GFG - Gender-Fair Generation:
A CALAMITA Challenge

Simona Frenda[1,2,*,†], Andrea Piergentili[3,4,*,†], Beatrice Savoldi[3], Marco Madeddu[5],
Martina Rosola[6], Silvia Casola[7], Chiara Ferrando[5], Viviana Patti[5], Matteo Negri[3] and
Luisa Bentivogli[3]

[1]Interaction Lab, Heriot-Watt University, Edinburgh, Scotland

[2]aequa-tech, Turin, Italy

[3]Fondazione Bruno Kessler, Trento, Italy

[4]University of Trento, Trento, Italy

[5]Computer Science Department, University of Turin, Turin, Italy

[6]Universitat de Barcelona, Barcelona, Spain

[7]MaiNLP & MCML, LMU Munich, Germany

Abstract

Gender-fair language aims at promoting gender equality by using terms and expressions that include all identities and avoid reinforcing gender stereotypes. Implementing gender-fair strategies is particularly challenging in heavily gender-marked languages, such as Italian. To address this, the Gender-Fair Generation challenge intends to help shift toward gender-fair language in written communication. The challenge, designed to assess and monitor the recognition and generation of gender-fair language in both mono- and cross-lingual scenarios, includes three tasks: (1) the detection of gendered expressions in Italian sentences, (2) the reformulation of gendered expressions into gender-fair alternatives, and (3) the generation of gender-fair language in automatic translation from English to Italian. The challenge relies on three different annotated datasets: the GFL-it corpus, which contains Italian texts extracted from administrative documents provided by the University of Brescia; GeNTE, a bilingual test set for gender-neutral rewriting and translation built upon a subset of the Europarl dataset; and Neo-GATE, a bilingual test set designed to assess the use of non-binary neomorphemes in Italian for both fair formulation and translation tasks. Finally, each task is evaluated with specific metrics: average of F1-score obtained by means of BERTScore computed on each entry of the datasets for task 1, an accuracy measured with a gender-neutral classifier, and a coverage-weighted accuracy for tasks 2 and 3.

Keywords

Gender-fair language, Inclusive language, Unfairness detection, Machine translation, Generation, Neomorphemes

1. Challenge: Introduction and Motivation

Gender-fair language, also known as inclusive language, consists in using linguistic expressions that promote gender equality, inclusion of non-binary identities, and avoid reinforcing gender stereotypes [1].

In order to pursue the goals of fairness and inclusiveness, measures that take into account the importance of the correlation between language and gender become central. Especially in heavily gender-marked languages such as Italian, the use and application of gender-fair strategies is an urgent and yet difficult challenge. Indeed, in these languages, several are the elements one has to take into account to ensure a gender-fair use of language. However, adopting a gender-fair language is crucial given the negative effects of the masculine generics, documented in a range of empirical studies [2, 3]; and recent years witnessed an increase in awareness and effort to address these issues by promoting gender-fair language [4].

In Italian, the masculine is not only used to refer to and address men but also generic or unknown individuals; mixed-gender groups, regardless of the proportion of genders of its members; women, typically when occupying prestigious roles; and genderqueer people, given that there is no codified grammatical gender for referring to them [5]. This use, though, makes women and gen-

✉ s.frenda@hw.ac.uk (S. Frenda); apiergentili@fbk.eu
(A. Piergentili); bsavoldi@fbk.eu (B. Savoldi);
marco.madeddu@unito.it (M. Madeddu);
martina.rosola@gmail.com (M. Rosola); s.casola@lmu.de
(S. Casola); chiara.ferrando@unito.it (C. Ferrando);
viviana.patti@unito.it (V. Patti); negri@fbk.eu (M. Negri);
bentivo@fbk.eu (L. Bentivogli)

🆔 0000-0002-6215-3374 (S. Frenda); 0000-0003-2117-1338
(A. Piergentili); 0000-0002-3061-8317 (B. Savoldi);
0009-0004-5620-0631 (M. Madeddu); 0000-0002-8891-352X
(M. Rosola); 0000-0002-0017-2975 (S. Casola); 0000-0001-5991-370X
(V. Patti); 0000-0002-8811-4330 (M. Negri); 0000-0001-7480-2231
(L. Bentivogli)

derqueer people invisible, giving rise to a proper injustice [6, 7, 8]. Extensive empirical literature also highlights how certain gendered expressions influence our cognition, with masculine terms evoking male images and reducing, e.g, the likelihood of women applying for or being considered suitable for a job position (for an overview see [9, 10]).

Crucially, such unfair linguistic practices are perpetuated in language technologies [11]. This becomes particularly evident in languages, like Italian, for which NLP tools often adopt masculine and stereotypical representations, making undue binary gender assumptions [12].

