Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1870–1884
Marseille, 20-25 June 2022

The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses

Bashar Alhafni, Nizar Habash, Houda Bouamor[†]
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
[†]Carnegie Mellon University in Qatar
{alhafni,nizar.habash}@nyu.edu, hbouamor@cmu.edu

Abstract

Gender bias in natural language processing (NLP) applications, particularly machine translation, has been receiving increasing attention. Much of the research on this issue has focused on mitigating gender bias in English NLP models and systems. Addressing the problem in poorly resourced, and/or morphologically rich languages has lagged behind, largely due to the lack of datasets and resources. In this paper, we introduce a new corpus for gender identification and rewriting in contexts involving one or two target users (I and/or You) – first and second grammatical persons with independent grammatical gender preferences. We focus on Arabic, a gender-marking morphologically rich language. The corpus has multiple parallel components: four combinations of 1st and 2nd person in feminine and masculine grammatical genders, as well as English, and English to Arabic machine translation output. This corpus expands on Habash et al. (2019)'s Arabic Parallel Gender Corpus (APGC v1.0) by adding second person targets as well as increasing the total number of sentences over 6.5 times, reaching over 590K words. Our new dataset will aid the research and development of gender identification, controlled text generation, and post-editing rewrite systems that could be used to personalize NLP applications and provide users with the correct outputs based on their grammatical gender preferences. We make the Arabic Parallel Gender Corpus (APGC v2.0) publicly available.

Keywords: Arabic, Gender Bias, Machine Translation, Controlled Generation, Text Rewriting, Gender Identification

1. Introduction

The great recent advances in many NLP applications have raised expectations about their end users' experiences, particularly in regards to gender identities. Gender negative and positive stereotypes are manifest in most of the world's languages (Maass and Arcuri, 1996; Menegatti and Rubini, 2017) and are propagated and amplified by NLP systems (Sun et al., 2019), which not only degrades users' experiences but also creates representational harms (Blodgett et al., 2020). Although human-generated data used to build these systems is considered the main source of these biases, balancing and debiasing the training data do not always lead to less biased systems (Habash et al., 2019). This is because the majority of NLP systems are designed to generate a single text output without considering any target user gender information. Therefore, to prevent this and to provide the correct user-aware output, NLP systems should incorporate their users' grammatical gender preferences when available. Of course, this becomes more challenging for systems targeting multi-user contexts (first, second, and third persons, with independent grammatical gender preferences). One example of this phenomenon is the machine translation of the sentence *I am a doctor and you are a nurse*. While English uses gender neutral terms leading to ambiguous gender references for the first and second persons (*I/doctor* and *you/nurse*), some morphologically rich languages use gender-specific terms for these two expressions. For instance, in Arabic, a gender-unaware single-output machine translation from English often results in أنا طبيب وأنت ممرضة *ÂnA Tbyb wÂnt mmrDħ*[1] 'I am a [male] doctor and you are a [female] nurse', which is inappropriate for female doctors and male nurses, respectively.

Alternatively, gender-aware personalized NLP systems should be designed to produce outputs that are as gender-specific as the user information they have access to. Users information could be either embedded as part of the input (e.g., 'she is a doctor and he is a nurse') or provided externally by the users themselves. Aside from context complexity, there is a lack of resources and datasets for morphologically rich languages, where multi-user expressed differences are ubiquitous. In this paper, we focus on Arabic, a gender-marking morphologically rich language. We introduce a new parallel corpus for gender identification and rewriting in contexts involving one or two users – first and second grammatical persons with independent grammatical gender preferences – I only, you only, *and* I and You. This corpus expands on Habash et al. (2019)'s Arabic Parallel Gender Corpus (APGC v1.0) by adding second person targets as well as increasing the total number of sentences over 6.5 times, reaching over 590K words. The Arabic Parallel Gender Corpus (APGC v2.0) also has multiple parallel components: four combinations of 1st and 2nd person in feminine and masculine grammatical genders, as well as English, and English to Arabic machine translation output.

We make this data publicly available hoping that it will

[1]Arabic transliteration is in the HSB scheme (Habash et al., 2007).

encourage research and development of gender identification, controlled generation, and post-editing rewrite systems that could be used to personalize NLP applications and provide users with the correct outputs based on their grammatical gender preferences.[2] While the work focuses on Arabic, we believe many insights and ideas are easily extensible to other languages, depending on their linguistic requirements.

Next, we discuss some related work (§2) and then give a background on Arabic linguistic facts (§3). We describe the selection process and the annotation guidelines of our newly created corpus in §4. We then present an overview and analysis of our corpus in §5. Lastly, we show how our corpus could be used to study gender bias in commercial machine translation systems (§6) and we conclude in §7.

