GATE: A Challenge Set for Gender-Ambiguous Translation Examples

Abstract

Although recent years have brought significant progress in improving translation of unambiguously gendered sentences, translation of ambiguously gendered input remains relatively unexplored. When source gender is ambiguous, machine translation models typically default to stereotypical gender roles, perpetuating harmful bias. Recent work has led to the development of "gender rewriters" that generate alternative gender translations on such ambiguous inputs, but such systems are plagued by poor linguistic coverage. To encourage better performance on this task we present GATE, a linguistically diverse corpus of gender-ambiguous source sentences along with multiple alternative target language translations. We also propose metrics and provide tools for evaluation. We also motivate the need for source-awareness in gender rewriting as a strategy for this task.

1 Introduction

Gender is expressed differently across different languages. For example, in English the word *lawyer* could refer to either a male or female individual, but in Spanish, *abogada* and *abogado* would be used to refer to a female or a male lawyer respectively. This frequently leads to situations where in order to produce a single translation, a translator or machine translation (MT) model tends to choose an arbitrary gender to assign to an animate entity in translation output where it was not implied by the source. In this paper, we refer to this phenomenon as *arbitrary gender marking* and to such entities as Arbitrarily Gender-Marked Entities (AGMEs).

Translation with arbitrary gender marking is a significant issue in MT because these arbitrary gender assignments often align with stereotypes, perpetuating harmful societal bias (Stanovsky et al., 2019; Ciora et al., 2021). For example, MT models will commonly translate the following (from English to Spanish):

The surgeon $\overset{MT}{\Longrightarrow}$ *El cirujano* (*m*)

The nurse $\overset{MT}{\Longrightarrow}$ *La enfermera* (*f*)

Progress has been made to remedy this using a "gender rewriter" – a system that transforms a single translation with some set of gender assignments for AGMEs into a complete set of translations that covers all valid sets of gender assignments for a source sentence into the target language (Johnson, 2020). Using a rewriter:

The surgeon

$\Downarrow$ MT

El cirujano (m)

$\Downarrow$ rewriter

La cirujana (f)
El cirujano (m)

Although a step in the right direction, these rewriters often have poor linguistic coverage and only work correctly in simpler cases. Google Translate has publicly released such a system for a subset of supported languages, and we observe two error cases[1]:

1. Does not rewrite when necessary: *The director was astonished by the response of the community.* produces only one translation corresponding to masculine director.

2. Rewrites partially, or incorrectly: *I'd rather be a nurse than a lawyer* produces two translations but only lawyer is reinflected for gender (nurse is feminine in both).

To facilitate improvement in coverage and accuracy of such rewriters and reduce bias in translation, we release GATE[2], a test corpus containing gender-ambiguous translation examples from

[1] as observed on Jan 14, 2022

[2] Data and evaluation code available at [anonymized URL]

English (en) into three Romance languages: Spanish (es), French (fr) and Italian (it). Each English source sentence[3] is accompanied by one target language translation for each possible combination of masculine and feminine gender assignments of AGMEs [4]:

*I know **a Turk** who lives in Paris.*
⇓ *Italian*(it)
Conosco una turca che vive a Parigi. (f)
Conosco un turco che vive a Parigi. (m)

GATE is constructed to be challenging, morphologically rich and linguistically diverse. It has ∼ 2000 translation examples for each target language, and each example is annotated with linguistic properties (coreferent entities, parts of speech, etc.). We additionally propose a set of metrics to use when evaluating gender rewriting modules.

Finally, this corpus was developed with the help of bilingual linguists with significant translation experience for each of our target languages (henceforth *linguists*). Each is a native speaker in their respective target language. We spoke in depth with our linguists about the nuances of gender-related phenomena in our focus languages and we share our analysis of the relevant aspects and how they impact our work and the task of gender rewriting.

2 Linguistic Background

[Add references]

2.1 Arbitrarily Gender-Marked Entities

In this paper, we use *animate entity* (or just *entity*) to refer to an individual or group for which a referential gender could be implied in either the source or target language[5]. Usually this will refer to humans, but may also be extended to some animals and mythical or sentient beings. For example, *cat* is generally translated into Spanish as *gato*, but *gata* is also frequently used to refer to a female cat. We use *referential gender* to refer to an entity's gender as a concept outside of any linguistic considerations.