We propose the **Gender-Fair Generation challenge** at CALAMITA 2024 [13], whose goal is to reduce the use of gender-unfair expressions in written Italian, focusing on both monolingual and cross-lingual scenarios (English-Italian). Our challenge is **structured into three tasks**—*i)* gendered language detection, *ii)* fair reformulation, and *iii)* fair translation—**across three different datasets**. Namely, the newly created GFL-it corpus, composed of Italian texts extracted from 35 documents provided by the academic administration office of the University of Brescia and annotated following specific guidelines [1]; GeNTE, a bilingual test set for gender-neutral rewriting and translation built on a subset of the Europarl dataset [14]; and Neo-GATE, a bilingual test set designed to evaluate the use of nonbinary neomorphemes in Italian [15].[1] We combine and repurpose these datasets across the three tasks envisioned in the Gender-Fair Generation challenge.

This report is structured as follows: in Section 2, we provide a description of our challenge; in Section 3, we present the three datasets in detail; in Section 4, we describe the metrics involved in our task; in Section 5, we describe the limitations of our work, and finally, in Section 6, we discuss the ethical issues.

2. Challenge: Description

The Gender-Fair Generation challenge is organized into three tasks, which we present in detail below.

1) Gendered language detection: the first task tests the models' ability to identify referentially gender-marked expressions within Italian sentences, namely those expressions whose (typically grammatical) gender is linked to their human referent. Referentially gendered (henceforth simply *gendered*) language includes:

- the `overextended` masculine or feminine, i.e., the use of a single gendered expression to refer to

persons belonging to a mixed-gender group - e.g., *i cittadini* (the.M citizens:M) used for a group of citizens of different genders;
- the `generic` masculine or feminine, i.e., the use of a single gendered expression to refer to a generic or unknown person - e.g., *il candidato deve avere tutti i requisiti* (the.M candidate:M has to possess all the requirements);
- the `incongruous` gender, i.e., the use of a grammatical gender that does not match the referent's gender - e.g., *il professore ordinario Maria Rossi* (the.M full.M professor:M Maria Rossi).

2) Fair reformulation: the second task tests models' ability to rewrite gendered expressions into alternative gender-fair expressions. To achieve this goal, various gender-fair language strategies can be employed. In particular, we will employ *obscuration* strategies:

- *conservative obscuration*, i.e., the use of expressions and constructions that avoid providing information on the referent's gender – e.g., *il corpo docente* (the teaching body) or *coloro che insegnano* (those who teach) instead of *i professori* (the.M professors:M);
- *innovative obscuration*, i.e., the use of novel, gender-neutral markers instead of the gendered ones – e.g., *la professorə* (the.INN professor:INN) instead of *il professore* (the.M professor:M) or *la professoressa* (the.F professor:F).[2]

As we further discuss in Section 3, the released version of GFL-it for this challenge and GeNTE include references and annotations designed for the former strategy, whereas Neo-GATE for the latter.

Note that the chosen strategies do not exhaust the full range of possibilities: we discarded, for the moment, *visibility* strategies such as the repetition of an expression in the feminine and the masculine - e.g., *i professori e le professoresse* (the.M professors:M and the.F professors:F) - and the repetition of in three gendered forms (feminine, masculine and innovative) – e.g., *i professori, la professorə e le professoresse* (the.M professors:M, the.INN professors:INN and the.F professors:F).

3) Fair translation: like the second task, the third one is designed to test the models' ability to generate gender-fair language texts, but in the *cross-lingual* context of *automatic translation* from English into Italian. For example, consider applying the two gender-fair language strategies described above to the translation of the sentence "I am glad to know such knowledgeable doctors":

- conservative obscuration: *Sono felice di conoscere un personale medico così preparato.* [medical staff]

---

[1] In this report, we refer to innovative gender-fair strategies such as the schwa as "neomorphemes". Although aware that this terminology is controversial, we adopted it for simplicity and do not intend our terminology to imply any substantive stance.

[2] We indicate the innovative forms with "INN" in the glosses.

Task	GFL-it	GeNTE	Neo-GATE	Task total
Detection	2,187	-	841	3,028
Reformulation	1,206	750	841	2,797
Translation	-	1,500	841	2,341

Table 1
Number of dataset entries used for each task.

- innovative obscuration: *Sono content₂ di conoscere medic₂ così preparat₂.*

3. Data description

For our challenge, we propose three benchmarks dedicated to the evaluation of gender-fair language generation, (GFL-it[3], GeNTE [14],[4] and Neo-GATE [15]),[5] and a total of 7 prompts to be used across the tasks and datasets. We describe the datasets in subsections 3.1, 3.2, and 3.3 respectively, and the prompts in subsection 3.4.

