gENder-IT: An Annotated English–Italian Parallel Challenge Set for Cross-Linguistic Natural Gender Phenomena

Eva Vanmassenhove[1], Johanna Monti[2]

[1]Department of CSAI, Tilburg University
[2]UNIOR NLP Research Group, University of Naples L'Orientale
e.o.j.vanmassenhove@tilburguniversity.edu, jmonti@unior.it

Abstract

Languages differ in terms of the absence or presence of gender features, the number of gender classes and whether and where gender features are explicitly marked. These cross-linguistic differences can lead to ambiguities that are difficult to resolve, especially for sentence-level MT systems. The identification of ambiguity and its subsequent resolution is a challenging task for which currently there aren't any specific resources or challenge sets available. In this paper, we introduce gENder-IT, an English–Italian challenge set focusing on the resolution of natural gender phenomena by providing word-level gender tags on the English source side and multiple gender alternative translations, where needed, on the Italian target side.

1 Introduction

Cross-linguistic differences between languages often require implicit information in the source to be made explicit on the target side. When faced with systematic structural differences between the source and target languages, human translators rely on the (broader) context (linguistic, extra-linguistic, world-knowledge) in order to infer the necessary information and adapt the target side accordingly.

One such way in which many languages systematically differ is in terms of grammatical gender. Languages not only differ in terms of the absence or presence of specific gender features but also in the number of (linguistic) gender classes, how and where gender features are marked, and in the underlying rules by which gender is assigned (Audring, 2016).[1]

In languages with grammatical gender all nouns have an (arbitrarily) assigned lexical gender.[2] In most cases, the lexical gender of a noun is covert and can only be inferred from the morphological agreement with other words (articles, verbs, adjectives...) (Corbett, 1991). However, when nouns refer to animate referents, overt gender markings corresponding to the so-called 'natural gender' (biological sex) of the referent are common (e.g. the Spanish word for 'nurse' is 'enfermero' (male) or 'enfermera' (female)). Such forms are generated using derivational suffixes and are often derived from the 'generic male'. This process is sometimes denoted as 'female marking' (Doleschal, 2000; Laleko, 2018).

While language learners encounter difficulties memorizing the lexically stored gender of foreign nouns (Rogers, 1987), Machine Translation (MT) technology, given the limited (linguistic and extra-linguistic) context most MT tools leverage, struggles with the explicitation of ambiguous forms, i.e. the process of disambiguation. So far, little research has been conducted on controlling the output of MT systems in terms of features such as gender and/or number that arise due to specific cross-linguistic differences. We believe that there are two main reasons for this: (i) The research that has been conducted in this area shows that controlling specific features is a technically very challenging problem. Especially given the fact that it often requires in-depth linguistic knowledge and specialized linguistic tools, the performance of the latter often depending on how well-researched and

---

[1]Linguistic gender classes can (and often do) correspond to what is referred to in linguistics as the natural gender of referents (i.e. 'masculine' and 'feminine'). However, within the field of linguistics the term 'gender class' is somewhat con-

fusing as it is often used as a synonym for noun class. There are, for instance, language with more than 3 gender classes (e.g. Kiswahili has 9) as the classes are based on different semantic distinctions. Likewise, there are languags with only two 'gender classes' which correspond e.g. to an animate vs inanimate distinction.

[2]The Dutch word for 'sun' is 'zon' (female), while the French word for 'sun' is 'soleil' (male).

well-resourced the languages in question are; (ii) The lack of high-quality, human-crafted challenge sets that target specific cross-linguistic phenomena.

In this paper, we present a word-level (human) annotated, adapted and cleaned version of a subset (English-Italian) of the MuST-SHE corpus (Bentivogli et al., 2020).[3] The main contribution of our work is threefold: (i) The MuST-SHE corpus focuses solely on gender agreement with the first person singular. Our extension provides simple word-level annotations for all nouns and pronouns referring to human referents for the English sentences; (ii) While the transcripts of MuST-SHE are accompanied with gender information (male, female) of the speaker on a short paragraph-level, our word-level tags can be either male or female, but also ambiguous, when the sentence itself does not provide any explicit clues with respect to the gender of the referents; (iii) We focus on the textually gender-ambiguous sentences and provide all the correct gender-alternative translations for Italian.

