned models, and automatically for the instruct-based model.

[11] https://pypi.org/project/jiwer/
[12] https://pypi.org/project/sacrebleu/

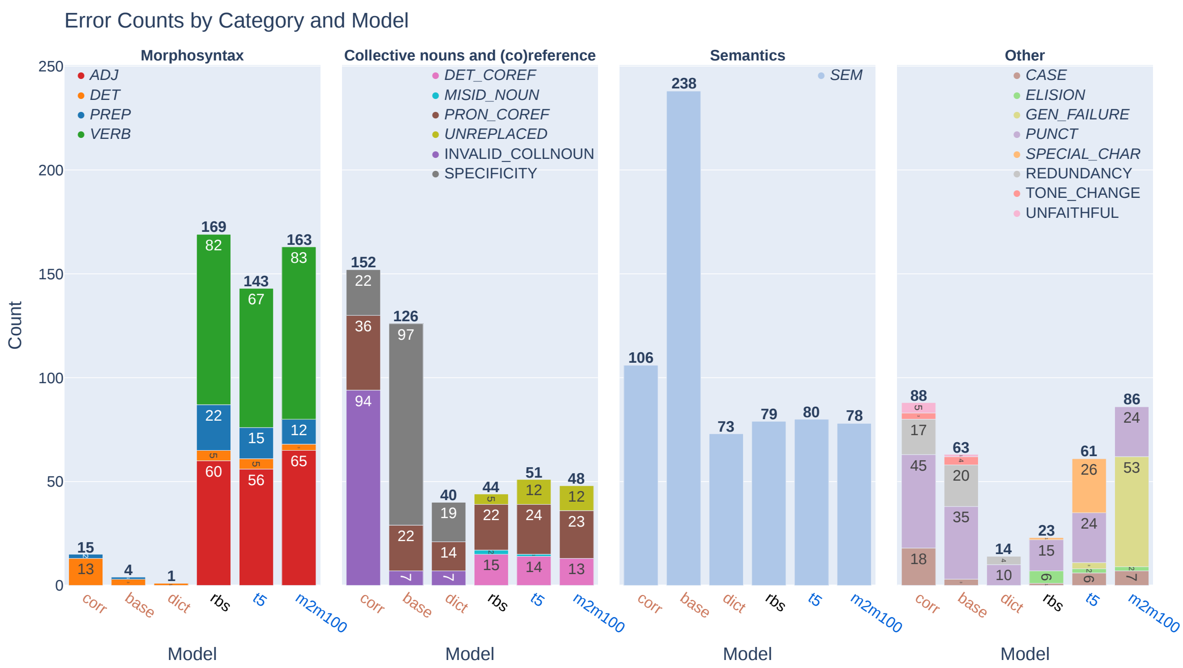

Figure 2: Error distribution for RBS, fine-tuned and instruct-based models

The latter method is chosen because LLMs tend to be significantly more permissive in their generations, resulting in a broader range of error types that would complicate manual annotation. We define four high-level categories of errors, as illustrated in Figure 2: morphosyntax; collective nouns and (co)reference; semantics; and other.

For GeNRe-RBS, GeNRe-T5, and GeNRe-M2M100, errors were manually annotated across both corpora and verified by two annotators. Annotation details are available in Appendix D.1. Based on the outputs from the RBS and fine-tuned models, we established 14 subcategories of errors (italicized in Figure 2), which were then applied to each sentence generated by each model (500 sentences × 3 models). Multiple error labels can be assigned to a single sentence. Descriptions of each error type can be found in Table 7 in the Appendix.

For Claude 3 Opus and its instructions (BASE, DICT, CORR), we use the "LLM-as-judge" approach (Zheng et al., 2023) to automate error annotation using GPT-4o mini (OpenAI, 2024) and in-context learning (Brown et al., 2020). First, GPT-4o mini is prompted to provide a brief explanation of the error(s) in each generated sentence compared to the reference. We generate 10 different outputs and identify the most frequent explanation by comparing sentence-level embeddings using SBERT (Reimers and Gurevych, 2019) and Sentence-CamemBERT-Large[13]. Subcategories of errors are then assigned to each sentence based on these explanations and the manually defined taxonomy. The most frequent subcategory across the 10 generations is retained. Prompts are included in Appendix E. We validate these automatically assigned subcategories by reviewing selected model outputs to ensure consistency, filtering out irrelevant or incorrectly attributed labels, and merging valid labels with those used in manual evaluation. Figure 2 shows the distribution of errors across models, grouped into the four main categories.

