Gender Neutralization for an Inclusive Machine Translation: from Theoretical Foundations to Open Challenges

Andrea Piergentili[1,2] *, **Dennis Fucci**[1,2] *,
Beatrice Savoldi[1], **Matteo Negri**[1], **Luisa Bentivogli**[1]
[1]Fondazione Bruno Kessler   [2]University of Trento
`{apiergentili,dfucci,bsavoldi,negri,bentivo}@fbk.eu`

Abstract

Gender inclusivity in language technologies has become a prominent research topic. In this study, we explore gender-neutral translation (GNT) as a form of gender inclusivity and a goal to be achieved by machine translation (MT) models, which have been found to perpetuate gender bias and discrimination. Specifically, we focus on translation from English into Italian, a language pair representative of salient gender-related linguistic transfer problems. To define GNT, we review a selection of relevant institutional guidelines for gender-inclusive language, discuss its scenarios of use, and examine the technical challenges of performing GNT in MT, concluding with a discussion of potential solutions to encourage advancements toward greater inclusivity in MT.

1 Introduction

Language technologies have become ubiquitous and play a significant role in our lives. In addition to their benefits, however, these technologies come with potential ethical shortcomings and risks (Blodgett et al., 2020). Among them is gender bias, whose presence in machine translation (MT) is well-documented (Savoldi et al., 2021). Indeed, MT systems were found to systematically prefer masculine forms (e.g., EN *The student* → IT <u>*Lo*</u> (M) *studente*) and stereotypical gender associations in their outputs (e.g., EN *The doctors*

and the nurses → IT <u>*I*</u> *dottori* (M) *e* <u>*le*</u> *infermiere* (F)), thus reinforcing bias and reiterating the under-representation of specific groups (Savoldi et al., 2021). As the role of gender is relevant on the social level (Kiesling, 2019) and for each individual's construction of identity (Crenshaw, 1991), the biased behaviors of MT systems give rise to concerns about the consequent risks. These risks rest on the power of language to reproduce and reinforce societal asymmetries (Lazar, 2005), and its impact on our perception (Boroditsky et al., 2003; Gygax et al., 2008).

Spurred by the ever-growing demand for a gender-inclusive language, in this work we explore gender-*neutral* language as a form of gender inclusivity. It conforms to standard and established linguistic resources that allow to avoid gendered forms (e.g., *chairperson* instead of *chairman*) – unlike innovative elements like neopronouns and neomorphemes, which are not considered acceptable in many contexts (see Section 2.1). Comprehensive inquiries on gender-neutral MT are largely absent and its implementation is a substantially uncharted territory. Such gap calls for dedicated work on methodological underpinnings, such as the definition of the objectives and an outline of the main challenges to be faced when developing gender-neutral MT systems.

In light of the foregoing, in the present work we discuss the implementation of inclusive language in MT, through the definition of a novel task for MT: *gender-neutral translation* (GNT). For this purpose, we first provide a brief account of the relation between gender and language, and gender bias in MT (Section 2). Then, we focus on English-Italian translation and start by analysing relevant guidelines for gender inclusivity in both languages to understand the current theoretical

frameworks (Section 3). We chose this language pair because it is representative of the challenges faced by MT systems when translating across languages that express gender differently. This mismatch can result in undesired and discriminatory phenomena, such as the misgendering of individuals or stereotyped translations. Thereafter, we integrate the main principles of the guidelines into the context of MT, thus outlining a set of desiderata which define the task of GNT in higher detail (Section 4). Finally, we discuss the open technical challenges that performing and evaluating GNT in MT entails, and examine the potential approaches to address them, thus sketching a road-map towards the implementation of GNT in MT (Section 5).

2 Background

Gender is a complex notion, which has been extensively debated across several disciplines. In the case of language, the relationship with gender is socially relevant (Section 2.1), with an impact on both the visibility of gender groups (Wasserman and Weseley, 2009) and the construction of personal identities (Gygax et al., 2019). Therefore, the appropriate use of gender expressions is critical in communicative practices, including those allowed by language technologies (Section 2.2).

2.1 Gender in Language

The concept of gender is so relevant to human experience that no language lacks expressions of femaleness or maleness altogether (Stahlberg et al., 2007). However, languages differ in how they encode gender. English, for example, is a notional gender language: it expresses the gender of human referents mostly through personal pronouns and possessive adjectives (e.g., *he/him/hers*; *she/her/hers*), and lexically gendered forms (e.g., *man*; *woman*). Grammatical gender languages like Italian, instead, are characterized by a system of morphosyntactic agreement, where several parts of speech beside the noun (e.g., verbs, determiners, adjectives) carry gender inflections, as in *I/Le bambini/bambine sono contenti/contente* (EN *The children are happy*). Such differences are particularly relevant in translation, especially when the source language does not provide gender information about a referent and the target is a grammatical gender language, as in the previous example.

Regardless of cross-lingual differences, however, linguistic practices can be discriminatory when they generate a disparity in the representation of the genders based on normative and stereotypical principles. Androcentric normativity promotes the masculine gender as the human prototype, encompassing the whole human experience (Hellinger and Pauwels, 2007), thus treating women as a gendered deviation from the norm. A typical manifestation of normativity in language is the masculine generic, i.e., the use of masculine forms as conceptually generic, neutral (e.g., *one must watch his language*), when referring to mixed-gender groups or when gender is unknown or unspecified. Stereotypes, instead, are reiterated in the assumption of someone's gender through associations of professional nouns and gender (e.g., *nurse* = feminine, *doctor* = masculine) (He, 2010), fostering unfair gender paradigms. Moreover, within binary gender linguistic frameworks, non-binary experiences are completely omitted from representation.