Statistics about the benchmarks and their use within this challenge proposal are available in Table 1. GFL-it contains a total of 2,187 texts, among which 5 expert annotators identified an average of 3.24 unfair spans (in total 3,908) in 1,206 texts. For each identified span, the annotators proposed various gender-fair alternatives, with an average of 3.8 alternatives per span. For more detailed statistics about GeNTE and Neo-GATE we refer to the respective papers.

3.1. GFL-it

GFL-it was built on documents and texts from University website pages provided by the University of Brescia. It constitutes an expansion of the corpus presented in Rosola et al. [1]. The corpus comprises a total of 35 documents in Italian, split into 2,187 texts. Each text was annotated by 5 paid expert annotators following the original annotation scheme [1]. First, the annotators identified all the spans that contained any gender-unfairness, distinguishing among: OVEREXTENDED (3,465), GENERIC (530) and INCONGRUOUS GENDER (31) (see 2). Overall, 3,908 spans were identified. Then, they provided at least one alternative per span. The alternatives could belong to any of the gender-fair strategies: conservative or innovative obscuration, conservative or innovative visibility, or hybrid alternatives (i.e., any combination of these types).

Given that GFL-it is annotated for spans, each text contains a list of different spans and their reformulations in different forms of gender-fair language[6]. More specifi-

cally, each entry is described by the following attributes:

- *id_text*: The unique ID for each text.
- *text*: The entire text of the entry.
- *list_spans*: The list containing all spans found in the text.
- *rewritten_texts_generico*: A reformulation of the entire text where spans labeled as GENERIC are replaced.
- *rewritten_texts_sovraesteso*: A reformulation of the entire text where spans labeled as OVEREXTENDED are replaced.
- *rewritten_texts_generico_e_sovraesteso*: A reformulation of the entire text where spans are randomly replaced by available options in *rewritten_texts_generico* or *rewritten_texts_sovraesteso*.

Each span in *list_spans* follows the structure:

- *span*: The textual representation of the span.
- *start*: The starting index of the span in the text.
- *end*: The ending index of the span in the text.
- *labels*: A list of the types of gendered language used in the selected spans; possible values are OVEREXTENDED, GENERIC and INCONGRUOUS GENDER.
- *key_span*: The concatenation of *span*, *start* and *end* attributes; it can be used as an ID for each span contained inside a text.

We propose to use the GFL-it corpus for tasks 1 and 2,[7] namely, those regarding **gendered language detection** and **fair reformulation**.

3.2. GeNTE

GeNTE is a parallel English → Italian test set [16]. Originally designed to evaluate MT models' ability to perform gender-neutral translations, GeNTE was built upon a subset of the Europarl corpus [17], which is representative of natural, formal communicative situations from the institutional domain, the context where gender-neutral language is most accepted and encouraged [16, 14]. Overall, it consists of 1,500 <*English source, gendered Italian reference, gender-neutral Italian reference*> triplets aligned at the sentence level, which always contain at least one mention of human referents. The gendered Italian reference (REF-G) comes from the original Europarl corpus, whereas the gender-neutral reference (REF-N) was produced by professional translators who edited gendered forms into gender-neutral alternatives.

---

[3]https://github.com/simonasnow/GFL-it-Dataset
[4]https://huggingface.co/datasets/FBK-MT/GeNTE
[5]https://huggingface.co/datasets/FBK-MT/Neo-GATE
[6]For the purpose of the task 2, only the conservative obscured reformulations have been released in this version of the dataset.

[7]For task 2, we used a classifier that distinguishes between gendered and gender-neutral texts (see Section 4). Hence, we only used the GFL-it texts where the annotators identified gendered expressions (= gendered class) and the texts for which annotators provided at least one conservative obscured reformulation (= gender-neutral class) for a total amount of 1,206 texts.

Text		Per **gli iscritti** agli anni successivi al primo tali valutazioni scendono rispettivamente a NUM , NUM (sotto la soglia critica) e NUM (vicino alla soglia critica).
Span		gli iscritti
Reformulated Text		Per **le persone iscritte** agli anni successivi al primo tali valutazioni scendono rispettivamente a NUM , NUM (sotto la soglia critica) e NUM (vicino alla soglia critica).
		[For those enrolled in years after the first, these ratings drop to NUM, NUM (below the critical threshold) and NUM (close to the critical threshold), respectively.]

Table 2
Example from the GFL-it dataset. Words in bold correspond to the identified unfair spans in the text, and the reformulated expressions in the reformulated text. A translation of the text is provided in square brackets.