Morphosyntactic errors are most frequent in the RBS model (169 instances), though slightly reduced in the fine-tuned models (163 for M2M100, 143 for T5). In both cases, errors involving adjectives (ADJ) and verbs (VERB) are most prevalent. 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 CN. 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 9a, the verb "veulent" (pl., *want*) was correctly changed by M2M100 to "veut" (sg.) to match with the new CN "actorat" (*group*

[13] https://huggingface.co/dangvantuan/sentence-camembert-large

of actors). Thus, despite their overall lower performance, these models show promise for correcting such errors. The instruction-tuned model variants exhibit significantly fewer such errors, reflecting their great linguistic capabilities.

For collective noun and (co)reference errors, the "INVALID_COLLNOUN" label denotes the failure to use a valid collective noun present in the reference sentence: it is typically replaced by a MG (Example 9b). The "SPECIFICITY" error captures shifts in noun specificity, such as changes from definite to indefinite forms.

Semantic errors are treated as a distinct category due to their frequency. All models tend to struggle with these, likely because the use of collective nouns is governed by strict semantic rules that limit their applicability in many contexts. The BASE instruction model exhibits a notably higher number of such errors (238), probably due to the increased generative freedom permitted by the prompt, which leads to the use of inappropriate or non-existent collective nouns (Example 9d).

When it comes to other types of errors, errors observed in the fine-tuned models and different from the RBS included token generation failures (M2M100 mostly, Example 9e, where *"Nebski" was generated instead of "Zemski"), and incorrect generation of special characters (T5 only, as in Example 9f 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 9g. Additional examples and their translations can be found in Appendix F.

(9) a. [...] le soutien apporté à la Commission à **l'actorat local** qui (**veulent**$^{\text{RBS}}$ | **veut**$^{\text{M2M100}}$) participer [...]
([...] the support given to the Commmission of the **local group of actors** who (**want [pl.]**$^{\text{RBS}}$ | **want [sg.]**$^{\text{M2M100}}$) to participate [...])

b. Le point culminant est une attaque contre (**la rébellion cachée**$^{\text{ORIG}}$ | **les rebelles cachés**$^{\text{OPUS-CORR}}$) dans les montagnes [...]
(The climax is an attack against (**the rebellion hidden [fem. sg.]**$^{\text{ORIG}}$ | **the rebels hidden [masc. pl.]**$^{\text{OPUS-CORR}}$) in the mountains [...])

c. L'armée nigérienne déplore également la perte d'un blindé type 92A, détruit par (**les soldats nigériens**$^{\text{ORIG}}$ | **l'armée nigérienne**$^{\text{RBS}}$), de 6 autres véhicules armés, de deux canons de 122mm [...]
(The Nigerian army also deplores the loss of a 92A armored vehicle, destroyed by (**Nigerian soldiers**$^{\text{ORIG}}$ | **the Nigerian army**$^{\text{RBS}}$), as well as 6 other armed cars, two 122-mm cannons [...])

d. En 1531, pour payer sa dette (**aux marchands**$^{\text{ORIG}}$ | **au négoce**$^{\text{OPUS-BASE}}$) de Lübeck, le roi Gustav Vasa [...]
(In 1531, to pay his debt (**to the dealers**$^{\text{ORIG}}$ | **to the business**$^{\text{OPUS-BASE}}$) of Lübeck, the king Gustav Vasa [...])

e. Juin, Russie : le (**Zemski**$^{\text{ORIG}}$ | **Nebski**$^{\text{M2M100}}$) sobor prend des décisions importantes.
(June, Russia: the (**Zemski**$^{\text{ORIG}}$ | **Nebski**$^{\text{M2M100}}$) Sobor makes important decisions.)

f. Il est allé à Cologne, où il est devenu président de l'association de la (**main-d'œuvre**$^{\text{ORIG}}$ | **main-d'uvre**$^{\text{T5}}$) [...]
((He went to Cologne, where he became president of the **labour** organization [...])

g. (**L'**$^{\text{ORIG}}$ | **l'**$^{\text{T5}}$)armée arriva avec une lance à eau pour disperser les détenus.
((**The**$^{\text{ORIG}}$ | **The**$^{\text{T5}}$) army arrived with a water hose to disperse the prisoners.)