2. Related Work

Several approaches have been proposed to mitigate gender bias in various NLP tasks including machine translation (Rabinovich et al., 2017; Elaraby et al., 2018; Vanmassenhove et al., 2018; Escudé Font and Costa-jussà, 2019; Stanovsky et al., 2019; Costa-jussà and de Jorge, 2020; Gonen and Webster, 2020; Saunders and Byrne, 2020; Saunders et al., 2020; Stafanovičs et al., 2020; Saunders et al., 2021; Savoldi et al., 2021; Ciora et al., 2021), dialogue systems (Cercas Curry et al., 2020; Dinan et al., 2020a; Liu et al., 2020a; Liu et al., 2020b; Sheng et al., 2021b; Sheng et al., 2021a), language modeling (Lu et al., 2018; Bordia and Bowman, 2019; Sheng et al., 2019; Vig et al., 2020), co-reference resolution (Rudinger et al., 2018; Zhao et al., 2018a), and named entity recognition (Mehrabi et al., 2019). The majority of these approaches focus either on debiasing word embeddings (contextualized or non-contextualized) before using them in downstream tasks (Bolukbasi et al., 2016; Zhao et al., 2018b; Gonen and Goldberg, 2019; Manzini et al., 2019; Zhao et al., 2020; Lauscher et al., 2020; Katsarou et al., 2022), classifying gender bias along multiple dimensions (Dinan et al., 2020b), adding additional information to the input to enable models to capture gender information correctly (Vanmassenhove et al., 2018; Moryossef et al., 2019; Stafanovičs et al., 2020; Saunders et al., 2020), or creating gender-balanced corpora through counterfactual data augmentation techniques (Lu et al., 2018; Hall Maudslay et al., 2019; Zmigrod et al., 2019).

In terms of rewriting, Vanmassenhove et al. (2021) and Sun et al. (2021) recently presented rule-based and neural rewriting models to generate gender-neutral sentences.

When it comes to morphologically rich languages, (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 when needed. For Arabic, Habash et al. (2019) created APGC v1.0 – a parallel corpus of first-person-singular Arabic sentences that are gender-annotated and reinflected. They selected the sentences from a subset of the English-Arabic OpenSubtitles 2018 dataset (Lison and Tiedemann, 2016). Each sentence is labeled based on the grammatical gender of its singular speaker as F (feminine), M (masculine), or B (ambiguous). For the M and F sentences, they introduced their parallel opposite gender forms. Moreover, they developed a two-step model to do gender identification and reinflection. They demonstrated the effectiveness of their approach by applying it to the output of a gender-unaware machine translation system to produce gender-specific outputs. In the same line of work, Alhafni et al. (2020) used APGC v1.0 to create a joint gender identification and reinflection sequence-to-sequence model. They treated the problem as a user-aware grammatical error correction task and showed improvements over Habash et al. (2019)'s system.

Our work expands APGC v1.0 by including contexts involving 1st and 2nd grammatical persons covering singular, dual, and plural constructions; and adding six times more sentences.

3. Arabic Linguistic Background

We provide background on the two main challenges that face Modern Standard Arabic (MSA) NLP systems when it comes to gender expressions: morphological richness and orthographic ambiguity.

Morphological Richness Arabic is a morphologically rich language that inflects for gender, number, person, case, state, aspect, mood and voice, in addition to various attachable clitics such as prepositions, particles, and pronouns (Habash, 2010). Gender in Arabic has two values: masculine (*M*) or feminine (*F*), whereas number has three values: singular (*S*), dual (*D*), and plural (*P*). Gender and number apply to verbs, nouns, and adjectives. They are commonly expressed using inflectional suffixes that represent some number and gender combination (for nominative indefinite): ∅+ (*MS*), ة+ +*ħ* (*FS*), ان+ +*An* (*MD*), تان+ +*tAn* (*FD*), ون+ +*wn* (*MP*), and ات+ +*At* (*FP*). For instance, the noun ممرض *mmrD* 'nurse' (*MS*) could have the following forms: ممرضة *mmrDħ* (*FS*), ممرضان *mmrDAn* (*MD*), ممرضتان *mmrDtAn* (*FD*), ممرضون *mmrDwn* (*MP*), and ممرضات *mmrDAt* (*FP*). Additionally, Arabic has many idiosyncratic templatic stem changes and inflectional suffixes that are not consistent in indicating a specific gender and number combination (Alkuhlani and Habash, 2011). In such cases, the functional (grammatical) gender and number do not match the form-based (morphemic) gender and number. One example of the so-called *Broken Plurals* in Arabic demonstrate this well: the plural of عبقري

[2]The corpus is available through the CAMeL Lab Resources page: `http://resources.camel-lab.com/`

ςbqry 'genius [m.sg]', عباقرة ςbAqrħ 'geniuses [m.pl]', has a feminine singular suffix but is a masculine plural noun. Similarly, the word خليفة *xlyfħ* 'caliph' is masculine but uses a feminine suffix.

Gender and number in Arabic participate in the morphosyntatic agreement between verbs and their subjects, and between nouns and their adjectives. However, they also interact with a morpholexical feature called *rationality* – a feature that is associated with human referring nouns such as man, princess, and doctor (Alkuhlani and Habash, 2011). For example, adjectives modifying rational nouns agree with them in gender and number, while adjectives modifying irrational plural nouns are always feminine and singular.

Orthographic Ambiguity and Noise In addition to its morphological richness and complexity, Arabic is also orthographically ambiguous as it uses optional diacritics to specify short vowels and consonantal doubling. As these diacritics are optional, Arabic readers deduce the meaning of words based on the sentential context. Since some gender-specific words only differ in diacritics, Arabic orthography makes such distinctions ambiguous.

In the context of text generation, this is sometimes a useful feature as it allows the same text to be interpreted differently by the target users. For example, the question ما اسمك؟ *mA Asmk?* 'what's your name?' can be diacritized as مَا اسْمُكَ؟ *mA Asmuka?* [2nd.m.sg] or ما اسْمُكِ؟ *mA Asmuki?* [2nd.f.sg].