Set-G	**SRC**	When you assumed office, <u>Mr</u> Schreyer, you assured us that you would strive to achieve this.
	REF-G	Al momento della sua nomina, **signor** [Mr] Schreyer, ci aveva promesso che si sarebbe **adoperato** [(would have) strived] in tal senso.
	REF-N	Al momento della sua nomina, Schreyer, ci aveva promesso *un impegno* [a commitment] in tal senso.
Set-N	**SRC**	To some extent, those of us who are politicians find ourselves in the middle.
	REF-G	In certa misura **quelli** [those (of us)] di noi che sono **politici** [politicians] si trovano in una posizione intermedia.
	REF-N	In certa misura *chi* di noi [who, among us,] *svolge attività politica* [carries out political activities] si trova in una posizione intermedia.

Table 3
Examples of Set-G and Set-N entries in GeNTE. <u>Underlined</u> words are linguistic cues informing about human referents' gender; words in **bold** are gendered mentions of human referents; words in *italic* are the gender-neutral reformulations of the gendered mentions. Glosses of relevant expressions are provided in square brackets.

As shown in Table 3, GeNTE represents two types of phenomena, which are equally represented within the corpus. Namely, *i)* Set-N, featuring 750 gender-ambiguous source sentences that require to be rendered gender-neutrally; and *ii)* Set-G featuring gender-unambiguous source sentences, to be properly rendered with gendered (masculine or feminine) forms. Crucially, these two sets are a key feature of GeNTE, as they allow benchmarking whether systems are able to perform gender-neutral translations, but only when desirable. As a matter of fact, when referents' gender is unknown or irrelevant, undue gender inferences should not be made and gender-neutral language (i.e., conservative obscuration strategy) should be used. However, gender-neutralization should not be always enforced, and when a referent's gender is known or relevant, models should not over-generalize to gender-neutral generations.

Each entry in GeNTE is organized into the following fields:

- *ID*: The unique GeNTE ID.
- *Europarl_ID*: The original sentence ID from Europarl's common-test-set 2.
- *SET*: Indicates whether the entry belongs to the Set-G or the Set-N subportion of the corpus.
- *SRC*: The English source sentence.
- *REF-G*: The gendered Italian reference translation.
- *REF-N*: The gender-neutral Italian reference, produced by a professional translator.
- *GENDER*: For entries belonging to the Set-G, it indicates if the entry is Feminine or Masculine.

We propose the use of the whole GeNTE for the **translation task 3**, testing models' ability to produce gender-neutral translations only when appropriate. For

SOURCE	After the accident, they took me to the hospital and I stayed there for a whole month.
REF-M	Dopo l'incidente, mi hanno portato all'ospedale e sono rimasto lì per un mese intero.
REF-F	Dopo l'incidente, mi hanno portata all'ospedale e sono rimasta lì per un mese intero.
REF-TAGGED	Dopo l'incidente, mi hanno <u>portatə</u> all'ospedale e sono <u>rimastə</u> lì per un mese intero.
ANNOTATION	portato portata <u>portatə</u>; rimasto rimasta <u>rimastə</u>;

Table 4
Example of a Neo-GATE entry, already adapted to the `schwa-simple` neomorpheme paradigm. <u>Underlined</u> words include the neomorpheme schwa (ə).

the **fair reformulation task 2**, we only repurpose part of the Italian portion of the corpus, i.e., REF-G references from SET-N.

3.3. Neo-GATE

Similarly to GeNTE, Neo-GATE is a parallel corpus designed for gender-fair English → Italian MT evaluation. Here, however, the focus is on the use of gender-fair neomorphemes (i.e., innovative obscuration strategy) rather than conservative gender-neutral language. Neo-GATE was built on GATE [18], a test set manually created specifically to evaluate gender reformulation and gender bias in MT. In GATE, the gender of human entities is unknown, i.e., there are no linguistic elements providing gender information about human referents in the (English) source sentences.

Neo-GATE includes an annotation that defines the words upon which the evaluation is based. It includes the three forms required for the evaluation, i.e., the masculine and feminine forms, and forms featuring placeholders in place of Italian overt gender markers. Before the evaluation, the placeholders must be replaced with the correct forms in the desired neomorpheme paradigm. For this task, Neo-GATE was adapted to a version of the 'schwa' paradigm [19, 20], to which we refer as schwa-simple here, i.e., the placeholders were replaced with the forms described in Appendix A.