To qualify as an AGME, an entity's gender must be ambiguous in the source sentence, but have a referential gender implied by one or more words in the target translation. Compared to Romance languages, there are relatively few ways that gender is denoted through word-choice in English. Most notably, English uses a handful of gendered pronouns and possessive adjectives (*she*, *her*, *hers*, *he*, *him*, *his*), as well as a relatively small number of animate nouns that imply a gender (e.g. *mother*, *father*, *masseuse*, *masseur*, etc). There is also often a correlation between certain proper names and referential gender (e.g. *Sarah* is traditionally a female name and *Matthew* is traditionally male), but we do not consider this a reliable enough signal for gender determination unless they are a well known public figure (e.g. *Barrack Obama* is known to be male). We follow Vanmassenhove and Monti (2021) in this.

Additionally, an AGME must have some gender marking in the translation. In the following English-Italian example,

*I heard the thief insult **his interlocutor**.*
⇓ *it*
Io ho sentito il ladro insultare la sua interlocutrice.
Io ho sentito il ladro insultare il suo interlocutore.

interlocutor→*interlocutrice* (f) / *interlocutore* (m) is an AGME, while *thief*→*ladro* and *I*→*Io* are not. *Thief* is unambiguously male because of its coreference with *his* in the source, while *I* has ambiguous gender which is not marked in the target.

2.2 Gender in Romance Languages

In Spanish, French and Italian, all nouns have a grammatical gender – either masculine or feminine. For inanimate objects, this gender is fixed and often arbitrary; for example, in French, *chaise* (chair) is feminine, while *canapé* (couch) is masculine. When a noun or pronoun refers to an animate entity, its grammatical gender will, with some notable exceptions, match the referntial gender of that entity.

In these languages, referential gender of entities is frequently marked through morphology of an animate noun (e.g. *en-es*: *lawyer* ⇒ *abogada*(f), *abogado*(m)) or through agreement with gendered determiners, adjectives and verb forms.

Some animate nouns are *dual gender*, meaning that the same surface form is used for both masculine and feminine, such as French *artiste* (artist). However, other clues to the artist's gender may exist in a French sentence through gender agreement with other associated words. For example, *The tall*

[3] A few non-sentence utterances are also included as well, such as noun-phrases and sentence fragments

[4] The majority of source sentences contain only one AGME and thus two translations

[5] For simplicity, we limit our discussion of gender and linguistics to masculine and feminine within the scope of this paper, but we do not intend to imply that gender is limited in this way.

artist could be translated into French as *La grande artiste*(f) or *Le grand artiste*(m). Here, grammatical gender of translations of *the* (*la* (f) / *le* (m)) and *tall* (*grande* (f), *grand* (m)) must match the referential gender of the referent noun.

Dual-gender determiners and adjectives exist as well, such as Spanish *mi* (my) and *importante* (important). So for example, Spanish *mi huésped importante* (My important guest) has no gender marking. Similarly, in French and Italian, some determiners may contract before vowels to lose their gender marking. Feminine and masculine forms of *the* in French, *le* and *la*, both contract before vowels (and sometimes *h*) to become *l'*, so *l'artiste* (the artist) is not marked for gender.

Similarly to English, some pronouns in Romance languages are inherently gendered, while others are not. Entities referred to by gender-neutral pronouns, such as Spanish *yo* (I) and *tú* (you) commonly become gender-marked through predicative gender-inflecting adjectives. Further complicating these cases, first and second person subject pronouns (analogous to *I*, *you*, *we*) are frequently omitted in Spanish and Italian (but notably not in French) as the subject is clear from verb morphology. This means that in some cases, the AGME in a sentence pair may be a zero-pronoun, such as English *I am* ***tired*** being translated to Spanish as *estoy cansada* (f) or *estoy cansado* (m). There is no word in these translations corresponding to *I*, but the subject is implied by the verb form *estoy*.

3 Related Work

Gender Bias in NMT has been studied extensively. Rabinovich et al. (2017) presented work on the preservation of author's gender. Prates et al. (2018) construct a templatized test set, using a list of job positions from the U.S. Bureau of Labor Statistics (BLS), in 12 gender-neutral languages and translate these sentences into English using Google Translate. They observe that Google shows a strong inclination towards defaulting to masculine forms, especially for fields related to science, engineering, and mathematics. Escudé Font and Costa-jussà (2019) studied the impact of using debiased and gender neutral word embedding on the gender debasing on NMT.