Finally, it has been shown that unedited MSA text could have a significant percentage (∼23%) of spelling errors (Zaghouani et al., 2014). The most common errors include Alif-Hamza spelling (أ، إ، آ، ا *A, Ā, Ǎ, Â*), Ya spelling (ي، ى *y, ý*), and the feminine singular suffix Ta-Marbuta (ة، ه *h, ħ*). Therefore, Alif/Ya normalization is a standard processing in Arabic NLP as it reduces some of the noise (Habash, 2010). This high degree of orthographic ambiguity and noise poses challenges for automatic learning systems due to confusability and data sparsity.

4. Extending the Arabic Parallel Gender Corpus

In this section, we describe the selection criteria and the annotation process of the APGC v2.0. To the best of our knowledge, no such corpus exists for Arabic or any other language.

4.1. Corpus Selection

As in Habash et al. (2019), we selected the original set of sentences from the English-Arabic OpenSubtitles 2018 dataset (Lison and Tiedemann, 2016), which includes 29.8 million English-Arabic sentence pairs. We chose OpenSubtitles because it has parallel sentences in English and because it is full of conversational (first and second person) texts in MSA. We extracted all the pairs that include first or second person pronouns on the English side: *I, me, my, mine, myself*, and *you, your, yours, yourself*. This selection process identified 13.4 million pairs: 2.8 million (21.1%) include first and second person pronouns, 5.7 million (42.5%) include only first person pronouns, and 4.9 million (36.4%) include only second person pronouns. Out of this set, we randomly selected 52,000 English-Arabic pairs to be manually annotated, while maintaining the original first and second person sentences proportions: 10,972 (21.1%) pairs contain first and second person pronouns on the English side, 22,100 (42.5%) pairs contain only first person pronouns on the English side, 18,928 (36.4%) pairs contain only second person pronouns on the English side. To be consistent with APGC v1.0's preprocessing, we ran the Arabic sentences through MADAMIRA (Pasha et al., 2014) to do white-space-and-punctuation tokenization and UTF-8 cleaning.

In addition to the above, we re-annotated all of the 11,240 sentences from APGC v1.0 to include second person references and match our extended guidelines completely. In total, this resulted in 63,240 English-Arabic sentence pairs for the next annotation step. In the final released corpus, we provide labels indicating the origins of all the sentences.

4.2. Corpus Annotation

Four professional linguists (three females and one male), all of whom are native speakers of Arabic, were hired through Ramitechs, a linguistic annotation firm, to complete the task.[3] We provided them with the annotation guidelines available in Appendix A.

Gender Identification First, the annotators were asked to identify the genders of the first and second person references in each sentence, then assign to each sentence a two-letter label, where each letter refers to the gender of the first and second person references, respectively. Each letter in the label can have one of four values: F (feminine), M (masculine), B (invariant/ambiguous), or N (non-existent). Therefore, each sentence will get a label from one of the 16 different label combinations – BB, FB, MB, BF, BM, BN, NB, NN, FN, MN, NF, NM, MM, FM, MF, or FF. Additionally, the annotators were asked to identify the dual and plural gendered references. In case they exist, the sub-label corresponding to the gender of the first or second person reference would get an extra mark: "!" (e.g., BF!, M!B!, etc.).

Gender Reinflection/Rewriting In the case of an F or M sub-label, the annotators were asked to copy the sentence and modify it to obtain the opposite gender forms. The modifications are strictly limited to morphological reinflections and word substitutions as was done in Habash et al. (2019). Therefore, the total num-

[3] `https://www.ramitechs.com/`

English	Arabic	Label	Reinflection Label	Reinflection	
I wanna thank you	أريد أن أشكرك	BB			(a)
I have something to say	لدي شيء لأقوله	BN			(b)
I'm so happy for you	أنا سعيدة من أجلك	FB	MB	أنا سعيد من أجلك	(c)
We were coming to see you	نحن قادمات لرؤيتك	F!B	M!B	نحن قادمون لرؤيتك	(d)
Because I'm your big brother	لأنني أخوك الكبير	MB	FB	لأنني أختك الكبيرة	(e)
We're ready	نحن مستعدون	M!B	F!B	نحن مستعدات	(f)
I know, babe	أعلم ذلك يا عزيزتي	BF	BM	أعلم ذلك يا عزيزي	(g)
I respect you [*plural*]	أنا أحترمكن	BF!	BM!	أنا أحترمكم	(h)
I'm right here dad	أنا هنا يا أبي	BM	BF	أنا هنا يا أمي	(i)
I love you [*plural*] so much	أحبكم كثيرا	BM!	BF!	أحبكن كثيرا	(j)
I'm sorry, you're going to have to leave	آسف يجب أن ترحل	MM	FM	آسفة يجب أن ترحل	(k)
			MF	آسف يجب أن ترحلي	(l)
			FF	آسفة يجب أن ترحلي	(m)
Baby, I'm so scared right now	أنا خائفة للغاية يا عزيزي	FM	MM	أنا خائف للغاية يا عزيزي	(n)
			FF	أنا خائفة للغاية يا عزيزتي	(o)
			MF	أنا خائف للغاية يا عزيزتي	(p)
I'm glad you made it home, mom	أنا سعيد بعودتك يا أماه	MF	FF	أنا سعيدة بعودتك يا أماه	(q)
			MM	أنا سعيد بعودتك يا أبتاه	(r)
			FM	أنا سعيدة بعودتك يا أبتاه	(s)
Don't call me a fool	لا تناديني بالغبية	FF	MF	لا تناديني بالغبي	(t)
			FM	لا تنادني بالغبية	(u)
			MM	لا تنادني بالغبي	(v)