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 CNs 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 CNs. Our experiment combining our manually created dictionary with the Claude 3 Opus instruct-based model also shows promise for the use of such models for the task of gender rewriting. We 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 other languages with collective nouns (such as Spanish) or similar gender neutralization techniques.

Limitations

French CNs 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 CN equivalents are both contextually relevant and semantically appropriate.

Furthermore, even though collective nouns have not been tested specifically, recent research works from Spinelli et al. (2023) and Tibblin et al. (2023) showed that gender neutralization appears to be less effective to counter gender biases induced by the use of MG. As previously stated, however, this writing technique is less contentious among the general population compared to others which explicitly highlight the feminine ending of words or separate it from the masculine ending.

The LLM-as-a-judge approach used for automatic instruct-based error analysis produces results that are not directly comparable with human annotation. Even if multiple steps were taken to generate consistent results, for instance by only retrieving the most common generations across multiple tries, some works have highlighted the limits of LLMs as judges (Gu et al., 2025).

Finally, this work is limited to the French language only, and the methodology we resorted to can only be used by languages with collective nouns acting as gender neutralizers (e.g., Spanish) and requires the creation of a language-specific human-member noun dictionary.

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 in sentences, the generation of such content seems unlikely. As discussed in the paper, instruct-based 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, LLMs may be trained on such data, which might lead to the generation of harmful or hateful content.

Acknowledgements

We sincerely thank the reviewers for their constructive comments and suggestions, which contributed substantively in improving this paper.
This research was supported by the LiRiC Interdisciplinary Thematic Institute at the University of Strasbourg.

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. http://arxiv.org/abs/2210.12410. *Preprint*, arXiv:2210.12410.

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

Marion Bartl and Susan Leavy. 2024. From 'Showgirls' to 'Performers': Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs. *Preprint*, arXiv:2407.04434.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models Are Few-Shot Learners. http://arxiv.org/abs/2005.14165. *Preprint*, arXiv:2005.14165.

Heather Burnett and Céline Pozniak. 2021. Political dimensions of gender inclusive writing in Parisian universities. *Journal of Sociolinguistics*, 25(5):808–831.

Adrien Chuttarsing. 2021. Inflecteur.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. Évaluation automatique des biais de genre dans des modèles de langue auto-régressifs. *TALN 2024*.

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. http://arxiv.org/abs/2010.11125. *Preprint*, arXiv:2010.11125.

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

Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. 2018. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858.

graelo. 2023. Graelo/wikipedia.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Pascal Gygax, Ute Gabriel, Arik Lévy, Eva Pool, Marjorie Grivel, and Elena Pedrazzini. 2012. The masculine form and its competing interpretations in French: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology*, 24(4):395–408.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, 23(3):464–485.

Pascal Mark Gygax, Lucie Schoenhals, Arik Lévy, Patrick Luethold, and Ute Gabriel. 2019. Exploring the Onset of a Male-Biased Interpretation of Masculine Generics Among French Speaking Kindergarten Children. *Frontiers in Psychology*, 10:1225.

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.

Harris Interactive. 2017. L'écriture inclusive : La population française connaît-elle l'écriture inclusive ? Quelle opinion en a-t-elle ? Technical report.

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

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.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity. *Preprint*, arXiv:2504.15941.

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

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation.

Sara Koeser, Elisabeth A. Kuhn, and Sabine Sczesny. 2015. Just reading? how gender-fair language triggers readers' use of gender-fair forms. *Journal of Language and Social Psychology*, 34(3):343–357.