The main motivations behind our work are the following: (i) First of all, there is a need for controlled diversity within the field of MT when it comes to controlling specific features of translations, specifically when dealing with structural cross-linguistic differences (Vanmassenhove, 2020). To allow for controlled diversity, we created the first test set that allows research on identifying ambiguity and generating multiple translation variants in terms of gender; (ii) Second, recent work by Saunders et al. (2020) indicates that even a (very) small synthetic set of high quality sentences annotated for gender can be leveraged to improve the accuracy of translations in terms of gender specific phenomena without decreasing the overall quality. Their work was limited to annotations for one referent per synthetic sentence and focused specifically on debiasing data in terms of gender. As highlighted in Vanmassenhove et al. (2019) and Saunders (2020), the effects of specific interventions need to be carefully examined on test sets that capture the complexity of a problem to its full extent. The manually annotated test set created does so by relying on 'natural' (as opposed to synthetic) data that is not limited to a single

human referent per sentence.

Bias statement (Blodgett et al., 2020)
In summary, this dataset is intended to encourage work on gender bias in MT, but could equally be leveraged for monolingual research on the generation of gender diverse translations (in Italian) and gender identification of referents (for English). The detailed analysis on English-Italian is intended to raise awareness on cross-linguistic differences between languages in terms of gender. NLP technologies are prone to the perpetuation (and possibly also the exacerbation (Vanmassenhove et al., 2021)) of inappropriate stereotypes and are currently unable to recognize or warn the user about the (gender) assumptions that have been made (e.g. by translating ambiguous source sentences systematically into one specific gendered variant on the target side). Furthermore, current systems lack the ability for the user to indicate and/or control the gender of referents if needed. As such, the gender of referents in the generated MT output depends entirely on the training data which might contain (un)conscious biases that are transmitted in (written and spoken) datasets.

2 Related Work

Recent years, several datasets were created that focus specifically on gender-related issues observed in (sub)fields of Natural Language Processing (NLP). Targeted gender datasets (test sets or corpora) exist for subfield such as coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Webster et al., 2018) and sentiment analysis (Kiritchenko and Mohammad, 2018). In this section, we will limit our discussion to datasets created specifically for mitigating and assessing gender bias in MT.

In the field of MT, Mirkin and Meunier (2015) used a recommender system approach to predicted user-based preferred translations based on preferences of similar users. Rabinovich et al. (2017) worked on personalized Statistical MT. Their work centers around the preservation of gender traits by treating gender as a separate domain. For their experiments, they created a bilingual parallel corpus (English–French and English–German) annotated, among others, with the gender of the speaker.[4] For Neural MT,

---

[3]The dataset is publicly available under a CC BY-NC-ND 3.0 through: `https://github.com/vnmssnhv/gENder-IT`.

[4]The dataset is publicly available: `http://cl.haifa.ac.il/projects/pmt/`

(Vanmassenhove et al., 2019; Vanmassenhove and Hardmeier, 2018) experimented with the integration of speaker gender-tags added to the source side of the parallel corpus. Using the demographic information released by Rabinovich et al. (2017), they compiled large datasets with gender information for 20 language pairs.[5] Both papers (Rabinovich et al., 2017; Vanmassenhove et al., 2019) focused specifically on gender agreement with the first person singular. As such, their corpora are limited to sentence-level gender-tags indicating the gender of the speaker.

Stanovsky et al. (2019) presented "WinoMT" a challenge set for the evaluation of gender bias in MT. The set is based on two existing data sets for gender bias in coreference resolution: WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018). WinoBias and Winogender consist of English sentences with two human entities in the form of two gender-neutral occupations (e.g. 'teacher', 'mechanic','assistant'...) and a gendered pronoun referring to one of the two human referents. WinoMT is a concatenation of WinoBias and Winogender and contains a total of 3,888 synthetic English sentences balanced for gender. The main contribution in Stanovsky et al. (2019) is an automatic evaluation of six popular MT systems on eight language pairs.[6] They provide an automatic gender bias evaluation protocol and show that the level of agreement with human annotations is above 85% for all languages.