In light of this, we look at gender-inclusive language[1] for the avoidance of discriminatory language. This is a form of *verbal hygiene* (Cameron, 1995) by which people attempt to regulate language in conformity to certain ideals, and promote linguistic policies that reflect them. The efforts to make language fair and inclusive of all gender identities can be distinguished in two main approaches: *i)* the introduction of innovative linguistic resources , and *ii)* the use of gender-neutral formulations. The first approach is the result of ongoing grassroots efforts, and includes neopronouns (EN *ze/zir* instead of *he/she/him/his/her*), neomorphemes (ES *-e/-es* instead of *-o/-os* and *-a/-as*), and other solutions (e.g., graphemic devices such as IT *-@* instead of *-a/-o/-e/-i*) that allow to mention referents without resorting to generic terms. The acceptance of such resources, however, is still highly debated and mostly restricted to informal communication channels like social media (Comandini, 2021). Thus, speakers who wish to use a gender-inclusive language in more formal contexts can turn to the second approach, which solely relies on established gender-neutral devices of the standard language. While some languages already

---

[1] The label "inclusive language" covers a wide range of linguistic practices aimed at avoiding discrimination and denigration on any basis (see `https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines`). Such practices have also been given different labels, such as 'neutral' and 'fair'. To set the object of our analysis within a larger scope of inclusivity, we hereby rely on the label *gender-inclusive language*.

feature convenient gender-neutral resources, such as the well established singular *they* in English,[2] speakers of other languages, such as Italian, cannot rely on similar devices. Then, they can resort to gender-neutralization strategies, such as the preference for epicene words, i.e. words that are not gender-marked and can be used regardless of the referent's gender (e.g., *spokesperson*, as opposed to *spokesman* and *spokeswoman*). Neutralization strategies range from simple word choices to complex sentence formulations without introducing innovative elements, thus being aligned with standardized forms and grammar. Consequently, we look at gender neutralization as a viable and grammatically acceptable form of gender-inclusive language, and a more solid ground for the exploration of gender-inclusive MT.

2.2 Gender (Bias) in Machine Translation

Although affecting many monolingual tasks in natural language processing (NLP), gender bias comes across more evident in cross-lingual scenarios, such as the case of MT, where different languages can encode very different gender marking mechanisms (Prates et al., 2020; Savoldi et al., 2021, *inter alia*). Most efforts to address gender bias in MT still operate in the binary perspective (Vanmassenhove et al., 2018; Stafanovičs et al., 2020, *inter alia*), thus ignoring the neutral solutions. By using gender-neutral forms, it is possible to avoid undue gendering when no information about the referents' gender is available, while also including all gender identities.

Recently, some works have started working on the processing of non-binary gender forms in NLP and highlighted the main challenges involved (Dev et al., 2021; Lauscher et al., 2022). They mainly focused on standard neutral solutions for text classification (Attanasio et al., 2021), coreference resolution (Cao and Daumé III, 2020), and natural language generation tasks, such as gender-neutral rewriting (Sun et al., 2021; Vanmassenhove et al., 2021; Attanasio et al., 2021). As regards MT, Cho et al. (2019) built a benchmark with template sentences to evaluate whether gender neutrality is preserved in Korean → English automatic translations. Working on English → German/Spanish, Saunders & Byrne (2020) also created a benchmark to assess the ability of MT systems to generate neutral target sentences. As the considered target languages do not have a neutral gender, they used gender-neutral placeholders for articles and inflectional morphemes. Finally, specific projects dedicated to gender-inclusive translation are also arising, like GenderFairMT,[3] with a focus on inclusive solutions for English → German MT (Burtscher et al., 2022).

Overall, adopting a neutral translation as a path towards gender inclusivity poses non-negligible challenges to MT. On the one hand, the complexity of implementing neutral forms comes from the inherent difficulties posed by grammatical gender languages, as also exemplified by the case study in (Saunders and Byrne, 2020). On the other hand, the application of an inclusive language must be carefully designed not to be perceived as intrusive nor as language policing.

In light of the foregoing, before we confront the technical challenges that arise from gender-neutralization in MT, we need to lay the groundwork for this endeavor. That is, framing the linguistic possibilities that could be adopted towards an automatic neutral translation, and identifying their suitable deployment.

3 Framing Gender-Inclusive Language

Looking for guidance to determine how MT systems should adopt gender-inclusive language, the MT scenario remains largely unexplored. Nonetheless, several resources intended for (human) communication are available and offer valuable linguistic knowledge for the understanding of gender-inclusive language and towards its adoption in MT. Among the most influential and accessible resources, there are the guidelines produced by renowned institutions to address gender discrimination in language. We consider them 'top-down' approaches in language, as opposed to the 'bottom-up' efforts of grassroots movements. Institutional guidelines currently only address monolingual communication while our domain of interest is translation. However, we analyze them to collect useful inclusive linguistic strategies, which let us investigate GNT and discuss its practical implications. More precisely, we intend to *i)* explore how gender inclusivity is conceptualized within such guidelines (Section 3.1), and *ii)* gain insights concerning *what* should be neutralized and *how* it

---

[2]See the American Psychological Association's style guidelines: https://apastyle.apa.org/style-grammar-guidelines/grammar/singular-they

[3]See https://genderfair.univie.ac.at/index_en.html

should be neutralized (Section 3.2).

To this aim, we selected 30 guidelines published online[4] by relevant institutions, equally divided between guidelines for English and Italian (see the full list of guidelines in Appendix A.1). Besides prestige, we prioritized comparability: we selected guidelines by international institutions (e.g., the European Union) that published the same document in both languages, or by national institutions (e.g., universities and governmental bodies) that share a similar status across countries, thus also ensuring that the selected guidelines belong to the same textual genre.

