ConGA: Guidelines for Contextual Gender Annotation. A Framework for Annotating Gender in Machine Translation

Argentina Anna Rescigno[1,2], Eva Vanmassenhove[3], Johanna Monti[2]
[1]University of Pisa, [2]University of Naples "L'Orientale", [3]Tilburg University
argentina.rescigno@phd.unipi.it, e.o.j.vanmassenhove@tilburguniversity.edu, jmonti@unior.it

Abstract

Handling gender across languages remains a persistent challenge for Machine Translation (MT) and Large Language Models (LLMs), especially when translating from gender-neutral languages into morphologically gendered ones, such as English to Italian. English largely omits grammatical gender, while Italian requires explicit agreement across multiple grammatical categories. This asymmetry often leads MT systems to default to masculine forms, reinforcing bias and reducing translation accuracy. To address this issue, we present the *Contextual Gender Annotation* (ConGA) framework, a linguistically grounded set of guidelines for word-level gender annotation. The scheme distinguishes between semantic gender in English through three tags, Masculine (M), Feminine (F), and Ambiguous (A), and grammatical gender realisation in Italian (Masculine (M), Feminine (F)), combined with entity-level identifiers for cross-sentence tracking. We apply ConGA to the gENder-IT dataset, creating a gold-standard resource for evaluating gender bias in translation. Our results reveal systematic masculine overuse and inconsistent feminine realisation, highlighting persistent limitations of current MT systems. By combining fine-grained linguistic annotation with quantitative evaluation, this work offers both a methodology and a benchmark for building more gender-aware and multilingual NLP systems.

Keywords: annotation, gender bias, machine translation

1. Introduction

Detecting and mitigating gender bias in Machine Translation (MT) systems, and particularly in Large Language Models (LLMs), remains a central challenge in computational linguistics and research. Bias often arises from the structural tension between gender-neutral and gender-marked languages. While English can leave the gender unspecified, Italian requires explicit gender agreement across grammatical elements, compelling the system to select between masculine or feminine forms, even when the source text is neutral. For instance, the English sentence "*I am a nurse*" may be translated into Italian as either "*Io sono un'infermiera* <F>" or "*Io sono un infermiere* <M>". Similarly, "*I am happy*" becomes either "*Io sono contenta* <F>" or "*Io sono contento* <M>".

Previous research has shown that MT systems tend to overuse masculine forms due to unbalanced data, where male references are disproportionately represented (Cho et al., 2019; Prates et al., 2020). When the source language lacks gender cues, MT systems may resort to the most statistically likely or stereotypical option, often defaulting to masculine forms or reinforcing gendered associations (e.g., *nurse* as female, *engineer* as male) because of social biases (Monti, 2017). These biases are further exacerbated in cross-linguistic contexts, where the system must infer gender information absent from the source (Rescigno et al., 2020; Rescigno and Monti, 2023; Vanmassenhove, 2024).

Despite advances in neural architectures and context-aware translation, such biases remain unresolved (Savoldi et al., 2025). Evaluating them also poses further significant challenges: quantitative methods rely on metrics such as precision, recall, or F1-score, while qualitative analysis explores context-rich examples, yet no unified framework exists to ensure comparability across languages or studies. Moreover, most existing research remains English-centric, neglecting low-resource and morphologically rich languages, as well as non-binary and gender-diverse identities.

The present study aims to address these gaps by introducing a linguistically informed framework for the annotation and evaluation of gender expression in translation. Specifically, it investigates how linguistic annotations can be leveraged to quantify bias in MT outputs. By combining linguistic insight with computational evaluation, this approach aims to support fairer and more context-aware MT systems.

2. Related Work

Research on gender bias in MT highlights persistent asymmetries between gender-neutral and gender-marked languages. Early statistical MT systems were trained on large, often unbalanced, corpora, which have been shown to contribute to gender bias in their outputs (Stanovsky et al., 2019). More recent neural MT systems have, in some cases, been shown to amplify gender-stereotypical

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4048–4057
11–16 May 2026.

translations by implicitly learning from such biased training data (Vanmassenhove et al., 2018; Garg et al., 2018; Savoldi et al., 2021). Studies have also shown that when translating from English into morphologically richer languages, systems tend to over-generate masculine forms, even when the source is ambiguous (Vanmassenhove et al., 2019; Rescigno et al., 2020; Rescigno and Monti, 2023).