Like GeNTE, Neo-GATE includes Italian references that differ exclusively in gender expression. Besides the English source sentence, all entries in Neo-GATE have three Italian references: REF-M, where the gender of words referring to human beings is masculine, REF-F, where human beings are referred to as feminine, and REF-TAGGED, where placeholders replace overt markers of gender – here adapted to the schwa-simple paradigm. However, differently from GeNTE, the English sentences in Neo-GATE never include gender cues. An example of a Neo-GATE entry is available in Table 4.

Each entry in Neo-GATE includes the following fields:

- *#*: The entry identifier within Neo-GATE.
- *GATE-ID*: A unique identifier of the original GATE entry, composed of a prefix indicating the subset of origin followed by a serial number.
- *SOURCE*: The English source sentence.
- *REF-M*: The Italian reference where all gender-marked terms are masculine.
- *REF-F*: The Italian reference where all gender-marked terms are feminine.
- *REF-TAGGED*: The Italian reference where all gender-marked terms are tagged with Neo-GATE's annotation.
- *ANNOTATION*: The word level annotation.

We propose to use all Neo-GATE entries for **all three tasks** of our challenge. While for tasks 1 (**gendered language detection**) and 2 (**fair reformulation**) we only use Italian references – namely both REF-M and REF-F for task 1, and REF-M only for task 2 – as input for the models, for task 3 (**fair translation**) we use the English SOURCE sentences.

3.4. Example of used prompts

This section describes the prompts we propose for our challenge, with examples available in Table 5.

In prompts **A** and **B**, we ask the model to identify the gendered expressions (introduced by the tag *[Espressione]:*) in the text given as input; if no gendered expression is detected in the text (initialized with the tag *[Genere marcato]:*) the model should output 0. The model can recognize more than one gendered expression.

In prompts **C**, **D**, and **E**, the shots include one line starting with the tag *[Genere marcato]:*, indicating that the following sentence is gendered. Then, in prompts **C** and **D** the following line starts with *[Neutro]:* followed by a gender-neutral reformulation, whereas in **E** it starts with *[Neomorfema]:* and includes the innovative obscuration alternative of the first sentence, with neomorphemes in place of the masculine forms.[8]

Prompts **F** and **G** start with the tag *[Inglese]:* followed by the English source sentence to be translated. In prompt **F**, the second line either starts with the tag *[Italiano, genere marcato]:* (see F - Exemplar format 1 in Table 5) if it is followed by a gendered translation or with the tag *[Italiano, neutro]:* if the subsequent translation is gender-neutral (see F - Exemplar format 2). Models are required to produce the correct tag and translation depending on the presence or absence of gender cues in the source. Finally, prompt **G** includes two different translations after the source sentence: the first, preceded by the tag *[Italiano, genere marcato]:*, includes a translation featuring masculine forms in reference to human beings, whereas the second translation starts with the tag *[Italiano, neomorfema]:* and uses neomorphemes in reference to human beings. Models are required to produce both translations, though only the second will be extracted in post-processing and used for the evaluation.

In particular, prompts D, E, F, and G are based on the ones used in previous experiments on the same datasets [12, 15], and were in turn inspired by the format proposed by Sánchez et al. [21].

---

[8]We here used *neutro* (neutral/neuter), despite being aware of its ambiguity with *neuter*, a grammatical gender not present in the Italian linguistic system. However, nothing substantive hinges on this terminological choice.