3.1 Conceptualization of Gender

Starting from *how these inclusive guidelines interpret gender*, and hence gender-based discrimination, we find clear differences between the English and the Italian documents. While the former mostly go beyond the binary gender framework, the Italian guidelines tend to address women and men only. Such a difference emerges clearly in the two versions of the European Parliament's guidelines (see documents E3, I5 in the reference list). This fundamental difference reflects different ideas of discrimination (e.g., E3: "achieving equality", I5: "achieving equality between men and women"). This conceptual discrepancy is reflected in the suggested strategies to address discrimination at the linguistic level. For instance, the Italian guidelines provide extensive lists of feminine counterparts for traditionally masculine professional nouns (e.g., EN *coordinator* as IT *coordinatore* [M] / *coordinatrice* [F]). Also, they often endorse gender specification to avoid masculine generics (e.g., EN *The professors* → *I professori* [M] *e le professoresse* [F]). Since such suggestions remain within a binary framework, they do not conform to our gender-neutral goal, and are hence discarded in the following discussion.

3.2 Neutralization Strategies

Moving on to the gender-neutralization strategies, here we discuss them through a multilingual perspective, focusing on their practical implications. In Table 1, we also offer a systematization that attempts to map strategies across English and Italian – except for highly language-specific solutions that are impossible to transfer.

Concerning *what should be neutralized*, we identify that these documents tend to largely focus on a particular form of gender discrimination: masculine generics. Masculine generics have been historically employed in administrative/legal texts to briefly refer to the public at large (e.g., see example B, where *he* refers to the whole occupational category of *judges*, and the Italian *il docente* [M], *professor* for the full teaching body). In the same vein, stereotypical associations and androcentric forms are discouraged (e.g., see example A in English). Overall, these guidelines are mostly concerned with generic referents. As we will discuss in Section 4, however, there are also circumstances where avoiding gender marks is necessary, e.g., to avoid misgendering individuals. Finally, and from a linguistic standpoint, we underscore that – as expected – English gender-inclusive strategies focus on the neutralization of pronouns (e.g., C, E), which are the main carrier of gender distinction in notional languages. Instead, the Italian guidelines prioritize the neutralization of nouns thus overlooking adjectives, pronouns, and verbs, which are subject to gender agreement too. Although the analyzed sentences are simple toy examples within an institutional genre, effective gender-inclusive solution should take into consideration the full range of gendered words in grammatical gender languages.

In light of the foregoing, we now delve into *how to avoid gender discrimination in language*. As previously anticipated, these top-down guidelines advocate for the use of neutralization strategies that conform to standardized, institutional language, over innovative, uncertain forms. As shown in Table 1, neutral solutions can vary greatly, ranging from omissions (e.g., E), and simple replacements of single words with epicene or collective nouns (e.g., A, B, D), to more complex reformulations that involve structural changes at the sentence level (e.g., F, G, H, I). On the one hand, though elegant, nouns replacement might be limiting if other gender-marked words are present, and only allow for a partial neutralization, e.g., as in IT *Il* [M] *professore* [M] *è tenuto* [M] *a rispondere* (EN *The professor must answer*) neutralized as *L'insegnante è tenuto* [M]. Moreover, the contextual nature of synonymy makes the choice of gender-neutral alternatives strictly case-specific (Edmonds and Hirst, 2002). When possible, however, the neutralization of short segments appears

		A. Epicene synonyms
EN	E5	*Chairman* → **Chair(person)**
IT	I3	*Professore* [Professor] → **Docente** [Teacher]

		B. Pluralization (towards generic or epicene forms)
EN	E2	A judge must certify that *he* has familiarized *himself* with...
		→ All **judges** must certify that **they** have familiarized **themselves** with...

		C. Relative and indefinite pronouns
EN	E5	If a staff member is not satisfied..., *he* can ask for a rehearing.
		→ Any staff member **who** is not satisfied... can ask for a rehearing
IT	I3	L'assicurazione... è a carico *del fruitore* [of the user].
		→ a carico di **chi fruisce** [of who uses].

		D. Collective and Role nouns
EN	§	Please contact one of the *waiters*.
		→ Please contact our **staff**.
IT	I3	Il palazzo ospita gli studi *dei professori* [of the professors] di slavo.
		→ Il palazzo ospita gli studi **del personale docente** [of the teaching staff] di slavo.

		E. Omission
EN	§	A person must reside... before *he* may apply for permanent residence.
		→ ...before ⎵ applying for permanent residence.
IT	I3	Un'accurata compilazione facilita *allo studente* [to the student] diverse
		→ ...facilita ⎵ diverse operazioni.

		F. Repetition
EN	E3	A manager may apply... if permission has been granted by *his* institution.
		→ ...if permission has been granted by **that manager**'s institution.

		G. Passive voice
EN	E5	Each action officer must send *his* document.
		→ Documents **must be sent**.
IT	I1	*Il* richiedente presenta la domanda [The applicant submits the application].
		→ La domanda **va presentata** [The application must be submitted].

		H. Imperative forms
EN	E5	Each staff member is requested to submit *his* information.
		→ Please **submit** all information.
IT	§	*Il cittadino* deve allegare [The citizen must attach] un documento.
		→ **Allega** [Attach] un documento.

		I. Impersonal forms
IT	I15	*Il candidato* decade [The candidate loses] dal diritto...
		→ **Si decade** [*One loses] dal diritto...