Though a slew of challenge sets (Mirkin and Meunier, 2015; Vanmassenhove and Hardmeier, 2018; Vanmassenhove et al., 2018; Stanovsky et al., 2019), containing both synthetic and natural sentences, have been published to evaluate the work on translating unambiguously gendered inputs, work on translating ambiguously gendered sentences has been limited.

Zmigrod et al. (2019) proposed a generative model that allows conversion between masculine inflected and feminine inflected sentences in four morphologically rich languages with a focus on animate nouns.

Habash et al. (2019) approach ambiguous input translation as a gender classification and reinflection task for target language sentences to address the first-person singular cases when translating from English into Arabic. Given a gender-ambiguous source sentence and its translation, their system provides an alternative translation using the opposite gender. Furthermore, they create a parallel corpus of first-person singular Arabic sentences that are gender-annotated and reinflected.

Alhafni et al. (2021) expand on Habash et al. (2019)'s Arabic Parallel Gender Corpus by adding second person targets as well as increasing the total number of sentences. More specifically, they include contexts involving first and second grammatical persons covering singular, dual, and plural constructions and we add six times more sentences. Vanmassenhove and Monti (2021) introduced an English-Italian dataset where the English sentences are gender annotated at the word-level and paired with multiple gender alternative Italian translations (for 148 sentences) when needed.

Google Translate announced[6] an effort to address gender bias for ambiguously gendered inputs by showing both feminine and masculine translations.

Our work sits in the intersection of efforts like Zmigrod et al. (2019) and Vanmassenhove and Monti (2021) . Similarly to Zmigrod et al. (2019), we are interested in reinflection but like Vanmassenhove and Monti (2021) we make use of source context as well.

4 GATE Corpus

We present GATE corpus, a collection of bilingual translation examples designed to challenge source-aware gender-rewriters. The linguists were asked to compile roughly 2,000 examples for each target language.

[6]: Google AI Blog: A Scalable Approach to Reducing Gender Bias in Google Translate (googleblog.com)

4.1 Anatomy of an Example

Each example in the data set consists of an English sentence with at least one AGME, and a set of alternative translations into the given target language corresponding to each possible combination male/female gender choices for each AGME. Variation among the alternative translations is restricted to the minimal changes necessary to naturally indicate the respective gender-markings.

We also mark several category features on each example, such as what class of animate noun AGMEs are (profession, relationship, etc), what grammatical role they play in the sentence (subject, direct object, etc), sentence type (question, imperative, etc) and several other phenomena. These are discussed further in Section 4.3, and detailed statistics are provided in

Additionally, each example is accompanied by a list of AGMEs as they appear in the English source, as well as their respective masculine and feminine translations found in the translated sentences. For multi-word phrases, only the head noun will be marked. For example, if *police officer* is translated to *policía* in Spanish, the English field would list only *officer*. Because of this, the English and target language fields may not be direct translations of one another. However, we found specifying the head noun to be more convenient for use with parsers and other natural language processing tools.

The same entity may be referred to multiple times in the same sentence through coreference. Coreferent mentions of AGMEs are listed joined by '='. For example, in the following *en-es* example, the English AGME field would contain "doctor=lawyer".

I'd rather be a ***nurse*** *than a* ***lawyer****.*

⇓ es

Prefiero ser enfermera que abogada. (f)

Prefiero ser enfermero que abogado. (m)

Finally, In cases where an AGME is represented by a pronoun that is elided in the translation, it will be represented by the nominative case form and be enclosed in parentheses. For example, in the following example, the Spanish AGME field would contain *(yo)*:

I am ***tired****.*

⇓ es

Estoy cansada. (f)

Estoy cansado. (m)

4.2 Corpus Development Process

The linguists were asked to aim for a distribution of sentences lengths ranging from very short (< 10 words) to complex (> 30) words. Actual example counts are shown in Table 1. Of the 2,000 examples for each language, linguists were asked to include roughly the following breakdown:

- 1,000 single animate noun AGME
- 500 single pronoun AGME
- 500 with two or more AGMEs

Linguists were given details of the various categories and attributes listed in section 4.3 and asked to try to find sentences such that each such category is well represented (depending on the relative ease of finding such sentences). Linguists were also asked to prioritize diversity of animate nouns where possible. They were allowed to pull examples sentences from natural text or construct them from scratch as they saw fit. However, except for a small number of toy examples, we asked that they include only sentences that were natural in both English and their target language, and could reasonably appear in some imaginable context.