Costa-jussà et al. (2020) presents the 'GeBioToolkit', a toolkit for the extraction of gender-balanced multilingual corpora with document-level gender annotations. They also introduce two versions of the 'GeBioCorpus'. The first one contains 16k sentences used for evaluating the automatically extracted parallel sentences. From the evaluation, it resulted that the human annotators gave the tool on average a 87.5% accuracy. The second version is a high-quality non-synthetic set of 2k English, Spanish and Catalan sentences post-edited by native speakers.

Saunders and Byrne (2020) created a small hand-crafted set of gender-balanced sentences for model adaptation. The set consists of 388 English synthetic sentences containing professions and their manually generated translations in each target language (Hebrew, German and Spanish). Saunders et al. (2020) explore the potential of explicit word-level gender inflection tags showing promising results. As such, gender tagging could be an effective tool for automatic translation tools where the user could specify the desired gender of the referents.

Our English-Italian parallel challenge set contains natural sentences (as opposed to synthetic) that do no follow a specific pattern[7] with word-level gender inflection tags. Since naturally occurring sentences are more complex and can contain multiple entities, animate nouns and pronouns have been annotated with word-level tags that indicate the gender given the limited sentence-level context. Unlike previous work, the challenge set is not limited to specific phenomena (e.g. 1st or 3rd person singular) but covers the full range of natural gender phenomena. It is specifically designed to encourage work on controlling output in terms of gender, the identification of gender ambiguous sentences and co-reference resolution.

3 Creation and Annotation of Dataset

In this section, we describe the pre-processing, cleaning and the gender annotations steps.

3.1 MuST-SHE

The gENder-IT challenge set is based on the MuST-SHE corpus comprising of naturally occurring sentences retrieved from TED Talks. We limited ourselves to the EN-IT parallel data and focused on data pertaining to what is referred to as 'category 2,3 and 4' in MuST-SHE, which are defined as sentences that contain contextual hints in terms of the gender of the speaker (category 2), sentences where both the audio signal and utterance context are needed to disambiguate the gender of referents (category 3) and sentences without contextual (audio or textual) gender information for disambiguation (category 4).

3.1.1 Corpus cleaning

While MuST-SHE contains segments (one or multiple sentences), we treated every sentence independently given that most state-of-the-art MT sys-

---

tems work on the sentence level. Aside from splitting the segments, sentences for which the target or source part was missing were removed, spelling mistakes corrected, and missing quotations marks and punctuation were added where missing. In total, 694 sentences were annotated.

3.2 Word-level gender tags

Annotations Word-level gender annotations are provided for all (pro)nouns referring to a person with exception to the few nouns in English that are already gender specific.[8] In example 3.2, the (pro)nouns are tagged with their respective genders based on the textual context, except for the noun 'dad'. The tags provided are <F> or <M> when it is clear from the sentential context that the referent should be referred to with male/female pronouns (see Ex. 3.2).

Example 3.1. 'So she turned and she looked at her dad, and she said, "Dad, I <F> know how you <M> feel, but I <F> don't believe in the death penalty."'

In all other cases, the <A> tag is used to indicate that within the given context, no assumption can or should be made with respect to the gender of the referent. When there are multiple <A> tags, we further distinguish between <A1>, <A2>, etc. to indicate that different entities are being referred to. This is important from a translational point of view, since it could imply that more than two translations need to be generated. For instance, in the following sentence (Ex. 3.2), two nurses (<A1> and <A2>) are mentioned refering to two different entities of which the gender, within this particular context, is unknown. In Italian, there is a male and female form for the English word *nurse*: infermiera (female) and infermiere (male), which implies that there are at least four correct translation alternatives in terms of gender.

Example 3.2. "And it was there that another nurse <A1>, not the nurse <A2> who <A2> was looking after Mrs. Drucker <F> before, but another nurse <A1>, said three words to me <A3> that are the three words that most emergency physicians <A4> I <A3> know dread."

Usually, annotating (pro)nouns suffices to indicate the contextual natural gender of referents, however in some cases, nounless adjectives can appear that refer to a human referent. Therefore, adjectives functioning as nouns (e.g. 'the rich'...) and/or adjectives used in a (conversational) constructions without a (pro)noun (e.g. 'And sporty <A>!) were tagged as well.