Table 1: Examples of neutralization strategies. In *red, italic* the generic masculine formulations; in **green, underlined** the gender-neutralizations. Column 2 provides the reference to the (E)nglish/(I)talian guidelines where each example was found (E1,2,3,..). If no example was found for a specific strategy within the guidelines, but the strategy is nonetheless be applicable, we fabricated an example (indicated with §). If a strategy is not applicable in one language, the corresponding example was omitted.

preferable as it makes the outcome more fluent, as opposed to more complex phrasings. This strategy is not always viable, though. Consider, for instance, the Italian term "*figlio/a*" (EN *child*): in lack of epicene synonyms, neutralization would require verbose periphrases, e.g., IT *minore a carico* (EN *underage, dependent child*) or *persona che si è concepita o adottata* (EN *person who was conceived or adopted*).

Neutralization strategies emerge as complex choices, to be carefully selected and weighted so as to preserve the effectiveness of communication and the acceptability of a text, i.e, features like fluency, style. Such choices, of course, highly depend on various constraints (e.g., register, length, context of use). Therefore, when adopting inclusive language, it is crucial to consider the possible trade-off between neutrality and the overall acceptability of the text where it is implemented. Moreover, as previously discussed, the feasibility and efficacy of adopting neutral strategies heavily depends on the context and the content of the source text. Therefore, such strategies are expected to be particularly pertinent in certain contexts, such as

the administrative-institutional domain – to which, it is worth noting, most monolingual guidelines belong. Different and less formal textual styles and contexts could present harder challenges in performing GNT because of the higher heterogeneity of their texts, where the strategies presented above might prove inapplicable or inappropriate. For instance, consider the translation of the simple, colloquial sentence EN *I have never been there* into Italian (*Non sono mai stato/stata lì*): none of the strategies in Table 1 is applicable here. However, compared to institutional and administrative communication, colloquial contexts tend to have greater tolerance to creative translations (e.g., IT *Non ho mai messo piede lì* – literally, EN *I have never set foot there*). Whether the system should resort to similar (or other) devices when straightforward solutions such as the strategies discussed above are not applicable is a decision that should be taken into account when building inclusive MT systems.

4 Desiderata for GNT in MT

In light of both the insights that emerged in Section 3, we now specifically address the use case of GNT, which allows MT systems to avoid discriminatory practices while conforming to standard linguistic forms. Specifically, **we define GNT as the task of automatically translating from one language into another without marking the gender of human referents in the target.** For example, given the English sentence *Your neighbors will thank you*, an inclusive MT system is required to translate *Il vostro vicinato[5] vi ringrazierà*, as opposed to *I vostri vicini vi ringrazieranno*, which features a masculine generic.

One crucial aspect of GNT is to determine when it should be performed, namely, when the marking of gender should be avoided or preferred. To this aim, and informed by our analysis of the existing guidelines, we devise three main desiderata to obtain a gender-neutral MT output, with specific examples in Table 2.

D1. Gender should not be expressed in the output translation when it cannot be properly assumed in the source. An inclusive MT system is expected to perform a gender-neutral translation in the target language when the gender of the referent(s) cannot be properly assigned from

---

[5]While the word *vicinato* is formally masculine, as a collective noun it is conceived as conceptually neutral.

the source. This scenario is quite frequent when translating from a notional gender language into a grammatical gender one, because of the gap in gender expression we discussed in Section 2.1. In these cases a gender-neutral translation refrains from any gratuitous assumptions, thus avoiding expressions which may: *i)* misgender a specific referent (Example 1); *ii)* exclude a social group, such as in the case of masculine generics (Ex. 2); *iii)* foster stereotypical associations (Ex. 3); adopt "androcentric" expressions (Ex. 4).

D2. Proper expressions of gender should be generated in the output translation if they are (indirectly) expressed in the source. The gender of some entities can be sometimes inferred through linguistic elements, which we may define as "gender cues". For example, in English, gender cues are 3rd person pronouns (*he/him/his*, *she/her/hers*), terms of address (e.g., *Mr./Mrs/Ms.*), gender-specific nouns (e.g, *boy, lady, lord, wife*). The presence of gender cues is crucial in determining whether a GNT is required or not. In (Ex. 5), the pronoun *herself* unequivocally identifies the referent as feminine. First names, surnames, or even nicknames, however, should not be included among these cues for several reasons. First names can hardly be considered a reliable index of someone's gender identity (Lauscher et al., 2022). Even in the attempt of any binary correlation, names and nicknames are highly ambiguous across genders and cultures (e.g., *Andrea*, which is typically masculine in Italian, but feminine in German). In addition, referents' gender could be known also through non-textual elements, such as explicit external information about who is speaking, which is sometimes provided to the translators. In all these cases, gender expressions are preferable in the translation.

D3. Masculine generics should not be propagated from the source language to the output translation. In spite of the seemingly straightforward definition of gender cues in D2, their recognition might not be clear-cut. This is the case of masculine generics used in the source, whose distinction from an actual gender cue might be equivocal. Hence, a MT system should be brought to carefully consider every information, in particular the word *man* along with its derivations and compounds so as to understand if they are used properly. For instance, to explicitly refer to the