To mitigate these issues, several strategies have been explored. Vanmassenhove et al. (2018) and Elaraby et al. (2018) introduced gender tagging and speaker-aware MT to improve agreement, while Escudé Font and Costa-jussà (2019) used debiased word embeddings. Furthermore, Stanovsky et al. (2019) proposed the *Target Gender Annotation*, marking tokens in the source with the grammatical gender of their aligned target translation to guide neural models. This method significantly improved accuracy on the WinoMT performance without lowering BLEU scores (Papineni et al., 2002). Despite these advances, most strategies focus on the mitigation of bias at the model level, rather than the evaluation and annotation side of the problem.

2.1. Evaluation and Annotation Frameworks

Existing resources for the evaluation of gender bias in MT fall into two categories: benchmark datasets for evaluation and annotation schemes. Benchmarks, such as WinoGender (Rudinger et al., 2018) and WinoMT (Stanovsky et al., 2019) provide templated sentences to test coreference and expose occupational stereotypes, but they lack linguistic diversity. Datasets like MuST-SHE (Bentivogli et al., 2020) and gENder-IT (Vanmassenhove and Monti, 2021) extend the evaluation to speech and text translation, by providing contextual gender annotation. In MuST-SHE, gender information is defined at the segment level and derived either from the speaker's gender or from contextual lexical cues, allowing analysis of gender agreement and preservation in longer sentences. gENder-IT adapts this framework by offering explicit gender tags for systematic evaluation of gender realisation in MT.

More recent approaches emphasise contextual, linguistically grounded annotation. Instead of just counting the *masculine vs. feminine* outputs, these approaches aim to describe how and where gender information is encoded in a text. Hitti et al. (2019) proposed a systematic taxonomy for gender bias, giving a definition of the phenomenon of *gender generalisation*, among others. Building on this, Havens et al. (2022) expanded the classification to eleven categories, though mainly for English monolingual data.

These taxonomies emphasise inclusivity and context but are largely limited to monolingual English data and to detecting bias rather than measuring it quantitatively across languages.

While evaluation corpora for MT mostly focus on word- or sentence-level tagging, recent research dealing with large-scale models alignment introduces datasets for bias mitigation. GenderAlign (Zhang et al., 2024) presents a large dataset of dialogues annotated with unbiased or biased responses to fine-tune LLMs via Direct Preference Optimisation (DPO). It covers four different categories (such as stereotypes or discriminatory language) and shows that alignment on balanced distributions significantly reduces bias across the considered benchmarks.

These alignment approaches differ from linguistic annotation schemes: rather than describing gender realisation in text, they train models to prefer non-biased outputs. Nevertheless, both fields converge on the need for transparent, well-annotated data and clear taxonomies to support reproducible bias evaluation.

3. Theoretical Background

Languages encode gender in different ways, using different linguistic elements. An European Parliament (2018) report identifies three main types:

1. genderless languages (e.g., Finno-Ugric ones), which express the gender lexically (Finnish *poika* "boy", *tyttö* "girl") or by suffixes (in Hungarian, *doctor* is the gender neutral form, and *doktornő* is a "female doctor", with the *-nő* suffix that means "woman"), and nouns and pronouns have no gender;

2. natural gender languages (e.g., English, Swedish, Danish), which have mainly genderless nouns, but the pronouns are specific for each gender and reflect the referents' biological sex;

3. grammatical gender languages, which have gendered nouns/pronouns and an inflectional pattern for them (e.g., Slavic and Romance languages).

The physical sex of animate entities, which is realised by the male/female polarity, varies across languages as it depends on the classification of reality, and is called ***natural gender***. However, some may argue that the concept of "natural gender" is a myth. Instead, McConnell-Ginet (2013) argues that English, and even many grammatical gender systems, reflect ***notional gender*** (Nevalainen and Raumolin-Brunberg, 1993), culturally constructed ideas about sex and gender, rather than purely biological sex (McConnell-Ginet, 2013).