		Task 1: Gendered language detection
A	Instruction	Identifica le espressioni che contengono dei marcatori di genere femminile o maschile. Se più di un'espressione è identificata, separale con " ;". Se nessuna espressione è identificata, restituisci 0.
	Exemplar format	[Genere marcato]: A livello di Ateneo ha partecipato il 76% degli iscritti del 1° anno del CdL e l'81% di tutti gli altri studenti. [Espressione]: degli iscritti ; tutti gli altri studenti
B	Instruction	Identifica le espressioni che contengono dei marcatori di genere femminile o maschile. Se più di un'espressione è identificata, separale con " ;". Se nessuna espressione è identificata, restituisci 0.
	Exemplar format	[Genere marcato]: Quest'anno mi sono ammalata già due volte. [Espressione]: ammalata
		Task 2: Fair reformulation
C	Instruction	Riformula la seguente frase utilizzando un linguaggio neutro rispetto al genere dei referenti umani, evitando l'uso di forme maschili e femminili.
	Exemplar format	[Genere marcato]: - Il 79% dei laureati lavora (a un anno dalla laurea). [Neutro]: - Il 79% delle persone laureate lavora (a un anno dalla laurea).
D	Instruction	Riformula la seguente frase utilizzando un linguaggio neutro rispetto al genere dei referenti umani, evitando l'uso di forme maschili e femminili.
	Exemplar format	[Genere marcato]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità dei parlamentari europei? [Neutro]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità dei membri del Parlamento Europeo?
E	Instruction	Riformula la seguente frase usando il neomorfema "ə". Il neomorfema "ə" deve essere usato come sostituto dei morfemi maschili e femminili nelle parole che si riferiscono agli esseri umani.
	Exemplar format	[Genere marcato]: Il partner di un mio amico ci ha invitati a cena. [Neomorfema]: Lə partner di unə miə amicə ci ha invitatə a cena.
		Task 3: Fair translation
F	Instruction	Traduci la seguente frase inglese in italiano seguendo queste regole: 1. Se la frase inglese indica chiaramente il genere dei referenti umani (maschile o femminile), traduci usando il genere corretto. 2. Se la frase inglese non indica il genere dei referenti umani, traduci usando un linguaggio neutro che non esprime genere, evitando forme maschili e femminili.
	Exemplar format 1	[Inglese]: However, it is important that the Commissioner has declared his loyalty to the President himself. [Italiano, genere marcato]: Tuttavia, è importante che il Commissario abbia dichiarato la sua fedeltà al Presidente stesso.
	Exemplar format 2	[Inglese]: Secondly, how far does it increase transparency and accountability of the MEPs? [Italiano, neutro]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità dei membri del Parlamento Europeo?
G	Instruction	Traduci la seguente frase inglese in italiano usando il neomorfema "ə". Il neomorfema "ə" deve essere usato come sostituto dei morfemi maschili e femminili nelle parole che si riferiscono agli esseri umani.
	Exemplar format	[Inglese]: The partner of a friend of mine invited us to dinner. [Italiano, genere marcato]: Il partner di un mio amico ci ha invitati a cena. [Italiano, neomorfema]: Lə partner di unə miə amicə ci ha invitatə a cena.

Table 5
Examples of the format of all prompts we propose for our challenge. Dataset-wise, prompts A and C are designed to be used with GFL-it data, prompts B, E, and G are designed for Neo-GATE, and prompts D and F are designed for GeNTE.

4. Metrics

For the evaluation of gendered language detection (i.e., with GFL-it and Neo-GATE in task 1) we used the F1-score obtained using BERTScore[9] [22] for each entry in the datasets. In particular, for each entry, we extract the most relevant correspondence between the gendered expressions identified by the annotators and the ones

---

[9]https://huggingface.co/spaces/evaluate-metric/bertscore

produced by the generative model, computing the maximum F1-score. Once the correspondences are set for each entry, we average the scores.

For the evaluation of gender-neutral reformulation—i.e., with GFL-it and GeNTE in task 2—and translation—i.e., with GeNTE and Neo-GATE in task 3—we propose an accuracy score based on the labels produced by the classifier introduced in Piergentili et al. [14]. More specifically, we use version 2 of the classifier, introduced in Savoldi et al. [12]. This classifier assigns a label to each model output, either gender-neutral or gendered. We then compare those labels against the true labels, i.e., always gender-neutral in the reformulation task and either gendered or gender-neutral for the translation task, depending on whether the entry belongs to Set-G or Set-N respectively. The final score is computed as the corpus-level percentage of correct labels.

For neomorpheme-based gender-fair reformulation (task 2) and translation (task 3) based on Neo-GATE, we propose the coverage-weighted accuracy described in Piergentili et al. [15] as the main metric. This metric takes into account both how accurately a model generates neomorphemes and the proportion of annotations (i.e., either of the masculine, feminine, or innovative forms) found during the evaluation, thus allowing for fair system comparisons and rankings. As complementary metric to assess models' ability to correctly generate neomorphemes, we propose reporting the mis-generation score [15] as well. This metric can flag undesired behaviors even despite good accuracy, as it counts cases where models generate neomorphemes inappropriately, for instance by applying the use of neomorphemes to words that do not refer to human entities (e.g., by generating 'tavolə' instead of 'tavolo', en: table).

5. Limitations

Our work presents some limitations. Firstly, the datasets employed only derive from specific domains: GFL-it exclusively contains data from administrative documents and official web pages of the University, GeNTE from documents of the European Parliament, and Neo-GATE data manually created by experts. The corpora could be expanded to other domains and annotated by more annotators in future research. Secondly, our metrics are only a first attempt and others should be explored in the future. Moreover, we only tested one paradigm of neomorphemes, namely the `schwa-simple`, while many others exist (e.g., the asterisk, the '-u', the '@' - see [23] for a complete list), and even more could be proposed. Furthermore, GeNTE and Neo-GATE do not contain mixed texts where rewriting is needed with respect to one entity but not others.