(1)	EN	I refuse to give up on **a single student** in my class.
	IT	Mi rifiuto di lasciare indietro ***un solo* studente** nella mia classe.
	GNT	Mi rifiuto di lasciare indietro **qualsiasi studente** nella mia classe.
(2)	EN	**A lot of innovative teachers** began bringing comics...
	IT	***Molti* insegnanti *innovativi*** iniziarono a portare i fumetti...
	GNT	**Un gran numero di insegnanti all'avanguardia** iniziarono a portare i fumetti...
(3)	EN	We train **nurses** to do it, and they use local anesthetics.
	IT	Formiamo ***le infermiere*** a farlo, e loro usano anestetici locali.
	GNT	Formiamo **il personale infermieristico** a farlo, e loro usano anestetici locali.
(4)	EN	Vehicles may only proceed **at walking pace**.
	IT	I veicoli possono procedere solo ***a passo d'uomo***.
	GNT	I veicoli possono procedere solo **a passo di persona**.
(5)	EN	Even **the lead singer herself** abandoned the project.
	IT	Persino **la stessa cantante solista** ha abbandonato il progetto.
(6)	EN	It affects one to two percent of the population, more commonly **men**.
	IT	Riguarda dall'uno al due percento della popolazione, ed è più comune negli **uomini**.
(7)	EN	Earth was pristine before **men** appeared.
	IT	La Terra era incontaminata prima della comparsa ***degli uomini***.
	GNT	La Terra era incontaminata prima della comparsa **degli esseri umani**.
(8)	EN	The **fishermen** were so upset about not having enough fish to catch that...
	IT	I ***pescatori*** erano così ***disperati*** per la mancanza di pesce da pescare che...
	GNT	**Le persone che pescavano** erano così **disperate** per la mancanza di pesce da pescare che...
(9)	EN	Now when I was a **freshman** in college, I took my first biology class.
	IT	Quando ero **uno** studente al primo anno di università, seguii il mio primo corso di biologia.

Table 2: Examples for D1-3. We mark binary gender-marked expressions in ***red***, and in **green** those that are neutral.

masculine gender group as a whole (Ex. 6), where a neutralization would effectively compromise the meaning of the sentence. On the contrary, when they are used to refer to the totality of human beings (Ex. 7), or to entire categories of mixed-gender people through terms such as *fishermen* (Ex. 8), thus effectively functioning as masculine generics, they should be translated with neutral forms in the output. As there is not always a clear-cut distinction between a masculine generic and a masculine term used to refer to an actual masculine referent, and given the short context window within which MT systems operate, ambiguous cases can occur rather frequently. In these cases, a GNT should be considered the safest option as it avoids the propagation of the potential masculine generic without compromising the meaning of the sentence. Nonetheless, there is a specific case where gender cues ought to be considered as trustworthy; namely, in relation to the speaker as 1st person singular referent (Ex. 9). Based on the assumption that speakers deliberately choose the most appropriate expressions while talking about themselves, such a choice should be respected in the output translation.

In conclusion, we have outlined a set of three overarching desiderata towards the purpose of gender-inclusive MT. Such a scaffolding represents our proposed set of guiding principles to be applied towards the development of more inclusive MT models based on gender-neutral translation. In the next Section, we discuss the technical challenges of implementing such desiderata in MT.

5 Challenges and Insights for a Gender-Neutral Machine Translation

The adoption of neutral forms in MT could be conceived as a condition to be met or not met, without any intermediate nuance. The efficacy with which the condition is satisfied, on the contrary, can be rather nuanced; for example, there might be alternative inclusive solutions which might be perceived as more elegant or semantically closer to the input text, and others that satisfy these conditions to a lesser extent. Therefore, from a formal perspective, gender inclusivity can be likened to the concept of *constraint* (Garbacea and Mei, 2022).

As a constraint, it shows a multifaceted character, which makes it comparable to other types of well-known constraints adopted in automatic language generation (Garbacea and Mei, 2022). First, as seen in Section 3, gender inclusivity can be linguistically realized through both specific lexical forms and syntactic constructions. For this rea-

son, it can be likened to *lexical* and *syntactic constraints*. Then, the requirement of producing automatic translations that are as readable and fluent as possible, which is not always easily guaranteed in the case of neutral reformulations, makes gender inclusivity analogous to *utility constraints* (i.e. the criteria by which a text must exhibit characteristics such as coherence, comprehensibility, and faithfulness) (van Deemter, 2009). Nevertheless, gender inclusivity also summarizes the manifold challenges of the aforementioned constraints, thereby demonstrating a higher level of complexity. Below we illustrate the major challenges of satisfying such a multifaceted constraint, focusing on the dynamicity of the neutralization strategies (Section 5.1), the dearth of adequate training data and methods (Section 5.2), and the lack of evaluation procedures (Section 5.3).

5.1 Addressing the Dynamic Nature of Gender Inclusivity

To prevent unintended neutralizations, it is not always advisable to ensure GNT at all times (see Section 4). This condition makes neutral translation a "dynamic constraint", requiring MT systems to determine when to apply GNT. This ability, however, may be challenging to acquire, especially when gender cues are available outside the limited sentence context (e.g., *He was talking with a young <u>man</u>. Only later I realized that <u>this person</u> was a <u>professor</u>*). This presents a problem for current state-of-the-art MT systems, which work at the sentence level, i.e., by translating each sentence in isolation.

Alternative solutions that account for larger textual context in translation (Lopes et al., 2020) might be more apt to decide when performing neutral translations. For example, the design of MT models that translate beyond the sentence level ought to be considered. Translating sentences in a wider context, indeed, has proven crucial for correctly handling discourse cohesion (Bawden et al., 2018), and was shown to a certain extent beneficial to mitigate gender bias (Basta et al., 2020). However, it remains occasionally dubious whether context provides a useful linguistically-motivated knowledge (Kim et al., 2019; Li et al., 2020). Before venturing into any document-level endeavor, it is thus recommended to verify whether there is a positive interpretable link between gender-neutral translation, context-

informed MT, and overall quality of the system.

Besides gender cues, explicit external information too may contribute to the disambiguation of gender in the source sentence, thus guiding the neutral translation. For instance, speakers' metadata can be supplied in the form of tags, either at the word level (Stafanovičs et al., 2020) or at the sentence level (Vanmassenhove et al., 2018; Basta et al., 2020). Such prior knowledge, therefore, can also provide assistance in addressing the dynamic nature of gender inclusivity.