Grammatical gender is a morphological characteristic of languages, and can be masculine, feminine, or neuter. Usually, for inanimate entities, it is assigned by linguistic criteria; for animate entities, it will be coherent with the natural gender of the entity. Consequently, evaluating gender bias, which is inherently tied to the social and biological dimensions of human referents, is only meaningful for animate entities. Since inanimate concepts lack a natural gender, their grammatical gender assignment in the target language is a strictly morphological process. Therefore, it is impossible to establish a correspondence between grammatical and natural gender for inanimate objects, making them inapplicable for the evaluation of social or representational bias. Corbett (1991) also distinguishes the type of grammatical gender among languages, e.g., Italian (or German) has grammatical gender for nouns, pronouns, articles, and adjectives, while languages like English, which only have expressed gender in pronouns, are pronominal gender languages.

3.1. Italian Grammatical Gender and Agreement

Italian, like many Romance languages, uses grammatical gender (masculine or feminine), which is often assumed to reflect biological sex, with the consequent misconception that one can freely choose which grammatical gender to use when referring to animate entities. Grammatical gender has precise attribution and agreement rules, just like for singular and plural.

Gender, but also number and person, is an inflectional category for Italian agreement. Inflection is a process of word formation in which words are modified to express different grammatical categories. Agreement occurs when a word changes form depending on the other words to which it relates. It is an instance of inflection, and usually involves making the value of some grammatical category (such as gender or person) "agree" between varied words or parts of the sentence. Agreement can be inherent, when the inherent traits of the word belong to the word *per se*, such as gender for the names. Inherent inflection doesn't require syntactic context (Booij, 1996), but the choice of the form is determined by what the speaker wants to express. For example, the noun *attrice* (en: *actress*) is inherently feminine as it describes an animate entity of female gender by definition. On the other hand, contextual traits depend on the context in which the word is; therefore, the agreement is contextual and its inflection is dictated by syntax (e.g., agreement markers for adjectives); for example, an adjective can have a masculine or feminine form contextually to the noun it refers to, for instance, it: *vecchia* <F> *insegnante* (en: *old female teacher*) and *vecchio* <M> *insegnante* (en: *old male teacher*).

4. Methodology

For this experiment, we used the English sentences of the gENder-IT dataset (Vanmassenhove and Monti, 2021), a bilingual English-Italian corpus derived from the MuST-SHE dataset (version 1.04) (Bentivogli et al., 2020). The dataset originally provides word-level annotations for all nouns and pronouns referring to human referents for the English sentences. The set of tags created for annotating the gENder-IT dataset were <F> or <M> when it was clear from the sentence context that the referent was female/male, and a <A> tag when, within the given context, no assumption can or should be made with respect to the gender of the referent. We also performed a light cleaning of the dataset for the removal of the pre-existing tags. This was necessary to ensure that the tags were consistent with the proposed annotation guidelines.

For the Italian data, we used the gENder-IT dataset as an input to generate the translations. To do so, we used two large-scale models, mBART and TowerInstruct-7B model (`thinkverse/towerinstruct`, from now on referred to as TowerLLM), available via the Ollama platform[1]. This model is part of the Tower family, which is designed for instruction-following tasks and optimised for high-quality text generation. The translations were produced using Ollama's local inference framework, allowing for on-device execution of LLMs. The TowerInstruct model was executed with a temperature setting of 0.2 to obtain more deterministic and consistent outputs, and also to ensure accuracy and semantic consistency in the translations, which are essential prerequisites for performing a reliable word-level alignment. For this specific study, the primary goal is to build a reliable word-level gold standard for gender annotations. To achieve this, we had to isolate the baseline, therefore we maintained a temperature of 0.2. Higher temperatures, while introducing lexical variety, also increase the risk of hallucinations or semantic divergence from the source text, complicating the entity-level alignment required by the framework. The annotation procedure was carried out using INCEpTION[2] (Klie et al., 2018), a tool that enables entity-level alignment between English and Italian sentences and is completely customisable to the experiment's needs.

The resulting parallel annotated corpus provides a fine-grained mapping between contextual gender cues in English and grammatical gender realisations in Italian, forming the gold-standard resource

[1] https://ollama.com/thinkverse/towerinstruct
[2] https://inception-project.github.io

to be used for subsequent quantitative evaluation of bias through precision, recall, and F1 scores.