We provided samples of web-scraped data that had been filtered with various heuristics to help identify sentences fitting some of the harder-to-satisfy criteria. For example, we used Stanza (Qi et al., 2020) to filter some web-scraped data for those containing an animate noun marked as an indirect object and provided this to the linguists. In some cases these sentences were used directly, and in others they were modified slightly to fit the requirements.

Throughout the process, we prioritized diversity of sentence structure, domain and vocabulary. Rather than produce a representative sample, our intention was to produce a corpus that would challenge any tested systems on a wide range of phenomena.

4.3 Sentence Categories

[ToDo. There's a large table and a bit of discussion in Appendix A now. Deciding how much to put here and how much in Appendix]

5 Proposed Use

5.1 Gender Rewriting

Our goal in development this corpus is to facilitate generation of multiple translations covering all

data set	**< 10**	**10-19**	**20-29**	**>= 30**	total
Spanish 1 AGME	477	722	197	105	1501
Spanish 2+ AGMEs	70	176	56	21	323
French 1 AGME	704	661	171	14	1550
French 2+ AGMEs	177	222	41	4	444
Italian 1 AGME	397	867	195	48	1507
Italian 2+ AGMEs	93	500	139	30	762

Table 1: Distribution of lengths (words) of English utterance per target language and AGME count

valid gender assignments. One strategy for producing such a set of translations is to first use an MT model to produce a default translation and then use a rewriter to generate one or more alternative translations with other gender assignments (Johnson, 2020).

$$\text{source} \xRightarrow{\text{MT}} \text{translation} \xRightarrow{\text{rewriter}} \{\text{all translations}\}$$

It is important to note that although all source-translation pairs in GATE have AGMEs, in more general translation scenarios most will not. In these cases the rewrite step should be a no-op. One could break such a rewriter into two components (Habash et al., 2019): (1) a gender-ambiguous entity detector that determines if the source and default translation contain any AGMEs, and optionally what those AGMEs are and (2) a gender-rewriter that is called only if any AGMEs are present. We explain how GATE can be used to evaluate the quality of both these components in following sections.

5.2 Rewriter Evaluation

We formalize the task of gender rewriting on a single-AGME sentence as follows: given the source sentence src, target translations corresponding to male and female referent entities, and a rewrite direction (M to F or F to M), produce an output target translation with the alternative gender from the original translation. We will refer to the original input translation as tgt_0, the desired/reference translation as tgt_1 and the output generated by the rewriter as hyp:

$$rewriter(src, tgt_0) = hyp \sim tgt_1$$

For this task, we believe that looking at exact full-sentence matches between hyp and tgt_1 is the most sensible approach for evaluation. We do not give partial credit of changing the gender markings on only a subset of the words to those found in tgt_1. Doing so will generally result in a sentence that is either grammatically incorrect due to newly introduced agreement errors, or for which the semantics has changed in an unacceptable way, such as a changed coreference. Because of this, we find sentence-similarity measures such as BLEU and words error rate not to reflective of a user's experience [citation for bleu].

For some production use-cases, the cost of surfacing an incorrect alternative translation may be higher than of surfacing only a single translation. To support such scenarios, we may allow the rewriter to choose not to output an alternative translation when confidence is low. Additionally, if operating on an expanded data set including negative examples, it may do say when it does not believe that an alternative translations is appropriate.

We can calculate precision as the proportion of correct alternatives among those attempted, while recall is equivalent to accuracy, i.e. the proportion of correct alternatives produced among all sentences, even if not attempted. Using these definitions of precision and recall, we find $F_0.5$ to be an ideal overall metric, prioritizing precision while still incorporating an idea of coverage.

We focus our discussion of evaluation on sentences containing a single AGME, which will typically lead to exactly two alternative translations. GATE also includes a smaller number of examples with more than one AGME, and these have more than two alternatives translations and thus more than one correct output for a rewriter. We do not formalize evaluation on this subset here but believe that the data set will be useful in evaluating rewriting systems capable of producing multiple outputs for multiple sets of gender assignments.