Proper names Many of the gender clues within the textual context referred to in the MuST-SHE corpus depend on the names of referents mentioned within the context. We opted for a slightly different approach in terms of proper names given the variety of naming conventions that exist in different cultures. Furthermore, a person's pronoun preferences might not necessarily match with the gender we traditionally or prototypically associated with a name. As such, proper names by themselves are not considered a gender clue (see Ex. 3.3).

Example 3.3. "Vera <A> was dead."

We make an exception for cases where the full name of a person is given and this person can be considered a 'public figure' for whom the pronouns can be retrieved, see Ex. 3.4.[9]

Example 3.4. 'The German physicist <M> Werner Heisenberg <M> said, "When I <M> meet... "'.

In total, 950 word-level tags are provided of which 138 are <F> (15%), 190 <M> (20%), and 622 <A(1-6)> (65%).[10]

3.3 Multiple Translations

Sentences that contain ambiguous referents, sometimes – depending on the target language – entail multiple equivalent translations in terms of gender. For 148 out of the 694 sentences annotated, this was the case and multiple gender alternative translations were provided in Italian.[11]

4 Analysis and Discussion

This section provides an analysis and discussion of the specific problems posed by the Italian language and the specific choices taken with respect

---

[8]Either due to the form: 'waitress', 'actress' or because of semantic features: 'mother', 'brother'...

[9]In practice and for consistency, we verified whenever a full name was given whether the referent has a Wikipedia page on which they are being referred to with specific pronouns.

[10]As outlined in the previous section, when there are multiple ambiguous referents we added an additional identifier (1-...) to indicate whether a sentence contains multiple ambiguous entities as this might have an influence on the amount of different gendered translations.

[11]Annotations were provided by linguists and the Italian translations were generated by a native Italian speaker/linguist.

to the gender translations proposed in the corpus. First of all, Italian is a pro-drop language and the subject pronoun is often omitted. Therefore in sentences where there are ambiguous subjects (I, you, we, they), like in:

Example 4.1. "Why did I <A> send her home?"

there is no need to produce alternative gender translations. However, if there are adjectives referring to ambiguous pronouns, gendered translations are needed, e.g.:

Example 4.2. "You <A1> know, I <A2> 'm really tired of this thing being called New Jersey."

for which we have two Italian sentences, namely "Disse: "Sono *stufo* di questa cosa chiamata New Jersey" for the masculine form and 'Disse: "Sono *stufa* di questa cosa chiamata New Jersey"' for the feminine form. The same applies when there are past participle forms in the sentence, since in Italian these forms sometimes require gender agreement with the noun they refer to such as in:

Example 4.3. "What did you <A> expect it to feel like?"

for which we produce the alternate gender translation: "Come pensavi che ti saresti *sentito*?" and "Come pensavi che ti saresti *sentita*?".

Gender translations were needed for bigender Italian nouns as well, such as for instance *insegnante* (teacher) or *paziente* (patient), which have a single invariable form for masculine and feminine and the gender becomes apparent only when there is a coordinated article or adjective , such as in

Example 4.4. "Do you <A1> remember that patient <A2> you <A1> saw with the sore throat?"

for which the sentences "Si ricorda *quel paziente* che ha visitato con il mal di gola?" and "Si ricorda *quella paziente* che ha visitato con il mal di gola?" were produced.

We also made a conscious decision in terms of the Italian 'non-marked' masculine form, also called the inclusive masculine - when the masculine form is used to refer, generically to males and females, such as for instance the use of the masculine form *bambini* (children) to refer to both male and female children. For this particular form, although the use of the inclusive masculine is acceptable when refering to a group of people whose gender is unknown (as proposed in e.g. Robustelli (2012)), we still opted to provide an alternative translation. For instance the sentence:

Example 4.5. "Man, I <A1> come home from work, drawers are open, clothes hanging outside the drawers, the kids <A2> are still in their pajamas..."

is translated as: "Amico, torno a casa dal lavoro, i cassetti sono aperti, i vestiti tutti fuori, *i bambini* sono ancora in pigiama..." and "Amico, torno a casa dal lavoro, i cassetti sono aperti, i vestiti tutti fuori, *le bambine* sono ancora in pigiama..."