5. Annotation Task

While previous works have addressed gender bias in MT through evaluation of the outputs against reference translations or challenge sets (Prates et al., 2020; Stanovsky et al., 2019; Vanmassenhove et al., 2018), this approach targets an earlier phase, by defining exactly what types of gender-related phenomena contribute to evaluate reliably gender bias in both MT systems and LLMs. The main aims of the task are:

- to create an efficient framework for linguistic gender annotation to better detect bias in LLMs. These annotations should clarify when and how gender is made explicit through agreement in the Italian language (e.g., adjectives, verbs, participles), and whether that realisation reflects a bias when derived from an otherwise ambiguous source;
- to annotate a dataset according to the framework described here, which will serve as a gold standard to assess biased orientation concerning the Italian language in an EN-IT machine translation task.

This task involves the annotation at the word level of the grammatical and contextual gender of words in a bilingual setting (English-Italian).

5.1. Task Explanation

Given the structural differences between English and Italian, the tagging approach was adapted to the specific role of each language in the translation process. In English (source), gender tagging is based on contextual meaning, as understanding gender cues in the source is essential for evaluating translation accuracy. In Italian (target), gender tagging relies on explicit grammatical realisations, such as agreement markers and gendered morphology, to identify how gender information is rendered when it is present or inferred. This distinction reflects a generalisable methodology for assessing gender in translation: capturing gender cues in the source to evaluate correctness, and analysing their realisations in the target to detect potential biases. See Table 1 for an example.

In this example, the entity's tag <F> for the English pronoun "we" results from the contextual agreement with the subject of the sentence, which is "*women*". For the Italian translation, the tags would be applied to the adjectives, as elements bearing the gender information; in this specific case, then the <F> tag would go only to "*reattive*", an explicit female form of the adjective, while "*sensibili*" and "*irragionevoli*" are left untagged, as these adjectives have grammatically the same form for both genders, and considering the grammatical gender of these adjectives feminine would mean making an assumption, because as far as we know, this could be an analogous case to the example in Table 2. Even though the subject is the same for both sentences (i.e., *you* <A1>), the two adjectives referring to it have two different genders in the Italian translation (i.e., the adjectives *asciutta* <F> / *bravo* <M>).

EN source	Women have been trained to think that **we** <F1> are overreacting or that **we** <F1> are being too sensitive or unreasonable.
IT target	Alle donne è stato insegnato a pensare che (noi) siamo troppo **reattive** <F1> o che (noi) siamo troppo sensibili o irragionevoli.

Table 1: Example of tagging between English and Italian sentences. The tag <F1> highlights corresponding focus elements across languages.

EN source	No, **you** <A1>'re still dry, **you** <A1>'re just being nice.
IT target	No, (tu) sei ancora **asciutta** <F1>, (tu) stai solo facendo il **bravo** <M1>.

Table 2: Example of tagging between English and Italian sentences. Tags such as <A1>, <F1>, and <M1> indicate corresponding referents or gender features across languages.

5.2. Tagging Strategy

For the task, a minimal and functional set of **gender tags** has been defined. The tags reflect the semantic and grammatical realisation of gender in translation. For feminine and masculine animate entities, the tags used are <F> and <M>, respectively. For those cases where the gender cannot be said or inferred by any other element of the sentence, or can refer to a male or female entity, the tag <A> is used. However, the tags are applied differently across English and Italian due to structural differences between the languages.

In English, tags are assigned based on contextual cues. If the referent's gender is explicit or strongly implied (e.g., via context or coreference), it is tagged as <F> or <M>. Whenever the gender is not inferable (e.g., first-person pronouns like *I*, *you*, or gender-neutral roles like *parent*), it is tagged as <A>.

In Italian, only <F> and <M> tags are used, and the annotation should reflect the grammatical gen-

der realisations (e.g., adjective or participle agreement). We chose not to use any <A> tag in Italian, as the presence or absence of gender realisation is precisely what allows us to compute bias. Gender-neutral or ambiguous references that remain grammatically unresolved are left untagged.

This simplified tag set ensures that all gender realisations in Italian can be interpreted with respect to ambiguous or specified gender in English, enabling the construction of a bias score that reflects unjustified gender disambiguation in translation.

To support fine-grained analysis and the development of a bias evaluation metric, we enrich each tag with an **entity-level identifier** that distinguishes referents within the same sentence. As such, tags are enriched with an additional numerical tag that indicates which entities connect to each other and how many different distinct referents there are. Their annotation as <A1>, <F1>, <M1>, and so on, indicates distinct animate referents.