Marlene Kollmayer, Andreas Pfaffel, Barbara Schober, and Laura Brandt. 2018. Breaking away from the male stereotype of a specialist: Gendered language affects performance in a thinking task. *Frontiers in Psychology*, Volume 9 - 2018.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

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.

Manuel Lardelli and Dagmar Gromann. 2023. Gender-Fair Post-Editing: A Case Study Beyond the Binary.

Michelle Lecolle. 2019. *Les Noms Collectifs Humains En Français. Enjeux Sémantiques, Lexicaux et Discursifs*. Lambert-Lucas, Université de Lorraine.

Paul Lerner and Cyril Grouin. 2024. INCLURE: A Dataset and Toolkit for Inclusive French Translation.

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. http://arxiv.org/abs/2107.13586. *Preprint*, arXiv:2107.13586.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender Bias in Neural Natural Language Processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security*, volume 12300, pages 189–202. Springer International Publishing, Cham.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte 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.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2024. spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.

Mara Nunziatini and Sara Diego. 2024. Implementing gender-inclusivity in MT output using automatic post-editing with LLMs. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 580–589, Sheffield, UK. European Association for Machine Translation (EAMT).

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

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. http://arxiv.org/abs/2203.02155. *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.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. From Inclusive Language to Gender-Neutral Machine Translation. http://arxiv.org/abs/2301.10075. *Preprint*, arXiv:2301.10075.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus. *Preprint*, arXiv:2310.05294.

David Pomerenke. 2022. INCLUSIFY: A Benchmark and a Model for Gender-Inclusive German. http://arxiv.org/abs/2212.02564. *Preprint*, arXiv:2212.02564.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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. http://arxiv.org/abs/1910.10683. *Preprint*, arXiv:1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Patrick Rothermund and Fritz Strack. 2024. Reminding May Not Be Enough: Overcoming the Male Dominance of the Generic Masculine. *Journal of Language and Social Psychology*, 43(4):468–485.

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.

Elsa Spinelli, Jean-Pierre Chevrot, and Léo Varnet. 2023. Neutral is not fair enough: Testing the efficiency of different language gender-fair strategies. *Frontiers in Psychology*, 14:1256779.

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.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. http://arxiv.org/abs/2112.14168. *Preprint*, arXiv:2112.14168.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 10.

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.

Julia Tibblin, Jonas Granfeldt, Joost Van De Weijer, and Pascal Gygax. 2023. The male bias can be attenuated in reading: On the resolution of anaphoric expressions following gender-fair forms in French. *Glossa Psycholinguistics*, 2(1).

Eva Vanmassenhove. 2024. Gender Bias in Machine Translation and The Era of Large Language Models. *Preprint*, arXiv:2401.10016.

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

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2021. Screening Gender Transfer in Neural Machine Translation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 311–321.

Hualin Xiao, Brent Strickland, and Sharon Peperkamp. 2023. How fair is gender-fair language? insights from gender ratio estimations in french. *Journal of Language and Social Psychology*, 42(1):82–106.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Preprint*, arXiv:1804.06876.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Fine-Tuning Details

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

A.1 GeNRe-T5

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

A.2 GeNRe-M2M100

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

B Instruction Details

B.1 Instruct-Based Model Hyperparameters

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

B.2 Types of Instructions

Table 4 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 instruct-based model. See Tables 5 and 6 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 instruct-based model is guided towards generating the final, gender-neutralized sentence.

C Few-shot sentences given to Claude 3 Opus

Tables 5 and 6 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].

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

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 5: Few-shot sentences for "BASE" and "DICT" instructions. Bold indicates the differences between sentences with MG 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 6: Few-shot sentences for "CORR" instruction. Bold indicates the differences between the RBS-generated ssentences with error and the manual, correct sentences.

D Manual Error Type Labelling Details

We give additional information about some of the error types below.