Table 1: Examples from the Arabic Parallel Gender Corpus v2.0 including the original sentence, its gender label, its reinflection gender label, and its reinflection/rewrite to the opposite grammatical gender where appropriate. First person gendered words are in blue and second person gendered words are in red. The "!" in the labels indicate plural forms were used. The two-letter reinflection label specifies gender information of first person (first letter) and second person (second letter). M is Masculine; F is Feminine; B is invariant; and N is non-existent.

ber of words is maintained along with a perfect alignment between each sentence and its parallel opposite gender forms. For example, the sentence in Table 1(c) includes a first person gender reference and is labeled by the annotators as FB, and therefore, the annotators would introduce its gender cognate MB. If the sentence includes both first and second person gender references (MM, FM, MF, or FF), the annotators would then introduce all its possible gender cognates, as in Table 1(k-m) for instance.

In the vast majority of cases, the opposite gender forms of most words end up sharing the same lemma (reinflection), e.g., والد *wAld* 'parent/father [M]' and والدة *wAldħ* 'parent/mother [F]'. However, there are cases where gender-specific words have to be mapped to different lemmas, resulting in a lexical change. For instance, أبي *Âby* 'my dad' and أمي *Âmy* 'my mom' (Table 1(i)), or أخوك *Âxwk* 'your brother' and أختك *Âxtk* 'your sister' (Table 1(e)). While technically these are instances of lexical rewriting and not morphological reinflection, we interchangeably refer to the process covering both phenomena as reinflection or rewriting.

Furthermore, the annotators were instructed to avoid any heterocentric assumptions during the annotation. For example, the sentence أنت زوجي *Ânt zwjy* 'you are my husband' is labeled as BM (ambiguous first person, masculine second person) and not FM (feminine first person, masculine second person) . The annotators were also instructed to treat all proper names as gender-ambiguous (B), even when they have strong gender-specific associations, and as such are not rewritten. Finally, the annotators were asked to flag bad translations and malformed sentences.

At the end of the annotation process, we did a quality check on the dataset and fixed some of the annotation errors manually. Most of these errors were either due to malformed Arabic subtitles or misalignment between the parallel sentences.

English	Arabic	Label	
I wanna thank you	أريد أن أشكرك B B B	BB	(a)
I have something to say	لدي شيء لأقوله B B B	BN	(b)
I'm so happy for you	أنا سعيدة من أجلك B 1F B B	FB	(c)
	أنا سعيد من أجلك B 1M B B	MB	(d)
I know, babe	أعلم ذلك يا عزيزتي B B B 2F	BF	(e)
	أعلم ذلك يا عزيزي B B B 2M	BM	(f)
Baby, I'm so scared right now	أنا خائفة للغاية يا عزيزي B 1F B B 2M	FM	(g)
	أنا خائف للغاية يا عزيزي B 1M B B 2M	MM	(h)
	أنا خائفة للغاية يا عزيزتي B 1F B B 2F	FF	(i)
	أنا خائف للغاية يا عزيزتي B 1M B B 2F	MF	(j)

Table 2: Examples of word-level gender annotation. First person gendered words are in blue and second person gendered words are in red.

4.3. Automatic Word-Level Annotations

Since the annotators were only allowed to perform grammatical inflections and word substitutions, all sentences and their parallels are perfectly aligned at the word level. This allowed us to obtain word-level gender annotations automatically as a byproduct. To do this, we look at the original sentence and all of its parallel forms. If the word is the same across all the parallel versions of a sentence, then we label it as B. Otherwise, we assign the word a label based on its sentence-level gender label. For example, in Table 2(g-j), the word أنا *Âna* 'I' is the same across all four parallel versions of the sentence and thus labeled as B. In contrast, the words خائفة *xaŷfħ* 'scared [F]' and خائف *xaŷf* 'scared [M]' change across the parallel versions. By looking at the sentence-level labels of the four parallel forms, we can deduce that the word خائفة *xaŷfħ* is first-person feminine and label it 1F, and that the word خائف *xaŷf* is first-person masculine and labeled it 1M. Similarly, we determine that the words عزيزي *ςzyzy* 'baby/dear [M]' and عزيزتي *ςzyzty* 'baby/dear [F]' are second-person masculine and second-person feminine and label them 2M and 2F, respectively. All words belonging to sentences that do not have any gender cognates (BB, BN, NB, etc. cases) as in Table 2(a & b) are labeled B. Therefore, each word can have one of following possible labels: B, 1F, 1M, 2F, 2M. We mark the dual/plural words by adding "!" to their corresponding labels.

5. Corpus Overview and Statistics

5.1. The Original Corpus

After the annotation, 8.2% of the sentences (5,205) were eliminated due to malformed Arabic and annotation errors. This resulted in 58,035 sentences (423,254 words), constituting our ORIGINAL CORPUS. We created a condensed version of the annotations for this corpus by mapping the N (non-existent) sub-labels to B (invariant/ambiguous) and removing the dual/plural marks ("!") from the labels across all the sentences.