5.3 Experiments

[Add description of our rewriter, and discussion of results]

	P	R	F0.5
es f2m	0.972	0.5	0.818
es m2f	0.951	0.397	0.743
fr f2m	0.962	0.212	0.563
fr m2f	0.933	0.145	0.446
it f2m	0.962	0.469	0.794
it m2f	0.911	0.319	0.665

Table 2: Our rewriter's scores on GATE for each target language and rewrite direction

5.4 End-to-End Evaluation

In our envisioned scenario, a gender rewriter would operate on the output of an MT system. It is unlikely, however, that direct MT output will consistently match GATE's translations word-for-word. Because of this, references cannot be used as is, and human evaluation is necessary to test a rewriter's output in conjunction with MT, or a combined system that simply outputs gendered-alternative translations from a single source sentence. One consideration is that a parallel sentence that contains an AGME in GATE, may not contain an AGME when machine translated, as the MT output may be unmarked for gender.

In order to test our combine system end-to-end, we use our production MT models to translate the source sentences from GATE into Spanish and pass that output to our rewriter [add numbers for other languages]. We then ask annotators for the following annotations.

- If one only translation is returned, is the target gender marked for an ambiguous source entity?
- If two are returned, correct if following are true:
  - Is the target gender-marked for an ambiguous source entity?
  - Were all the words marking gender of AGME changed correctly?
  - Were only the words marking gender of AGME changed?
- throw out sentences for special cases:
  - Translation quality is too poor to judge
  - There are multiple AGMEs (possible if there were multiple ambiguous source entities, but only one was marked in the GATE reference)

We also retrieve translations for these sentences from Google Translate's English-Spanish translation system. For this translation direction, Google Translated does in many cases produce two gender-alternative translations. We asked annotators to annotate these translations in the same manner. Results are presented in Table 3.

	P	R	F0.5
Our System	0.927	0.495	0.789
Google Translate	0.982	0.267	0.640

Table 3: end-to-end scores for our system and Google translate's

[To Do: recalculate with the updated test set and add analysis. These numbers come from a small portion (about 200 sentences) of a previous iteration of the data]

A full, end-to-end evalution should include testing on both sentences with and without AGMEs. Because each example in GATE contains at least one AGME, we recommend supplementing GATE with examples from Renduchintala and Williams (2021) and Vanmassenhove and Monti (2021), which contain unambiguously gendered source entities. In future work, we intend to develop a supplemental data set for GATE containing various types of negative examples: unambiguous source entities, entities that are unmarked in both source and target, and inanimate objects whose surface forms are distractors (e.g. depending on context, *player* and *cleaner* may refer to either objects or people).

6 Test Corpus Analysis

6.1 Source-Aware Rewriting

[remove or reduce this section]

In most prior formulations of the rewriting task, rewriters only take in the original target translation to determine whether to rewrite (one could imagine you check for candidates for reinflection in the target). However, we believe that in order to be successful in this task, one must employ **source-aware rewriting** that uses both the source sentence and target translation to determine what, if any, rewrite is appropriate.

The target sentence is required because you can only rewrite if there are tokens that need to be changed. Often an English word will have multiple possible translations into a target language but not all require reinflection. For example, the Spanish words *triste* and *descontento* are both reasonable translations of "*unhappy*" but *triste* can be used for either gender while *descontento*/*descontenta* are masculine and feminine forms respectively.

The source sentence is also required. Take, for example, the three sentences below. All three could reasonably be translated into Spanish as *La abogada se comió el almuerzo*, in which *La abogada* is marked for feminine referential gender. In the final English sentence (containing *her*) the lawyer is unambiguously female and the translation therefore should not be rewritten. In the first two sentences, it is appropriate to rewrite the sentence to support a hypothetical male lawyer.

The lawyer ate lunch (neutral)
The lawyer ate their lunch (neutral)
The lawyer ate her lunch (fem)
$\Downarrow$ es
La abogada se comió el almuerzo (fem)

This demonstrates the necessity of source-awareness in determining what rewrites of a target sentence are valid and shows that both source and target are necessary in the general case.

6.2 Multi-Entity Interactions

One interesting pattern we observed is the "gender-linking" of distinct nouns when they implicitly constitute a single group. For example in the following translation set, *soldiers or mercenaries* can be thought of as a singular group entity with regards to gender marking.

No ***soldiers*** *or* ***mercenaries*** *were injured.*
$\Downarrow$ es
ningún soldada o mercenaria resultó herido.(f)
ningún soldado o mercenario resultó herido.(m)

After considering such sentences, our linguists concluded that translations that mark *soldiers* with a different gender from *mercenaries* would be unusual here, and so, at their discretion, cases like this may be treated as single entities. These examples are marked with the GLNK category marker (see Section 4.3).