5.2 Constraining MT systems towards GNT

Future GNT-capable models are expected to learn to map words referring to human referents to corresponding neutral translations in order to satisfy the desiderata D1-3. Ideally, these models should be able to learn this mapping based on extensive training sets that include pairs of sentences with gender-neutral translations. To the best of our knowledge, however, training data that consistently have neutral forms in the target side (with a grammatical gender language as target) is lacking. It is necessary, then, to think of training methods that can overcome this lack of data, for example by taking inspiration from methods already applied to MT to satisfy other types of constraints.

Although various strategies have been proposed to make systems meet constraints (for an overview see (Garbacea and Mei, 2022)), it is crucial to evaluate which ones are applicable to the objective of gender inclusivity and how they can be adapted accordingly. The most straightforward method, for example, is to make the constraint explicit to the model directly in the input data. In the case of lexical constraints, this has been done by appending the constraint in the form of a target word or lemma to the source input so as to encourage the model to copy it in the output (Dinu et al., 2019; Song et al., 2019; Chen et al., 2020, *inter alia*). However, this approach is designed to work mainly at word-level, hence it would not be suitable when neutralization should involve several segments of the sentence. Moreover, this method requires bilingual dictionaries to map source words to target words. For gender inclusivity, however, such terminologies are not available, yet. Upon their creation, this technique could be taken into account when dealing with neutral source words which may be suitably translated with target epicene words.

Another line of solutions consists in restricting the search space at decoding time to sequences that contain the pre-defined constraints, such as specific words or phrases in the lexically-constrained MT. For example, Hokamp & Liu (2017) and Post & Vilar (2018) proposed modified versions of the beam search, which ensure that the translation hypotheses have met all the constraints before concluding the search. Similarly, Saunders & Byrne (2020) and Saunders et al. (2022) designed a constrained beam search pass to improve gender diversity – but for masculine/feminine forms only – in the *n*-best list by producing synthetic gendered alternatives of the original best hypothesis. Alternatively, some approaches were proposed to re-rank the *n*-best hypotheses according to additional scores, which informed whether or to which extent the constraints were satisfied, like in the dubbing-optimized MT (Saboo and Baumann, 2019) or in gender-specific translations (Saunders et al., 2022). Decoding and re-ranking methods, however, may also entail outputs of lower quality (Saboo and Baumann, 2019; Chousa and Morishita, 2021), due to the restriction of the search space and the trade-off between the need to satisfy the constraint and to faithfully reproduce the source text. Therefore, although such approaches may be a promising way to ensure gender inclusivity in automatic translations, their adoption too should be carefully evaluated.

5.3 Evaluating Gender-Neutral Outputs

The lack of dedicated test sets and metrics prevents the possibility of determining whether systems are actually making any advancements towards the resolution of a given task. In the case of gender-neutral MT, the benchmarks – traditionally designed as parallel data for reference-based evaluations – should comprise a range of source sentences aligned with target ones expressing either gender-marked or gender-neutral forms. As a suitable starting point, the domain of such a test set could be based on the institutional/administrative texts, since the guidelines available for gender-inclusive language belong to this domain (see Section 3). In addition to parallel data, specific protocols should also be designed to effectively evaluate whether the neutrality constraint has been satisfied.

Typically, MT evaluation methods involve comparing the output with a reference and measuring the degree of overlap between n-grams (Papineni et al., 2002; Popović, 2015) or the distance between the generated sentence and the reference in terms of edit operations required to make them equal (Snover et al., 2006). Some more sophisticated metrics take into account not only exact matches but also stems, synonyms, and paraphrases when comparing the MT output with the reference translation (Banerjee and Lavie, 2005). Alternatively, neural metrics use models to predict the similarity between the output and reference (or even directly between the source and output) (Rei et al., 2020). Although metrics that do not rely solely on surface similarity may be more appropriate for evaluating gender neutrality, it may be preferable to develop accuracy-like scores that isolate the evaluation of gender neutrality from the overall translation quality. This could involve annotating such expressions in the reference translation and attempting to match them, as done in MuST-SHE (Bentivogli et al., 2020). In such cases, accuracy is determined through string matching between expressions in the reference and in the output. Hence, the risk of mismatch remains present, as automatic neutralizations may be difficult to detect in an evaluation pipeline based on a single reference and may require extensive manual analysis to be identified (Savoldi et al., 2022). Using multiple references (Qin and Specia, 2015) that contain different neutral realizations to account for language variability could alleviate this difficulty. Another option would be to calculate accuracy without exploiting reference translations, as designed in WinoMT (Stanovsky et al., 2019). In WinoMT the aim is to identify the gendered translation through word alignment with the source, determine its gender through a morphological analyzer, and then check whether it corresponds to that of the source. However, our scenario includes an additional challenge, as in grammatical gender languages gender-neutral expressions may carry a formal gender (e.g. *la persona interessata* is a gender-neutral alternative of the masculine generic *interessato*, but it is formally feminine). Thus morphological analysis may be problematic.

Overall, effectively evaluating whether the output of an MT system is gender-neutral or gender-marked presents several challenges. These challenges need to be addressed to develop an accurate approach that can overcome the limitations of overall translation quality metrics and account for the intrinsic variability of gender-neutral solutions.

6 Conclusions

As a promising route forward to counter gender bias, in this work we have taken the first steps towards the adoption of gender-inclusive language in MT, focusing on the use of neutral forms devoid of gender marking for an English-Italian setting. To this aim, we reviewed various gender neutralization strategies presented in English and Italian guidelines for inclusivity, and outlined a definition of gender-neutral translation (GNT). Finally, we identified and discussed the technical challenges involved in implementing GNT in MT.