Corpus Statistics Table 3(a) includes the statistics about the ORIGINAL CORPUS. Out of all sentences, 36,980 (63.7%) are labeled as BB. There are 17,374 (30%) sentences that include only second-person gendered references (BF and BM). This is five times more than sentences with only first-person gendered references (FB and MB), which accounts for 5.3% (3,063 sentences) of all sentences. Moreover, the number of sentences including first or second person masculine references is more than the ones including feminine references (12,164 BM vs 5,210 BF, and 1,940 MB vs 1,123 FB). There are 618 (1.1%) sentences that have both first and second gendered references. All of the sentences which have first or second (or both) person gendered references are rewritten to introduce their opposite gender forms. This resulted in 21,055 manually added sentences (162,055 words). The word-level statistics of our ORIGINAL CORPUS are shown in Table 4(a). Among the newly added sentences, about 17%

(a)

Sentences		Original Corpus Label	Reinflection Label		
36,980	63.7%	BB			
1,123	1.9%	FB	MB		
1,940	3.3%	MB	FB		
5,210	9%	BF	BM		
12,164	21%	BM	BF		
68	0.1%	FF	MF	FM	MM
135	0.2%	FM	MM	FF	MF
117	0.2%	MF	FF	MM	FM
298	0.5%	MM	FM	MF	FF
58,035					

(b)

Balanced Corpus Input	Target$_{MM}$	Target$_{FM}$	Target$_{MF}$	Target$_{FF}$	Sentences	
BB	BB	BB	BB	BB	36,980	46%
FB	MB	FB	MB	FB	3,063	3.8%
MB	MB	FB	MB	FB	3,063	3.8%
BF	BM	BM	BF	BF	17,374	21.6%
BM	BM	BM	BF	BF	17,374	21.6%
FF	MM	FM	MF	FF	618	0.8%
FM	MM	FM	MF	FF	618	0.8%
MF	MM	FM	MF	FF	618	0.8%
MM	MM	FM	MF	FF	618	0.8%
					80,326	

Table 3: Sentence-level statistics of the original corpus (a) and the balanced corpus (b) with its five versions.

(a)

Words		Original Corpus Label	Reinflection Label
395,658	93.5%	B	
1,511	0.4%	1F	1M
2,716	0.6%	1M	1F
6,844	1.6%	2F	2M
16,525	3.9%	2M	2F
423,254			

(b)

Balanced Corpus Input	Target$_{MM}$	Target$_{FM}$	Target$_{MF}$	Target$_{FF}$	Words	
B	B	B	B	B	538,733	90.3%
1F	1M	1F	1M	1F	4,923	0.8%
1M	1M	1F	1M	1F	4,923	0.8%
2F	2M	2M	2F	2F	24,110	4%
2M	2M	2M	2F	2F	24,110	4%
					596,799	

Table 4: Word-level statistics of the original corpus (a) and the balanced corpus (b) with its five versions.

(27,596) of the words are gender-specific, constituting around 6.5% of all the words in the corpus.

Table 7(a) and Table 8(a) in Appendix B present statistics on the non-condensed annotations of our ORIGINAL CORPUS at the sentence and word levels, respectively.

Morphological Reinflection vs Lexical Rewriting To quantify the proportions of the morphological reinflections and lexical changes introduced as part of the manual annotation process (§4.2), we analyzed the gender-specific words across all parallel sentences using the CALIMA$_{Star}$ Arabic morphological analyzer (Taji et al., 2018) included in the CAMeL Tools toolkit (Obeid et al., 2020). We consider the manually introduced gender cognate of a specific word to be its reinflection, if both words share at least one lemma. If no lemmas are shared, then the gender cognate is a result of a lexical change. If the word or its gender cognate does not get recognized by the morphological analyzer, we look at them manually. Out of the 27,596 newly introduced gender specific words, 26,728 (96.9%) resulted from morphological reinflection, whereas 868 words (3.1%) resulted from lexical rewriting.

5.2. The Balanced Corpus

Similarly to Habash et al. (2019), to ensure equal gender representation in our dataset, we force balance the corpus by adding the manually rewritten sentences to the ORIGINAL CORPUS and using their original forms as their rewritten forms. This constitutes our BALANCED CORPUS.

Corpus Statistics The sentence-level statistics of the BALANCED CORPUS are presented in Table 3(b). This corpus has 80,326 sentences in total. Out of all sentences, 46% (36,980) are marked as BB, whereas sentences with gendered references constituted 54% (43,346 sentences). We introduce five versions of the BALANCED CORPUS: Input, Target$_{MM}$, Target$_{FM}$, Target$_{MF}$, and Target$_{FF}$. The balanced Input corpus, includes all the sentences from the ORIGINAL CORPUS in addition to their rewritten forms. The Target$_{MM}$ corpus is the masculine-only corpus and it includes sentences that are either invariant/ambiguous or have a first or second person (or both) masculine references. Therefore, it only contains BB, MB, BM, and MM sentences. The Target$_{MF}$ corpus is the masculine-feminine corpus and it contains sentences that are either invariant/ambiguous or have first person masculine references, second person feminine references, or first person masculine and second person feminine references (i.e., BB, MB, BF, and MF sentences). The Target$_{FM}$ corpus is the feminine-masculine corpus and it contains BB, FB, BM, and FM sentences. Finally, the Target$_{FF}$ corpus is the feminine-only corpus and it contains BB, FB, BF, and FF sentences. All five corpora have the same number of sentences, words, and gendered-specific words. The word-level statistics of the BALANCED CORPUS are shown in Table 4(b). Table 7(b) and Table 8(b) in Appendix B present statistics on the non-condensed annotations of the BALANCED CORPUS at the sentence and word levels, respectively.