This indexing helps track how each ambiguous source entity is resolved in Italian. Moreover, this enables computation of how many ambiguous entities are present, and how many are realised with a gender in translation, providing a quantifiable basis for bias detection. For instance, if an English sentence includes two ambiguous entities, they would be annotated as <A1> and <A2>, respectively. The same criteria would be applied in the Italian sentence, with the tags corresponding to the gendered realisations of indexed English referents. If no gender is grammatically expressed, then no tag is applied. This neutrality is significant when assessing whether a gendered realisation was necessary.

5.3. Specific Rules for Italian Annotation

Italian annotations are restricted to cases where grammatical gender is explicitly realised. Only elements that morphologically encode gender are tagged with feminine or masculine labels, while ambiguous or unmarked forms remain untagged to highlight instances where gender is unnecessarily introduced in translation. Tags are applied to agreement targets such as adjectives, past participles with the verb *essere* (to be), articles, possessives, and relative pronouns that display gender inflection.

Personal pronouns pose a challenge since Italian often omits the subject pronoun, as it is a pro-drop language; thus, tagging depends on agreement markers rather than the pronoun itself. Past participles agree in gender either with the subject (for intransitive verbs) or with the preceding object pronoun (for transitive verbs).

Proper names are generally annotated as <A> in the English sentences, since names alone can not be considered reliable gender cues, given the cross-cultural naming variation and potential mismatch between a person's name and their pronoun preferences. An exception is made for cases where the full name of a well-known public figure is provided and their gender can be unambiguously verified through external sources (e.g., documented pronouns or biographical data). This approach follows the annotation criteria described in the gENder-IT corpus as well (Vanmassenhove and Monti, 2021). However, in the Italian sentences, proper names are not directly tagged, but we tagged the gendered entity (article, adjective, past participle) that explicitly and directly refers to them.

Finally, instances of overextended masculine forms are explicitly labelled with <M>, as masculine is not treated as a gender-neutral default.

6. Annotation Procedure

The annotation was carried out manually on a sentence-by-sentence basis using the INCEpTION platform, which supports token-level tagging and entity alignment between the English and Italian texts. Each English sentence was first annotated to determine the contextual gender of all animate referents, following the tripartite tag set mentioned before. The corresponding Italian translations were then annotated by identifying the grammatical realisations of gender, such as agreement in adjectives, past participles, determiners, and pronouns. Ambiguous or unmarked forms were deliberately left untagged, as their neutrality provides essential evidence for gender omission or introduction in translation.

The overview of the aggregated annotation statistics shows the number of <M>, <F>, and <A> tags obtained from the annotation. The total number of tags for the English sentences is 1559, significantly higher than the number of tags for the Italian sentences (see Table 3). The main cause for that is that the majority of pronouns tagged in the source text are not present in the target sentences, as Italian is a pro-drop language; furthermore, English annotations also include the <A> tag, which was not considered for the Italian annotation task.

To ensure consistency, the annotation followed detailed linguistic rules governing agreement, morphology, article-noun concordance, and overextended masculine forms. A trained linguist annotated the sentences for gender using a rule-based framework grounded in grammatical and contextual cues. The annotation followed explicit linguistic criteria, ensuring objective and consistent labelling across the dataset.

7. Evaluation

To quantitatively assess how faithfully gender information is preserved in translation, this study ap-

Gender Tags	EN	IT Tower-LLM	IT mBART
M	356	544	570
F	294	210	181
A	909	0	0
Total Tags	1559	754	751

Table 3: Tag distribution by gender in English (EN) and Italian outputs from TowerLLM and mBART.

plies standard classification, such as Precision, Recall, and F1-score, to the annotated bilingual dataset. The comparison is carried out at the entity level, aligning tokens that share the same identifier across the two languages. The English annotations serve as the gold standard, while the Italian outputs produced by TowerLLM and mBART are treated as the models' predictions.