Error Type	Description
ADJ	Errors related to adjective agreement with the modified noun. Past participles used as adjectives are included in this category.
CASE (N)	Errors related to an incorrect use of lowercase/uppercase characters.
DET	Errors related to determiner agreement with the modified noun.
DET_COREF	Errors related to coreferent possessive determiner agreement with the modified noun.
ELISION	Errors related to elision.
GEN_FAILURE (N)	Errors related to incorrect text-to-text model generation, most particularly with proper nouns or words that are not part of the model's vocabulary.
MISID_NOUN	Errors occurring when a member noun's form in the collective-member noun dictionary was wrongly detected as a noun in the original sentence, and was thus incorrectly changed into a CN.
PREP	Errors related to preposition usage.
PRON_COREF	Errors related to coreferent pronoun agreement with the modified noun.
PUNCT	Errors related to punctuation (e.g. missing or double spaces).
SEM	Errors occurring when changing the member noun into its CN counterpart leads to an asemantic sentence.
SPECIAL_CHAR (N)	Errors related to special characters (e.g. accents).
UNREPLACED	Errors occurring when the member noun was not replaced with its CN counterpart.
VERB	Errors related to verb or auxiliary agreement.

Table 7: Error types and descriptions for the RBS and fine-tuned models

Label	**Wikipedia**			**Europarl**		
	Agree	**Disagree**	**Rate (%)**	**Agree**	**Disagree**	**Rate (%)**
ADJ	88	26	77.19	56	58	49.12
PREP	24	10	70.59	8	6	57.14
PRON_COREF	8	8	50.00	46	55	45.54
SEM	128	104	55.17	40	109	26.85
VERB	91	45	66.91	71	74	48.97

Table 8: Main labels interannotator agreement for Wikipedia and Europarl corpora across all 3 non-instruct-based models (RBS, T5, M2M100)

The ELISION error is related to how elision works in French: in the sentences that we are modifying, the masculine determiner "le" and the feminine determiner "la" (*the*) should be elided and written as "l'" when the word that follows begins with a vowel or a mute "h".

The MISID_NOUN error may occur when the form of a member noun shares several different grammatical categories. For example, "jeunes" (*young*), the member noun's form of the CN "jeunesse" (*youth*), can be both a noun and an adjective. When the adjective form was wrongly detected as a noun, it was included in our dataset and produced an ungrammatical result sentence.

Finally, when it comes to the SEM error type, as discussed by Lecolle (2019), CNs in French, and more specifically human CNs, 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 CNs 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.

We report the interannotator agreement for main labels for Wikipedia and Europarl corpora annotations in Table 8. It should be noted that agreement for the SEM label is the lowest: this is due to the fact that this type of error has a subjectivity componet, insofar as some sentences, depending on the context, may sound semantic to one person but asemantic to another one. This leads to a difficulty of pinpointing accurately what parts of the sentence may sound odd to a native speaker or may not convey the intended meaning. Moreover, the fact that CNs are relatively uncommon in everyday language and are not as popular as other techniques for gender neutralization makes undiscussed agreement even more difficult. For other labels, interannotator agreement may also be affected by the subtle nature of the changes between sentence pairs (in many cases, the differences are minimal and not immediately perceptible) as well as classification disagreements (past participles categorized as VERB vs. ADJ for example) that were later resolved. All applied labels for all sentences and models were agreed upon after discussion.

D.1 Labelling Instructions Given to Annotators (French)

La tâche consiste à annoter les erreurs présentes dans les trois types de phrases générées (par le système à base de règles, T5 et M2M100) pour les deux corpus (Wikipédia et Europarl). Une même phrase peut être annotée avec plusieurs types d'erreurs différents. Si une erreur est répétée, un seul label de même type est appliqué. Les types d'annotations d'erreurs sont définis ci-dessous :

ADJ : erreurs qui portent sur les adjectifs (accord) (p. ex. *« les soldats allemand »)

CASE : erreurs qui portent sur l'utilisation des majuscules/minuscules (p. ex., nom propre non mis en majuscule)

DET : erreurs qui portent sur les déterminants (p. ex. *« le police »)

DET_COREF : erreurs survenues lorsqu'un déterminant possessif qui se rapporte au nom collectif n'a pas été modifié ou a été modifié à tort, (p. ex. « Il a rejoint la police dans leurs locaux » [au lieu de « ses »])

ELISION : erreurs qui portent sur l'élision (p. ex. *« la armée »)

GEN_FAILURE : erreurs qui émanent d'une mauvaise génération du modèle, notamment par exemple sur les noms propres ou les mots inconnus du modèle (p. ex. « Society of Friends » → « Society ofamina »). Dans certains cas, cela peut aussi mener à des déformations de mots par l'ajout ou la suppression d'espaces (p. ex. *« L'organisationPromet »). Dans ce cas, annoter avec GEN_FAILURE et non pas avec SPACE.