Corpus Splits To aid reproducibility when using APGC v2.0 for various research experiments, we provide train, development, and test splits for all five balanced corpora. Following Habash et al. (2019), all five corpora were divided randomly as follows: training (TRAIN: 70% or 57,603 sentences), development (DEV: 10% or 6,647 sentences) and testing (TEST: 20% or 16,076 sentences). We made sure that the splits are balanced and all parallel versions of the sentences are in the same split.

6. Revisiting the Motivation: Quantifying Bias in Gender-Unaware Machine Translation

The efforts to develop APGC v1.0 and APGC v2.0 were motivated by the observation of common gender bias in gender-unaware NLP systems targeting morphologically rich languages, specifically Arabic in our case. In this section, we revisit this motivation and use our newly created corpus to quantify and detect gender bias in machine translation. We translated the English side of the Input balanced corpus to Arabic using the Google Translate API.[4] We chose to use Google Translate because of its popularity, but these experiments can be easily done on any machine translation output.[5] We include Google Translate's outputs in the release of our corpus to encourage research and development on corrective post-editing.

We evaluate Google Translate's output against all four balanced target corpora (i.e., Target$_{MM}$, Target$_{FM}$, Target$_{MF}$, Target$_{FF}$) separately as well as in a multi-reference setting. In Tables 5 and 6, we present the results in terms of BLEU (Papineni et al., 2002) using the latest version of SacreBLEU (Post, 2018). All the results are reported in an orthographically normalized space for Alif, Ya, and Ta-Marbuta (Habash, 2010). The results are organized around different subsets of the BALANCED CORPUS to allow us to determine the effect of different gender-specificity factors in Arabic and English on the results.

6.1. Overall Results

To start off, looking at the BLEU scores of all sentences (ALL) in Table 5, we notice that the score against the Target$_{MM}$ corpus is higher than the score against Target$_{FF}$ (by 2.5 BLEU absolute). Moreover, we notice that the multi-reference BLEU score is a little higher than the score of Target$_{MM}$ (0.1 BLEU). This indicates that scores from the evaluation against Target$_{FM}$, Target$_{MF}$, and Target$_{FF}$ are contributing to the overall increase in the multi-reference evaluation but not that much. From these basic results, we observe that every time an M participant is switched to F, the BLEU scores drop. This strongly suggests that the machine translation output is biased towards masculine grammatical gender preferences.

6.2. Results on Arabic Gender Specific Subsets

The remainder of Table 5 presents the results organized by Arabic gender-specificity factors. The Arabic invariant/ambiguous (BB) sentences have the same BLEU scores in all conditions because they do not vary across the different Target references. When we compare the BLEU scores of BB sentences with gender specific sentences (ALL - BB), we notice a 0.6 drop in the multi-reference evaluation. This indicates that sentences with gender specific words are harder to translate for Google Translate than gender invariant/ambiguous sentences.

The drop in BLEU scores from BB to ALL - BB for Target$_{FF}$ is 5.4 BLEU, which is six times the corresponding drop for Target$_{MM}$. Also, the difference between the Target$_{MM}$ and Target$_{FF}$ BLEU scores for (ALL - BB) is almost double the difference for ALL (4.5 vs 2.5). By grouping the Arabic gender-marked sentences (ALL - BB) based on the variation of the first and second person gendered references (i.e., BM, BF, MB, FB, MM, FM, MF, and FF), we again observe that every time an M participant is switched to F, the BLEU scores drop. In the most extreme case of gender specific references in both first and second person (last row in Table 5), the difference between Target$_{MM}$ and Target$_{FF}$ in BLEU scores is 6.2. Therefore, we can deduce that Google Translate's Arabic outputs are biased against feminine target users compared to masculine users.

6.3. Results on English Gender Specific Subsets

While our evaluation setup assumes that Google's Arabic translations could have come from translating sentences in any language into Arabic, we acknowledge that some of the English sentences may be gender-specific and the bias we are observing might be caused by such sentences. In this section, we delve into studying the bias in the Arabic translations of gender specific English sentences.

Gender in English is usually expressed referentially through third person singular pronouns (he and she) or lexically using gender specific nouns (e.g., mother, son, etc.) (Cao and Daumé III, 2020). As our parallel gender corpus was only annotated for first and second person gendered Arabic references, we only focus on English sentences that contain gender specific nouns. We focus on the OpenSubtitles 2018 English sentences corresponding to the gender specific (ALL - BB) Arabic sentences in our BALANCED CORPUS (43,346 sentences). We tokenized each English sentence and obtained the part-of-speech tags of its tokens using spaCy (SpaCy, 2017). Out of 43,346 sentences, 24,350 contained at least one noun. We annotated these

[4] `https://cloud.google.com/translate` on September 15th, 2021

[5] It should be noted with admiration that the Google Translate team has done a lot of work on the front of fighting gender bias, e.g., generating multiple gendered translation for some language pairs (Johnson, 2020). To date, Arabic is not one of these languages.