7 Acknowledgements

References

Attanasio, Giuseppe, Salvatore Greco, et al. 2021. E-mimic: Empowering multilingual inclusive communication. In *IEEE Big Data 2021*, pages 4227–4234.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Basta, Christine, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *The Fourth Widening Natural Language Processing Workshop*, pages 99–102.

Bawden, Rachel, Rico Sennrich, et al. 2018. Evaluating discourse phenomena in neural machine translation. In *16th NAACL HLT*, pages 1304–1313.

Bentivogli, Luisa, Beatrice Savoldi, et al. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *58th ACL*, pages 6923–6933.

Blodgett, Su Lin, Solon Barocas, et al. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *58th ACL*, pages 5454–5476.

Boroditsky, Lera, Lauren A. Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. In Getner, Dedre and Susan Goldin-Meadow, editors, *Language in Mind: Advances in the Study of Language and Thought*, pages 61–79. MIT Press.

Burtscher, Sabrina, Katta Spiel, et al. 2022. "Es Geht Um Respekt, Nicht Um Technologie": Erkenntnisse Aus Einem Interessensgruppen-üBergreifenden Workshop Zu Genderfairer Sprache Und Sprachtechnologie. In *Mensch Und Computer 2022*, page 106–118.

Cameron, Deborah. 1995. *Verbal Hygiene (The Politics of Language*. Routledge.

Cao, Yang Trista and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *58th ACL*, pages 4568–4595.

Chen, Guanhua, Yun Chen, et al. 2020. Lexical-constraint-aware neural machine translation via data augmentation. In *29th IJCAI*, pages 3587–3593.

Cho, Won Ik, Ji Won Kim, et al. 2019. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *First Workshop on Gender Bias in Natural Language Processing*, pages 173–181.

Chousa, Katsuki and Makoto Morishita. 2021. Input augmentation improves constrained beam search for neural machine translation: NTT at WAT 2021. In *8th Workshop on Asian Translation*, pages 53–61.

Comandini, Gloria. 2021. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, 23:43–64.

Crenshaw, Kimberle Williams. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–99.

Dev, Sunipa, Masoud Monajatipoor, et al. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *EMNLP 2021*, pages 1968–1994.

Dinu, Georgiana, Prashant Mathur, et al. 2019. Training neural machine translation to apply terminology constraints. In *57th ACL*, pages 3063–3068.

Edmonds, Philip and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Garbacea, Cristina and Qiaozhu Mei. 2022. Why is constrained neural language generation particularly challenging? *ArXiv e-prints arXiv:2206.05395*.

Gygax, Pascal M., Ute Gabriel, et al. 2008. Generically Intended, but Specifically Interpreted: When Beauticians, Musicians and Mechanics are all Men. *Language and Cognitive Processes*, 23:464–485.

Gygax, Pascal Mark, Daniel Elmiger, et al. 2019. A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10.

He, Guimei. 2010. An analysis of sexism in english. *Journal of Language Teaching and Research*, 1(3):332–335.

Hellinger, Marlis and Anne Pauwels, editors. 2007. *Handbook of Language and Communication: Diversity and Change*. De Gruyter Mouton.

Hokamp, Chris and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *55th ACL*, pages 1535–1546.

Kiesling, Scott F. 2019. *Language, Gender, and Sexuality: An introduction*. Routledge.

Kim, Yunsu, Duc Thanh Tran, and Hermann Ney. 2019. When and Why is Document-level Context Useful in Neural Machine Translation? In *Fourth Workshop on Discourse in Machine Translation*, pages 24–34.

Lauscher, Anne, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In *29th COLING*, pages 1221–1232.

Lazar, Michelle M. 2005. Politicizing gender in discourse: Feminist critical discourse analysis as political perspective and praxis. In Lazar, Michelle M., editor, *Feminist Critical Discourse Analysis: Gender, Power and Ideology in Discourse*, pages 1–28.

Li, Bei, Hui Liu, et al. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *58th ACL*, pages 3512–3518.

Lopes, António, M. Amin Farajian, et al. 2020. Document-level neural MT: A systematic comparison. In *22nd EAMT*, pages 225–234.

Papineni, Kishore, Salim Roukos, et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th ACL*, pages 311–318.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *WMT15*, pages 392–395.

Post, Matt and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *16th NAACL HLT*, pages 1314–1324.

Prates, Marcelo O. R., 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.

Qin, Ying and Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *18th EAMT*.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *EMNLP 2020*, pages 2685–2702.

Saboo, Ashutosh and Timo Baumann. 2019. Integration of dubbing constraints into machine translation. In *WMT19*, pages 94–101.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *58th ACL*, pages 7724–7736.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. *Findings of 60th ACL*, pages 3814–3823.

Savoldi, Beatrice, Marco Gaido, et al. 2021. Gender bias in machine translation. *TACL*, 9:845–874.

Savoldi, Beatrice, Marco Gaido, et al. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *60th ACL*, pages 1807–1824.

Snover, Matthew, Bonnie Dorr, et al. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th AMTA*, pages 223–231.

Song, Kai, Yue Zhang, et al. 2019. Code-switching for enhancing NMT with pre-specified translation. In *17th NAACL HLT*, pages 449–459.

Stafanovičs, Artūrs, Mārcis Pinnis, and Toms Bergmanis. 2020. Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In *WMT20*, pages 629–638.

Stahlberg, Dagmar, Friederike Braun, et al. 2007. Representation of the Sexes in Language. *Social communication*, pages 163–187.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *57th ACL*, pages 1679–1684.

Sun, Tony, Kellie Webster, et al. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv preprint arXiv:2102.06788*.

van Deemter, Kees. 2009. Utility and language generation: The case of vagueness. *Journal of Philosophical Logic*, 38(6):607–632.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *EMNLP 2018*, pages 3003–3008.