Another common pattern is that of coreferent mentions of a single entity, which must by definition have the same referential gender, and usually but not always the same grammatical gender. For example, in the following sentence, *friend* and *nurse* are the same individual and we would typically expect them to share the same referential gender in a direct translation into any of the target languages.

My best friend is a nurse

While typically an entity's referential gender will align with its grammatical gender, there are a handful of animate nouns in our focus languages for which grammatical gender is fixed, regardless of the referential gender of the individual. These are referred to as epicene nouns. Most notable among these is the direct translation of *person* into each of the target languages, which is always grammatically feminine: *La persona* (*es,it*) or *La personne* (*fr*). We also find some language-specific words with fixed grammatical gender. For example, these Italian words are always grammatically feminine: *la guardia* (guard), *la vedetta* (sentry), *la sentinella* (sentry), *la recluta* (recruit), *la spia* (spy).

In cases where one coreferent mention uses a word with fixed grammatical gender as discussed above, the grammatical genders of coreferent mentions may in fact differ. In the following sentence, the described individual is unambiguously male. The phrase *una buena persona* (a good person) is grammatically female, while *un mal amigo* (a bad friend) and *él* (he) are grammatically male.

He is a good person but a bad friend.
$\Downarrow$ es
Él es una buena persona, pero un mal amigo.

6.3 Masculine Generics

Traditionally, many languages, including Spanish, French and Italian, employ a paradigm known as masculine generics. Under this paradigm, feminine formas are considered to be explicitly gender-marked, while masculine forms should be used in situations where referential gender is unclear. Specifically, when referential gender is unknown by the speaker, or a mixed-gender group is known to contain at least one male individual, defaulting to grammatically masculine forms is generally considered correct in the language standard[7]. In this sense, masculine gender marking does not imply the exclusion of female-identifying individuals, but a feminine gender marking would imply the exclusion of male-identifying individuals.

In most cases where a masculine generic might be used, we nonetheless ask our linguists to provide an alternative translation with feminine gender-marking. However, we annotate such generic men-

[7] In recent years there is some explorations of using novel, gender-neutral forms in these contexts

tions with the label INDF (*indefite gender*), so that users who wish to to follow a stricter interpretation can exclude these examples in their evaluations.

We also observe cases where a gendered pronoun (most often masculine) may be used in English in reference to an indefinite entity, but where referential gender of the entity is not implied by the gender of the pronoun. For example, in phrasings such as the following, there may be no implication that *the reader* is in fact male.

We ask the reader to close his eyes.

7 Conclusion

We have presented GATE, a corpus of hand-curated test cases designed to challenge gender rewriters on a wide range of vocabulary, sentence structures and gender-related phenomena. Additionally, we provide an in-depth analysis of many of the nuances of grammatical gender in Romance languages and how it relates to translation. We also suggest metrics for gender rewriting and provide tools to aid with their calculation. Through this work we aim to improve the quality of MT output in cases of ambiguous source gender, as well as facilitate the development of better and more inclusive natural language processing (NLP) tools in general.

We look forward to future work in improving GATE and related projects. We aim to add additional languages pairs to GATE and investigate translation directions into English. We also hope to supplement with additional data, including negative examples. Finally, we plan to explore use of gender-neutral language use in various languages and how it can be incorporated into NLP applications.

8 Bias Statement

In this work, we propose a test set to evaluate translation of ambiguously gendered source sentences by NMT systems. Our work only deals with English as the source and is currently scoped to Romance languages as the target. To construct our test set, we have worked with bilingual linguists for each target language. We plan to increase scope of both source and target languages in future work.

Through this work, we hope to encourage and facilitate more inclusive use of natural language processing technology, particularly in terms of gender representation. In recent years, there is significant ongoing movement in the way gender manifests in languages use. One form that this takes is in new gender-neutral language constructs in Romance languages such as French, Spanish and Italian to accommodate gender underspecificity and non-binary gender identities. We wholeheartedly support the development of this more representative and inclusive language, and endeavor to find ways to support it through technology. In this work, however, for the sake of simplicity we restrict our scope to language as used to express gender along more conventionally binary lines, and we therefore do not consider non-binary language or word forms. We are working with both language experts and non-binary-identifying individuals to expand the scope to include non-binary and gender-underspecified language in future work.