For each gender, masculine or feminine, Precision is calculated as the ratio between the correctly generated gendered forms in the outputs and all gendered forms produced by the models.

$$Precision = \frac{Correct\ Gendered\ Outputs}{Total\ Gender\ Tags\ (in\ Target)}$$

Recall is the ratio between gendered entities in the outputs and all the gendered forms in the source text.

$$Recall = \frac{Correct\ Gendered\ Outputs}{Total\ Gender\ Tags\ (in\ Source)}$$

F1-score, the harmonic mean of precision and recall, provides an overall index balancing accuracy and coverage.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The evaluation metrics were computed automatically through a custom Python scrip designed to measure how faithfully each system preserved explicit gender tags and how ambiguities were resolved from source to target language. To achieve this, the script aligns the English source entities with the Italian target realisations based on their unique entity identifiers. The comparison is performed on a sentence-by-sentence basis across the datasets. For each aligned triplet (English source, TowerLLM, and mBART targets), the code categorises the translation into one of the three outcomes:

1. ***matches***, where the Italian translation reproduces the gender correctly as it is in the source;
2. ***mismatches***, whenever the gender differs from source to target texts; this was further distinguished between (a) *error* and (b) *bias*. We considered errors those cases where the gender tag in the target does not match with the gender tag in the source (i.e., M→F or F→M), while biases refer to the cases of mismatches when a gendered tag in the target sentence is used instead of an ambiguous entity in the source (i.e., A→M or A→F). In this latter case, translating an ambiguous source entity into a specific gender is not inherently an error and yields an acceptable sentence. However, it is exactly the arbitrary and asymmetric choice of one gender over the other that allows us to measure the model's bias;
3. ***unmatched entities***, regarding those entities that are present in the source but absent in the target or vice versa.

Table 4 provides an example for each category and Table 5 provides a detailed comparison of these three different categories identified after the tagging process. On the basis of these comparisons, the script computes the metrics of Precision, Recall, and F1-score, separately for masculine and feminine realisations, treating the English annotations as the reference labels. These statistics are aggregated over all sentences to provide model-level measures for the accuracy of gender preservation (Table 5).

An additional component of the pipeline also addresses where the English source is ambiguous or unmarked (<A>). It analyses how each model resolves gender ambiguity by observing whether the Italian output adopts *masculine* (A→M) or *feminine* (A→F) forms. This step examines the models' behaviour when gender marking is not explicitly present in the source, thereby quantifying potential gender bias in translation. Instances of A→M or A→F indicate that an ambiguous entity has been rendered with an explicit gender. While such realisations are not inherently erroneous, systematic asymmetries between them may signal the presence of gender bias. The code additionally produces per-sentence logs and a .csv file reporting all matches, mismatches, and unresolved entities, facilitating a possible qualitative inspection.

8. Results

The metrics introduced in the previous section were applied to compare the performance of TowerLLM and mBART, using the manually annotated corpus as a reference standard. This section presents the results of our evaluation, focusing first on gender-wise accuracy and subsequently on the resolution of ambiguous cases.

✓ **MATCH** - Sentence 110 (F→F)	
EN source	Women have been trained to think that **we** <F1> are overreacting or that **we're** <F1> being too sensitive or unreasonable.
IT TowerLLM	Alle donne è stato insegnato a pensare che **siamo** troppo **reattive** <F1> o che **siamo** troppo sensibili o irragionevoli.
IT mBART	Le donne sono state addestrate a pensare che **siamo** troppo **reattive** <F1> o che **siamo** troppo sensibili o irragionevoli.
✗ **MISMATCH ERROR** - Sentence 164 (F→M)	
EN source	Many of the women working with me <A1> had to leave once **they** <F2> got married, because their husbands wouldn't let **them** <F2> work.
IT TowerLLM	Molte delle donne che lavoravano con me hanno dovuto lasciare il lavoro non appena si sono **sposate** <F2> perché i mariti non **li** <M2> volevano vedere lavorare.
IT mBART	Molte delle donne che lavoravano con me dovevano andare via una volta **sposate** <F2> perché i loro mariti non **li** <M2> lasciavano lavorare.
✗ **MISMATCH BIAS** - Sentence 125 (A→F/M)	
EN source	And he said, "No, **you** <A1>'re still dry, **you** <A1>'re just being nice."
IT TowerLLM	"No, sei ancora **asciutta** <F1>, stai solo facendo il **bravo** <M1>."
IT mBART	"No, sei ancora **secco** <M1>, stai solo facendo bene."
⚠ **UNMATCH** - Sentence 526 (A1→unmatched ID)	
EN source	**Lindsay Malloy** <A1>: **They** <A2> told **Brendan** <M3> that honesty would "set **him** <M3> free," but **they** <A2> were completely convinced of his guilt at that point.
IT Tower	Lindsay Malloy: A Brendan dissero che l'onestà lo avrebbe "**liberato**" <M3>, ma erano completamente **convinti** <M2> della sua colpevolezza a quel punto.
IT mBART	Lindsay Malloy: Hanno detto a Brendan che l'onestà lo avrebbe **liberato** <M3> ma erano completamente **convinti** <M2> della sua colpa a quel punto.