Selected Sentences$_{ar}$	Count	Target$_{MM}$	Target$_{FM}$	Target$_{MF}$	Target$_{FF}$	Multi-Reference
ALL	80,326	13.5	13.1	11.4	11.0	13.6
BB	36,980	14.0	14.0	14.0	14.0	14.0
ALL - BB	43,346	13.1	12.4	9.3	8.6	13.4
BM BF	34,748	13.1	13.1	8.6	8.6	13.3
MB FB	6,126	12.9	9.6	12.9	9.6	13.6
MM FM MF FF	2,472	12.9	9.5	9.5	6.7	13.5

Table 5: BLEU results (all Alif/Ya/Ta-Marbuta normalized) of the English-Arabic Google Translate output for the balanced input corpus against the four balanced target corpora.

Selected Sentences$_{ar}$	Selected Sentences$_{en}$	Count	Target$_{MM}$	Target$_{FM}$	Target$_{MF}$	Target$_{FF}$	Multi-Reference
ALL - BB	ALL$_{en}$	43,346	13.1	12.4	9.3	8.6	13.4
ALL - BB	B$_{en}$	39,484	13.1	12.4	9.2	8.5	13.3
ALL - BB	M$_{en}$	2,606	14.1	13.0	9.6	8.5	14.2
ALL - BB	F$_{en}$	1,256	10.8	11.1	10.0	10.4	13.2

Table 6: BLEU results (all Alif/Ya/Ta-Marbuta normalized) of the English-Arabic Google Translate output for the English sentences corresponding to gender specific (ALL - BB) Arabic sentences as in Table 5.

English nouns' lemmas (4,138 unique lemmas) manually as either B$_{en}$ (ambiguous, e.g., *teacher, scientist, prostitute*), M$_{en}$ (masculine, e.g., *father, policeman, brother*), or F$_{en}$ (feminine, e.g., *mother, queen, waitress*). Out of the 4,138 lemmas, 97.4% (4,032) were labeled as B$_{en}$, 1.4% (60) were labeled as M$_{en}$, and 1.1% (46) were labeled as F$_{en}$. The list of all the English gender-specific noun lemmas is in Appendix C.

Using the annotated English nouns, we labeled the English sentences based on the gender of their nouns as follows: if a sentence has only F$_{en}$ or (B$_{en}$ and F$_{en}$) nouns, it is labeled as F$_{en}$. If a sentence has only M$_{en}$ or (B$_{en}$ and M$_{en}$) nouns, it is labeled as M$_{en}$. All other sentences are labeled as B$_{en}$. This is clearly a rough approach since we do not label the sentences by the first and second person gender reference. This resulted in 39,484 (91.1%) B$_{en}$ sentences, 2,606 (6%) M$_{en}$ sentences, and 1,256 (2.9%) F$_{en}$ sentences. We include the English sentence labels in our released corpus to support further research on this topic. The results on these subsets are in Table 6, which presents the BLEU scores of Google's Arabic translations against our four balanced Target corpora (including multi-reference).

Examining the BLEU scores of the B$_{en}$ sentences, we notice that they are almost identical to the BLEU scores of the ALL$_{en}$ sentences (i.e. ALL - BB), which is expected as they are the majority subset. Moreover, although all of the English sentences in this subset did not have any gender specific nouns, the BLEU score of Google Translate's output against the Target$_{MM}$ sentences is significantly higher than the score against the Target$_{FF}$ sentences (4.6 BLEU). This highlights the bias of Google Translate towards masculine users compared to their feminine counterparts when English sentences are invariant or ambiguous gender-wise.

When we consider the BLEU scores of the M$_{en}$ sentences, we observe an expected increase in the scores across the all target corpora with M targets. As for the F$_{en}$ sentences, we also observe an expected increase for Target$_{FF}$ paired with a decrease for Target$_{MM}$. But even here, the Target$_{MM}$ BLEU score is still slightly higher than the Target$_{FF}$ BLEU score.

7. Conclusion and Future Work

We presented APGC v2.0, a new Arabic parallel corpus for gender identification and rewriting in contexts involving one or two target users (I and/or You) with independent grammatical gender preferences. We provided a detailed description of the selection and annotation process we followed to create our corpus. Furthermore, we showed that our corpus can be used to study and quantify the degree and type of gender biases and stereotypes that are embedded in and amplified by one of state-of-the-art commercial machine translation systems.

In future work, we plan to extend our corpus to other dialectal varieties and languages. By building our corpus and making it publicly available, we hope to encourage research on gender identification, controlled generation, and post-editing rewrite systems that could be used to personalize NLP applications based on their end users' preferences.

Acknowledgements

We thank Ramitechs for their help in the annotation process. We would like to thank Go Inoue and Salam Khalifa for the helpful and insightful conversations.

8. Bibliographical References

Alhafni, B., Habash, N., and Bouamor, H. (2020). Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Alkuhlani, S. and Habash, N. (2011). A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, Oregon, USA, June.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (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, July. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, et al., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bordia, S. and Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June.

Cao, Y. T. and Daumé III, H. (2020). Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July. Association for Computational Linguistics.

Cercas Curry, A., Robertson, J., and Rieser, V. (2020). Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Ciora, C., Iren, N., and Alikhani, M. (2021). Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK, August. Association for Computational Linguistics.

Costa-jussà, M. R. and de Jorge, A. (2020). Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., and Weston, J. (2020a). Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online, November. Association for Computational Linguistics.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020b). Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November. Association for Computational Linguistics.