6. Ethical issues

The proposed tasks in this challenge have the purpose of reducing the use of gender-unfair expressions in heavily gender-marked languages (i.e., Italian) that affect the visibility of other genders (in particular, feminine and non-binary). Although the datasets have been built by experts of gender-fair language, the group of annotators of GFL-it was not gender-balanced as only 2 out of 5 annotators were men.

Moreover, we are aware of the fact that the use of neomorphemes like the schwa ə makes reading harder for people with dyslexia or visual impairments [4, 24, 25]. This issue, however, is mitigated thanks to the possibility of selecting the most suitable neomorpheme according to each user's needs. In particular, both people with dyslexia or visual impairments can rely on screen readers, which differ in their ability to correctly interpret specific neomorphemes: the possibility to select different neomorphemes allows each user to select the one(s) their screenreader interpret best.

7. Data license and copyright issues

8. Acknowledgments

References

[1] M. Rosola, S. Frenda, A. T. Cignarella, M. Pellegrini, A. Marra, M. Floris, et al., Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in italian, in: CLiC-it 2023.

Proceedings of the 9th Italian Conference on Computational Linguistics. Venice, Italy, November 30-December 2, 2023., volume 3596, CEUR-WS, 2023, pp. 1–10.

[2] J. Silveira, Generic Masculine Words and Thinking, Women's Studies International Quarterly 3 (1980) 165–178. URL: https://www.sciencedirect.com/science/article/pii/S0148068580921132.

[3] P. Gygax, S. Sato, A. Öttl, U. Gabriel, The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system, Language Sciences 83 (2021) 101328.

[4] G. Sulis, V. Gheno, The debate on language and gender in italy, from the visibility of women to inclusive language (1980s–2020s), The Italianist 42 (2022) 153–183. doi:10.1080/02614340.2022.2125707.

[5] G. Visibility, N. across Languages, Beyond pronouns, The Oxford Handbook of Applied Philosophy of Language (2024) 320.

[6] M. Rosola, Linguistic hermeneutical injustice, Social Epistemology (2024). doi:10.1080/02691728.2024.2401143.

[7] S. J. Kapusta, Misgendering and its moral contestability, Hypatia 31 (2016) 502–519.

[8] R. Dembroff, D. Wodak, He/she/they/ze, Ergo (2018).

[9] S. Sczesny, M. Formanowicz, F. Moser, Can gender-fair language reduce gender stereotyping and discrimination?, Frontiers in psychology 7 (2016) 154379.

[10] P. Gygax, S. Zufferey, U. Gabriel, Le cerveau pense-t-il au masculin, Cerveau, langage et représentations sexistes, Paris, Le Robert (2021).

[11] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485. doi:10.18653/v1/2020.acl-main.485.

[12] B. Savoldi, A. Piergentili, D. Fucci, M. Negri, L. Bentivogli, A prompt response to the demand for automatic gender-neutral translation, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 256–267. URL: https://aclanthology.org/2024.eacl-short.23.

[13] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[14] A. Piergentili, B. Savoldi, D. Fucci, M. Negri, L. Bentivogli, Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 14124–14140. URL: https://aclanthology.org/2023.emnlp-main.873. doi:10.18653/v1/2023.emnlp-main.873.

[15] A. Piergentili, B. Savoldi, M. Negri, L. Bentivogli, Enhancing gender-inclusive machine translation with neomorphemes and large language models, in: C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia, H. Moniz (Eds.), Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), European Association for Machine Translation (EAMT), Sheffield, UK, 2024, pp. 300–314. URL: https://aclanthology.org/2024.eamt-1.25.

[16] A. Piergentili, D. Fucci, B. Savoldi, L. Bentivogli, M. Negri, Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges, in: E. Vanmassenhove, B. Savoldi, L. Bentivogli, J. Daems, J. Hackenbuchner (Eds.), Proceedings of the First Workshop on Gender-Inclusive Translation Technologies, European Association for Machine Translation, Tampere, Finland, 2023, pp. 71–83. URL: https://aclanthology.org/2023.gitt-1.7.

[17] P. Koehn, Europarl: A Parallel Corpus for Statistical Machine Translation, in: Proceedings of the tenth Machine Translation Summit, AAMT, Phuket, TH, 2005, pp. 79–86. URL: http://mt-archive.info/MTS-2005-Koehn.pdf.

[18] S. Rarrick, R. Naik, V. Mathur, S. Poudel, V. Chowdhary, GATE: A challenge set for gender-ambiguous translation examples, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 845–854. URL: https://doi.org/10.1145/3600211.3604675. doi:10.1145/3600211.3604675.