MISID_NOUN : cas où une forme présente dans le dictionnaire a été incorrectement identifiée comme nom et donc modifié comme nom collectif, ce qui rend la phrase asémantique/agrammaticale. Dans ce cas, annoter avec MISID_NOUN et non pas avec SEM.

PREP : erreurs relatives à l'utilisation des propositions (p. ex. *« invasion par de l'armée ennemie »)

PRON_COREF : erreurs survenues lorsqu'un pronom possessif qui se rapporte au nom collectif n'a pas été modifié ou a été modifié à tort (p. ex. « Au gouvernement, les manifestants leur ont exprimé leur colère » [au lieu de « lui »])

PUNCT : erreurs qui portent sur la ponctuation, les espaces

SEM : phrase jugée asémantique après le remplacement du nom de membre par un nom collectif (par exemple, dans des structures numérales de type « l'un des auteurs » → *« l'un de l'autorat » / « les millions de soldats » → ?« les millions de l'armée »)

SPECIAL_CHAR : erreurs qui portent sur les caractères spéciaux (p. ex. accents)

UNREPLACED : cas où le nom de membre n'a pas été remplacé

VERB : erreurs qui portent sur les verbes (conjugaison) ou les auxiliaires (p. ex. *« les soldats parte »)

E GPT 4o-mini Automatic Error Type Labelling

E.1 Error Type Explanations Generation Prompts

Listing 1: System Prompt

```
You are an assistant whose task is to quickly summarize error types
    comparing a golden and an LLM-rewritten sentence.
```

Listing 2: User Prompt

```
Given a rewritten golden sentence and an LLM-rewritten sentence,
   compare the two and write a few words about the errors found in
   the LLM-rewritten sentence (if any) compared to the golden
   sentence. Each error type explanation should be separated with a
   semicolon and written in lowercase. Each explanation should
   start with a past tense verb. Only output the explanations and
   keep it short.

<examples>
Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]
</examples>

<info>
For your information, the LLM-rewritten sentence was generated
   after the LLM received this prompt: [PROMPT GIVEN TO INSTRUCTION
   -TUNED MODEL]
</info>

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations:
```

E.2 Error Labels Generation Prompts

Listing 3: System Prompt

```
You are an assistant whose task is to generate short error labels
    based on error descriptions.
```

Listing 4: User Prompt

```
Given a rewritten golden sentence, an LLM-rewritten sentence and an
    error type explanation, generate error labels for the LLM-
    rewritten sentence. Each error label should be separated with a
    semicolon (without spaces) and written in uppercase. If no error
    is found, output nothing. Only output the labels.

For reference, the previous labels have already been generated: [
    PREVIOUSLY GENERATED LABELS, IF ANY]
If any of the labels corresponds to the error type explanation, use
    it. Otherwise, feel free to generate a new label.

<examples>
Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]
Error labels: [LABEL EXAMPLES]

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]
Error labels: [LABEL EXAMPLES]

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]
Error labels: [LABEL EXAMPLES]
</examples>

<info>
For your information, the LLM-rewritten sentence was generated
    after the LLM received this prompt: [PROMPT GIVEN TO INSTRUCT-
    BASED MODEL]
</info>

Golden sentence: [GOLDEN SENTENCE EXAMPLE]
LLM-rewritten sentence: [LLM-REWRITTEN SENTENCE EXAMPLE]
Error type explanations: [EXPLANATION EXAMPLE]
Error labels:
```

F Generation Examples

(F10) 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.)

(F11) a. Je vous invite à informer **les députés 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.)

(F12) 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-M2M100]
(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.)

(F13) 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.)

(F14) 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.)

(F15) 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.