Vanmassenhove, Eva, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *EMNLP 2021*, pages 8940–8948.

Wasserman, Benjamin D. and Allyson J. Weseley. 2009. ¿qué? quoi? do languages with grammatical gender promote sexist attitudes? *Sex Roles: A Journal of Research*, 61:634–643.

A Appendix

A.1 Guidelines

The following guidelines for gender-inclusive language were analyzed for this study:

E1 United Nations Economic Commission for Western Asia, 2014
`https://archive.unescwa.org/sites/www.unescwa.org/files/page_attachments/1400199_0.pdf.`

E2 United Nations, 2018
`https://www.un.org/en/gender-inclusive-language/guidelines.shtml.`

E3 General Secretariat, Council of the European Union, 2018.
`https://www.consilium.europa.eu/media/35446/en_brochure-inclusive-communication-in-the-gsc.pdf`

E4 European Parliament, 2018
`https://www.europarl.europa.eu/cmsdata/187115/GNL_Guidelines_EN-original.pdf`

E5 North Atlantic Treaty Organization, 2020
`https://www.nato.int/nato_static_fl2014/assets/pictures/images_mfu/2021/5/pdf/210514-GIL-Manual_en.pdf`

E6 Australian Government, 2021
`https://www.stylemanual.gov.au/accessible-and-inclusive-content/inclusive-language/gender-and-sexual-diversity`

E7 University of Houston, 2022
`https://www.uh.edu/marcom/guidelines-policies/inclusive-language/_files/inclusive-language-guide.pdf`

E8 Australian National University, n.a.
`https://services.anu.edu.au/human-resources/respect-inclusion/gender-inclusive-language`

E9 United Nations Women, n.a.
`https://authoring.prod.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Gender-inclusive%20language/Guidelines-on-gender-inclusive-language-en.pdf`

E10 University of North Carolina at Chapel Hill, n.a.
`https://writingcenter.unc.edu/tips-and-tools/gender-inclusive-language/`

E11 University of Pittsburgh, n.a.
`https://www.gsws.pitt.edu/resources/faculty-resources/gender-inclusive-non-sexist-language-guidelines-and-resources`

E12 Royal Melbourne Institute of Technology, n.a.
`https://www.rmit.edu.au/content/dam/rmit/au/en/students/documents/services-support/lgbtiq/guide-inclusive-language.pdf`

E13 California State University San Marcos, n.a.
`https://www.csusm.edu/ipa/surveys/inclusive-language-guidelines.html`

E14 University of Otago, n.a.
`https://www.otago.ac.nz/humanresources/working-at-otago/equity/inclusive-language/index.html`

E15 The University of Texas at Austin, n.a.
`https://intranet.dellmed.utexas.edu/public/inclusive-language-guidelines`

I1 Cancelleria Federale Svizzera, 2012
`https://www.bk.admin.ch/dam/bk/it/dokumente/sprachdienste/Sprachdienst_it/02/objekt_40366.pdf.download.pdf/guida_al_pari_trattamentolinguisticodidonnaeuomo.pdf`

I2 Università di Torino, 2015
`https://www.unito.it/sites/default/files/linee_guida_approccio_genere.pdf`

I3 Università degli Studi di Padova, 2017
`https://www.unipd.it/sites/uni`

pd.it/files/2017/Generi%20e%20
linguaggi.pdf

I4 Segretariato Generale, Consiglio dell'Unione
Europea, 2018
https://www.consilium.europa.e
u/it/documents-publications/pu
blications/inclusive-comm-gsc/

I5 Parlamento Europeo, 2018
https://www.europarl.europa.eu
/cmsdata/187102/GNL_Guidelines
_IT-original.pdf

I6 Università degli Studi di Verona, 2020
https://docs.univr.it/document
i/Documento/allegati/allegati0
44384.pdf

I7 Università di Bologna, 2020
https://www.unibo.it/it/allega
ti/linee-guida-per-la-visibil
ita-del-genere-nella-comunic
azione-istituzionale-dell2019u
niversita-di-bologna/@@downloa
d/file/Linee-Guida-Genere-202
0.pdf

I8 Università degli Studi dell'Aquila, 2020
https://www.univaq.it/include/
utilities/blob.php?item=file&t
able=allegato&id=4925

I9 Università di Siena, 2021
https://www.unisi.it/sites/def
ault/files/allegatiparagrafo/L
INEE_GUIDA_Linguaggi_e_Generi.
pdf

I10 Istituto Universitario Federale per la For-
mazione Professionale, 2021
https://www.suffp.swiss/sites/
default/files/guida_per_un_lin
guaggio_inclusivo_20200610.pdf

I11 Università della Calabria, 2021
https://www2.unical.it/portale
/strutture/dipartimenti_240/fi
sica/pariopportunita/Linee%20g
uida%20Linguaggio%20di%20gener
e_15%20marzo%202021.pdf

I12 Università degli Studi di Milano, 2021
https://www.unimi.it/sites/def
ault/files/2021-12/Vademecuml

inguaggio%20di%20genere_Univer
sit%C3%A0%20degli%20Studi%20di
%20Milano.pdf

I13 Università Mediterranea di Reggio Calabria,
n.a.
https://www.unirc.it/documenta
zione/media/files/ateneo/pari_
opportunita/File_allegato_2.pd
f

I14 Università di Trento, n.a.
https://www.unitn.it/alfresco/
download/workspace/SpacesStore
/1185b2b5-dcfe-48ef-882b-e70
42fe4ff1a/documentolinguaggio2
9mar%20(1).pdf

I15 Università di Ferrara, n.a.
https://drive.google.com/file/
d/1P5Eq2jjoJtTjXGEV7TzyM4XJTcV
2PRyp/view