A Category Labels

There are a wide range of linguistic phenomena that can interact with gender translation. We have devised several category labels that can apply to examples. In order to ensure diversity within the data set, linguists were asked to ensure that a certain minimum number of examples are included for each such label. This also has the benefit of helping pinpoint weaknesses in an evaluated system. For example, a rewriting system may perform well when the ambiguous noun is the subject of the sentence, but do poorly when it is a direct object. Unless otherwise stated, category labels are determined based on the target sentence set rather than the source sentence, as this is generally the more important input to the rewriter. A single example will typically have multiple labels.

- **Grammatical Role categories:** A key noun is a subject (SUBJ), direct object (DOBJ) indirect object (IOBJ), subject complement (SCMP), or object of a preposition (OPRP, excluding indirect objects)

- **Animate Noun categories:** profession (PROF, e.g. doctor), Religion (REL, e.g. Bhuddist), Nationality (NAT, e.g. Italian), Family and other relationships (REL, e.g. neighbor), Non-Human (NHUM, e.g. cat, vampire), Other (OTH, e.g. winner, accused)

- **Adjectives and past participles:** attributive (AATR), predicative (APRD), past-participle form as an adjective (PPA), past-participle form not as an adjective (PPNA), Adjective modifies non-ambiguous noun (ANAN). Most of these distinctions are included to test a rewriter's ability to distinguish between adjective surface forms that should be modified along with key nouns and those that should not.

- **Sentence Types categories:** Headline (HEAD), sentence fragment (FRAG), question (QUES), imperative (IMPR), Ambiguous noun in a subordinate clause (SUBC)

- **Other categories:** Plural ambiguous noun (PLUR), indefinite i.e. does not refer to an entity concretely known by the speaker, e.g. "Where can I find a good doctor?" (INDF), Requires agreement across different clauses with noun that was ambiguous in source (DFCL), Distinct animate nouns behave as a single group and are *gender-linked* (GLNK)

[include stats per category and add some discussion]

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2021. The arabic parallel gender corpus 2.0: Extensions and analyses.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in turkish and english machine translation models. *CoRR*, abs/2108.10379.

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.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate.

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. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2021. Investigating failures of automatic translation in the case of unambiguous gender. *CoRR*, abs/2104.07838.

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.

E. Vanmassenhove and C. Hardmeier. 2018. Europarl datasets with demographic speaker information. In *EAMT 2018 - Proceedings of the 21st Annual Conference of the European Association for Machine Translation*.

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. Genderit: An annotated english-italian parallel challenge set for cross-linguistic natural gender phenomena.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

ID	es	fr	it	description
Semantic Type				
PROF	1168	490	1208	Profession word
NAT	118	249	157	Nationality or locality membership
REL	25	150	29	Religious affiliation
FAM	327	250	192	Family or other relationship
NHUM	2	40	–	Non-Human
OTH	580	941	708	Other
Grammatical role				
SUBJ	1638	1221	1573	Subject
SCMP	118	185	121	Subject complement
DOBJ	181	328	399	Direct object
IOBJ	136	275	165	Indirect object
OPRP	250	279	518	Object of preposition
VOC	3	–	4	Vocative
POSC	80	–	289	Possessive complement
Sentence Type				
QUES	124	–	–	Question
FRAG	49	101	–	Sentence Fragment
IMPR	14	135	–	Imperative
Adjective-related				
APRD	82	359	213	Predicative adjective agreeing with AGME
AATR	293	190	315	Attributive adjective agreeing with AGME
ANAN	97	1026	–	Adjective modifying a word other than AGME
PPA	361	172	290	Adjective has same surface form as a past participle
APPS	–	35	22	post-positive adjective – remove??
Pronoun subtype				
PERS	–	219	146	Personal pronoun
RELA	–	15	13	Relative pronoun
DEMO	–	64	28	Demonstrative pronoun
POSS	80	–	–	Possesive pronoun
DROP	157	–	–	AGME is a dropped/zero pronoun
IPRO	–	369	53	Indefinite pronoun
Other				
PLUR	991	1110	1042	Plural
INDF	–	–	229	Indefinite/masculine generic could apply
DFCL	136	113	–	Changed words in alternatives cross clause boundaries
GLNK	–	94	–	"gender-link" – AGMEs are not coreferent but conceptually linked, different genders would be unnatural

Table 4: Counts of sentences with each category label per language. ’–’ indicates that this language was not annotated for this label