The generic masculine form is also used for the agreement of adjectives/past participles/nouns in agreement with the natural gender of referents that have different genders, e.g.: "Giovanni e Lucia sono *bravi insegnanti*" (Giovanni and Lucia are good teachers). In this case, we kept the masculine form as no possible alternatives are currently accepted. Recently, the use of the schwa, "ə", was proposed (Gheno, 2019), precisely to solve these types of problems related to the use of the inclusive masculine form but also to take into account non-binary people representation needs, nevertheless this solution has not yet been widely adopted and is not accepted as a linguistic norm.

A further problem addressed in providing gender translation is related to the so-called agentive nouns, namely those nouns that are used to classify people that have specific functions, roles, professions. This type of nouns represent the main problem of sexism in the Italian language, and it is currently widely debated, since the tendency is to use male forms also to refer to professions or roles played by women. This is especially true for nouns which refer to particularly prestigious roles, such as *direttore* (director), *presidente* (president), *ministro* (minister), *professore* (professor) and the like, for which feminine nouns exist: *direttrice* (female director), *presidentessa* (female president), *ministra* (female minister), *professoressa* (female professor), etc. These forms are not always used (including by women) as some consider the feminization of a profession a loss of prestige.[12] For these cases, we opted to provide both masculine and feminine translations:

---

[12]Recently, the Accademia della Crusca, one of the most important research bodies for the Italian Language, discussed this problem with reference to the request by Beatrice Venezi to be presented as "direttore d'orchestra" (orchestra director) and not as "direttrice d'orchestra" (female orchestra director) during an important Italian song contest, namely the 71st Festival of Sanremo: https://accademiadellacrusca.it/it/consulenza/direttori-dorchestra-e-maestri-del-coro-anche-se-donne/2917

Example 4.6. "So I <A1> one day decided to pay a visit to the manager <A2>."

for which we provide the following alternate translations: "E così un giorno decisi di andare a trovare *il direttore*" and "E così un giorno decisi di andare a trovare *la direttrice*."

5 Conclusions and Future Work

In this paper, we present and describe gENder-IT: an English-Italian annotated parallel challenge set. The English source side is annotated with word-level gender tags, while for the Italian target side the translations –including correct gender alternatives– are provided. We present a detailed description of the annotations as well as a contrastive analysis of translation specific gender challenges for English–Italian. In future work, we envisage working on: (i) an extension of the corpus to other languages, (ii) the identification of gender ambiguous sentences in English, and (iii) the subsequent generation of multiple gender alternatives where necessary, including paraphrases to adopt more gender-neutral solutions. With our challenge set and analysis, we hope to encourage research on ambiguity detection and the controlled generation of gender diverse alternatives for translations.

Acknowledgments

References

Jenny Audring. 2016. *Gender*. Oxford Research Encyclopedia of Linguistics (online) accessed: 20-03-21.

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.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2020. GeBioToolkit: Automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, France. European Language Resources Association.

Ursula Doleschal. 2000. Gender assignment revisited. *Trends in Linguistics Studies and Monographs*, 124:117–166.

Vera Gheno. 2019. *Femminili singolari: il femminismo è nelle parole*. Effequ.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana.

Oksana Laleko. 2018. What is difficult about grammatical gender? evidence from heritage russian. *Journal of Language Contact*, 11(2):233–267.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain.

Cecilia Robustelli. 2012. Linee guida per l'uso del genere nel linguaggio amministrativo. *Progetto Accademia della Crusca e Comune di Firenze, Comune di Firenze*.

Margaret Rogers. 1987. Learners difficulties with grammatical gender in german as a foreign language. *Applied Linguistics*, 8(1):48–74.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL-HLT (2)*.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online).

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.

Eva Vanmassenhove. 2020. On the Integration of Linguistic Features into Statistical and Neural Machine Translation. *arXiv preprint arXiv:2003.14324*.

Eva Vanmassenhove and Christian Hardmeier. 2018. Europarl datasets with demographic speaker information. In *Proceedings of the 21st Annual Conference of the European Associations for Machine Translation (EAMT)*, Alicante, Spain.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3003–3008.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

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