[19] A. M. Thornton, Genere e igiene verbale: l'uso di forme con ə in italiano, Annali Del Dipartimento Di Studi Letterari, Linguis-

tici E Comparati. Sezione Linguistica 11 (2020) 11–54. URL: http://www.serena.unina.it/index.php/aionlin/article/view/9623. doi:https://doi.org/10.6093/2281-6585/9623.

[20] R. Baiocco, F. Rosati, J. Pistella, Italian proposal for non-binary and inclusive language: The schwa as a non-gender–specific ending, Journal of Gay & Lesbian Mental Health 27 (2023) 248–253. URL: https://doi.org/10.1080/19359705.2023.2183537. doi:10.1080/19359705.2023.2183537.

[21] E. Sánchez, P. Andrews, P. Stenetorp, M. Artetxe, M. R. Costa-jussà, Gender-specific machine translation with large language models, 2024. URL: https://arxiv.org/abs/2309.03175. arXiv:2309.03175.

[22] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[23] V. Gheno, Lo schwa tra fantasia e norma, La falla (2020). URL: https://lafalla.cassero.it/lo-schwa-tra-fantasia-e-norma/.

[24] L. Iacopini, Lo schwa (ə) che rende l'inclusione inaccessibile, Web accessibile (2021). URL: https://webaccessibile.org/approfondimenti/lo-schwa-%C7%9D-che-rende-linclusione-inaccessibile/.

[25] C. D. Santis, L'emancipazione grammaticale non passa per una e rovesciata, 2022. URL: https://www.treccani.it/magazine/lingua_italiana/articoli/scritto_e_parlato/Schwa.html.

A. The `schwa-simple` paradigm

Table 6 reports the forms used in the `schwa-simple` paradigm, along with the corresponding tags in Neo-GATE and masculine and feminine equivalents.

TAG	Description	Masculine	Feminine	Schwa
<ENDS>	portion of the word differentiating gendered forms, singular	o, e, tore	a, essa, trice	ə, torə
<ENDP>	portion of the word differentiating gendered forms, plural	i, tori	e, esse, trici	ə, torə
<DARTS>	definite article, singular	il, lo, l'	la, l'	lə
<DARTP>	definite article, plural	i, gli	le	lə
<IART>	indefinite article	uno, un	una, un'	unə
<PARTP>	partitive article, plural	dei, degli	delle	deə
<PREPdiS>	articulated preposition with root 'di', singular	del, dello, dell'	della, dell'	dellə
<PREPdiP>	articulated preposition with root 'di', plural	dei, degli	delle	dellə
<PREPaS>	articulated preposition with root 'a', singular	al, allo, all'	alla, all'	allə
<PREPaP>	articulated preposition with root 'a', plural	agli, ai	alle	allə
<PREPdaS>	articulated preposition with root 'da', singular	dal, dallo, dall'	dalla, dall'	dallə
<PREPdaP>	articulated preposition with root 'da', plural	dagli	dalle	dallə
<PREPinP>	articulated preposition with root 'in', plural	negli	nelle	nellə
<PREPsuS>	articulated preposition with root 'su', singular	sul, sullo, sull'	sulla, sull'	sullə
<PREPsuP>	articulated preposition with root 'su', plural	sugli	sulle	sullə
<DADJquelS>	demonstrative adjective (far), singular	quel, quello, quell'	quella, quell'	quellə
<DADJquelP>	demonstrative adjective (far), plural	quegli	quelle	quellə
<DADJquestS>	demonstrative adjective (near), singular	questo, quest'	questa, quest'	questə
<DADJquestP>	demonstrative adjective (near), plural	questi	queste	questə
<POSS1S>	possessive adjective, 1st person singular, singular	mio	mia	miə
<POSS1P>	possessive adjective, 1st person singular, plural	miei	mie	miə
<POSS2S>	possessive adjective, 2nd person singular, singular	tuo	tua	tuə
<POSS2P>	possessive adjective, 2nd person singular, plural	tuoi	tue	tuə
<POSS3S>	possessive adjective, 3rd person singular, singular	suo	sua	suə
<POSS3P>	possessive adjective, 3rd person singular, plural	suoi	sue	suə
<POSS4S>	possessive adjective, 1st person plural, singular	nostro	nostra	nostrə
<POSS4P>	possessive adjective, 1st person plural, plural	nostri	nostre	nostrə
<PRONDOBJS>	direct object pronoun, singular	lo	la	lə
<PRONDOBJP>	direct object pronoun, plural	li	le	lə

Table 6
The full tagset used in Neo-GATE, mapped to the Italian gendered forms and the `schwa-simple` nomorpheme paradigm.