Table 4: Examples of gender agreement across translation systems (TOWER, MBART) for each evaluation category.

	TowerLLM	mBART
✓ **Match categories**		
Match M	173	165
Match F	104	88
Total Matches	**277**	**253**
✗ **Mismatch categories**		
Bias A→M	215	221
Bias A→F	35	29
Error M→F	0	2
Error F→M	8	25
Total Mismatches	**258**	**277**
⚠ **Unmatched EN IDs**	488	493

Table 5: Detailed comparison of matching and mismatch categories for TOWERLLM vs.MBART.

8.1. Gender-wise Performance

The gender-wise analysis evaluates how accurately each model reproduces explicit masculine or feminine references from the English source text in the Italian translation. Both models were assessed using the scores mentioned in the previous section.

Table 6 reports the Precision, Recall, and F1-scores of masculine and feminine outputs across the two systems. Overall, TowerLLM demonstrates slightly higher precision than mBART for both gender forms, indicating that its translations contain fewer erroneous outputs. However, mBART exhibits marginally lower recall, suggesting that it more frequently fails to reproduce gendered forms present in the source text.

These results highlight a systemic masculine bias across both models: masculine gender forms are reproduced more frequently but with lower precision, whereas feminine gender forms are generated more selectively and often under-represented. This imbalance reflects a persistent male-default tendency in translation, where masculine forms serve as the default choice. Consequently, while TowerLLM offers modest improvements in accuracy, both systems exhibit asymmetric gender distribution in translation that reinforces existing gender biases in

Model	Match	Total Matches	Total Tags (EN)	Precision	Recall	F1-Score
Male						
TowerLLM	173	544	570	31.8%	48.6%	38.3%
mBART	165	570	356	28.9%	46.3%	35.5%
Female						
TowerLLM	104	210	294	49.5%	35.4%	41.3%
mBART	88	181	294	48.6%	29.9%	36.9%

Table 6: Performance Metrics of Precision, Recall, and F1-score for TowerLLM and mBART by Gender. *Alternating columns are shaded for readability*.

MT (Table 6).

8.2. Ambiguous Case Analysis

Beyond explicit gender references, a crucial aspect of this evaluation concerns the translation of ambiguous entities: cases in which the source text provides no explicit lexical or contextual cues about the gender of an entity. Therefore, translating those entities into a morphologically richer target language compels the model to "make a choice", thereby exposing the gender bias.

The results in Table 5 (see "*Mismatches*") indicate that both models default to masculine realisations when ambiguity is present, showing high confidence in masculine predictions but very limited sensitivity to feminine alternatives. This male-defaulting tendency suggests that contextual cues in the English source are not sufficiently exploited to guide gender selection. The disproportionate number of A→M (ambiguous-to-masculine) over A→F (ambiguous-to-feminine) cases quantitatively confirms that gender introduction bias persists even in advanced LLM-based translation systems. This imbalance highlights a persistent structural asymmetry in how neutrality is resolved across languages, where masculine forms continue to function as the default. Moreover, the prevalence of A→M translations suggests that models do not just lack information, but they actively apply a masculine-centric heuristic when in absence of source cues.

9. Discussion and Conclusions

This study introduced the *Contextual Gender Annotation* (ConGA) framework as a linguistically grounded methodology for detecting and evaluating gender bias in MT. By aligning semantic gender in English with grammatical gender realisation in Italian, our evaluation reveals a systematic pattern of asymmetry in how LLMs handle gender. Both TowerLLM and mBART consistently over-generate masculine forms and under-generate feminine ones, particularly when handling gender-neutral or ambiguous input. These findings confirm that gender bias persists even in advanced LLM-based translation systems, reflecting a systemic gender imbalance rather than random variation.

From a linguistic perspective, this pattern reinforces the notion of a "*male-default*" translation bias observed in previous research (Cho et al., 2019; Vanmassenhove, 2024). It also exemplifies what Kahneman (2011) described as a *systematic bias*, a consistent deviation that skews judgments in one direction. When such bias operates at scale, it poses fairness concerns, as repeated biased outputs may be reproduced or re-absorbed into future training cycles, amplifying gender imbalance over time.

Beyond the empirical findings, this work aims to address a major gap in the field, namely the lack of a standardised and scalable evaluation for gender bias in MT. While ConGA is primarily an evaluation framework, the annotated data it produces can directly support mitigation strategies, such as providing high-quality preference pairs for Direct Preference Optimisation (DPO) or fine-tuning models to improve gender-aware agreement.

By integrating linguistic human expertise with reproducible computational evaluation, ConGA offers a scalable approach for systematic bias assessment. The guidelines represent an important step in automatically annotated datasets. Finally, while we employed a low generation temperature to ensure high semantic consistency for our gold-standard annotation, we acknowledge that different sampling parameters might influence gender resolution. Exploring the impact of higher temperature settings on the stability of the male-default bias remains an important direction for future research. Ultimately, linking linguistic insight and computational reproducibility makes the resulting dataset a reliable and reusable benchmark fro cross-system comparisons and broader cross-linguistic bias assessment.

10. Acknowledgements

This work has been funded by the National PhD programme in Artificial Intelligence, partnered by

the University of Pisa and the University of Naples "L'Orientale", through the doctoral grant 39-411-24-DOT23A27WJ-7219 established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan. Acknowledgements are also due to the support of The Eliza Centre for Humanities, Social Sciences, and Artificial Intelligence, whose interdisciplinary environment and resources have significantly contributed to the development of this research.

11. Bibliographical References

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Geert Booij. 1996. Inherent versus contextual inflection and the split morphology hypothesis. In Geert Booij, Jaap Van Marle, Stephen Anderson, Mark Aronoff, Laurie Bauer, Mark Baker, Rudie Botha, Joan Bybee, Andrew Carstairs-McCarthy, Wolfgang Dressler, Jack Hoeksema, Rochelle Lieber, Peter Matthews, Franz Rainer, Sergio Scalise, Henk Schultink, Arnold Zwicky, Geert Booij, and Jaap Van Marle, editors, *Yearbook of Morphology 1995*, pages 1–16. Springer Netherlands, Dordrecht. Series Title: Yearbook of Morphology.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Greville G. Corbett. 1991. *Gender*, 1 edition. Cambridge University Press.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender Aware Spoken Language Translation Applied to English-Arabic. ArXiv:1802.09287 [cs].

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

European Parliament. 2018. Gender-neutral language in the european parliament. Technical report, European Parliament, Luxembourg. Publications Office of the European Union.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16). ArXiv:1711.08412 [cs].

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57, Seattle, Washington. Association for Computational Linguistics.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Sally McConnell-Ginet. 2013. ' Gender and its relation to sex: The myth of 'natural' gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. DE GRUYTER.

Johanna Monti. 2017. Questioni di genere in traduzione automatica. In *Al femminile: scritti linguistici in onore di Cristina Vallini*, number 139 in Quaderni della rassegna. Franco Cesati Editore, Firenze.

Terttu Nevalainen and Helena Raumolin-Brunberg. 1993. Its strength and the beauty of it: The standardization of the third person neuter possessive in Early Modern English. In Dieter Stein and Ingrid Tieken-Boon van Ostade, editors, *1600 - 1800*, pages 171–216. De Gruyter Mouton, Berlin, Boston.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing Gender Bias in Machine Translation: a Case Study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

A. A. Rescigno, E. Vanmassenhove, J. Monti, and A. Way. 2020. A case study of natural gender phenomena in translation A comparison of google translate, bing microsoft translator and deepl for english to Italian, French and Spanish. In *CEUR Workshop Proceedings*, volume 2769. ISSN: 16130073.

Argentina Anna Rescigno and Johanna Monti. 2023. Gender Bias in Machine Translation: a statistical evaluation of Google Translate and DeepL for English, Italian and German. In *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, pages 1–11. INCOMA Ltd., Shoumen, Bulgaria.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. ArXiv:1804.09301 [cs].

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

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. Place: Cambridge, MA Publisher: MIT Press.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender Bias in Machine Translation and The Era of Large Language Models. ArXiv:2401.10016 [cs].

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An Annotated English-Italian Parallel Challenge Set for Cross-Linguistic Natural Gender Phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. 2024. GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models. ArXiv:2406.13925 [cs].