Elaraby, M., Tawfik, A. Y., Khaled, M., Hassan, H., and Osama, A. (2018). Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Escudé Font, J. and Costa-jussà, M. R. (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, August. Association for Computational Linguistics.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Gonen, H. and Webster, K. (2020). Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online, November. Association for Computational Linguistics.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In A. van den Bosch et al., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Habash, N., Bouamor, H., and Chung, C. (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, August.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Hall Maudslay, R., Gonen, H., Cotterell, R., and Teufel, S. (2019). It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China, November.

Johnson, M. (2020). A scalable approach to reducing gender bias in google translate. Google AI Blog.

Katsarou, S., Rodríguez-Gálvez, B., and Shanahan, J. (2022). Measuring gender bias in contextualized embeddings. In *Computer Sciences and Mathematics Forum*, volume 3, page 3. MDPI.

Lauscher, A., Takieddin, R., Ponzetto, S. P., and Glavaš, G. (2020). AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Liu, H., Dacon, J., Fan, W., Liu, H., Liu, Z., and Tang, J. (2020a). Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Liu, H., Wang, W., Wang, Y., Liu, H., Liu, Z., and Tang, J. (2020b). Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online, November. Association for Computational Linguistics.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2018). Gender bias in neural natural language processing.

Maass, A. and Arcuri, L. (1996). Language and stereotyping. *Stereotypes and stereotyping*, pages 193–226.

Manzini, T., Yao Chong, L., Black, A. W., and Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Mehrabi, N., Gowda, T., Morstatter, F., Peng, N., and Galstyan, A. (2019). Man is to person as woman is to location: Measuring gender bias in named entity recognition.

Menegatti, M. and Rubini, M. (2017). Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. Oxford University Press.

Moryossef, A., Aharoni, R., and Goldberg, Y. (2019). Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy, August.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October.

Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., and Wintner, S. (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, April.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana, June.

Saunders, D. and Byrne, B. (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, July. Association for Computational Linguistics.

Saunders, D., Sallis, R., and Byrne, B. (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), December. Association for Computational Linguistics.

Saunders, D., Sallis, R., and Byrne, B. (2021). First the worst: Finding better gender translations during beam search.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 08.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November.

Sheng, E., Arnold, J., Yu, Z., Chang, K.-W., and Peng, N. (2021a). Revealing persona biases in dialogue systems.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2021b). "nice try, kiddo": Investigating ad hominems in dialogue responses.

SpaCy. (2017). spacy - industrial-strength natural language processing in python.

Stafanovičs, A., Bergmanis, T., and Pinnis, M. (2020). Mitigating gender bias in machine translation with target gender annotations.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (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, July.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Sun, T., Webster, K., Shah, A., Wang, W. Y., and Johnson, M. (2021). They, them, theirs: Rewriting with gender-neutral english.

Taji, D., Khalifa, S., Obeid, O., Eryani, F., and Habash, N. (2018). An Arabic Morphological Analyzer and Generator with Copious Features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON)*, pages 140–150.

Vanmassenhove, E. and Monti, J. (2021). gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online, August. Association for Computational Linguistics.

Vanmassenhove, E., Hardmeier, C., and Way, A. (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, October-November.

Vanmassenhove, E., Emmery, C., and Shterionov, D. (2021). NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). 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, June.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018b). Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W., and Hassan Awadallah, A. (2020). Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July. Association for Computational Linguistics.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (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, July.

9. Language Resource References

Habash, Nizar and Bouamor, Houda and Chung, Christine. (2019). *Automatic Gender Identification and Reinflection in Arabic*.

Lison, Pierre and Tiedemann, Jörg. (2016). *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles*. European Language Resources Association (ELRA).

A. Annotation Guidelines

Gender Identification and Rewriting

Input:
The input has three columns: **SentenceID**, **English** and **Arabic**.
These columns should not change throughout the annotation process.

Output:
The output has two columns:
SentenceType, and **RewrittenSentence**

The value of **SentenceType** should be a two-letter code, where the first code refers to the status of the 1st person reference in the sentence, and the second code refers to the status of the 2nd person reference in the sentence.
The status should be one of four values:

- M: Masculine
- F: Feminine
- B: Ambiguous
- N: does not exist

Here are all the possible combinations with examples:

Code	Meaning	Example
M-M	The sentence has a masculine 1st person reference and a masculine 2nd person reference	أنا صبور مثلك يا صديقي
M-B	The sentence has a masculine 1st person reference and an ambiguous 2nd person reference	أنا صبور مثلك
F-B	The sentence has a feminine 1st person reference and an ambiguous 2nd person reference	مرحبا، أنا جارتك الجديدة
B-M	The sentence has an ambiguous 1st person reference and a masculine 2nd person reference	أنا من تنتظره
B-F	The sentence has an ambiguous 1st person reference and a feminine 2nd person reference	أحبك يا صديقتي
B-B	The sentence has ambiguous 1st person and 2nd person references	أحبك انت
M-N	The sentence has a masculine 1st person reference but NO 2nd person reference	أنا قوي
F-N	The sentence has a feminine 1st person reference but NO 2nd person reference	أنا سعيدة جدا
N-M	The sentence has a masculine 2nd person reference but NO 1st person reference	! قف
N-F	The sentence has a feminine 2nd person reference but NO 1st person reference	انت تكتبين
B-N	The sentence has an ambiguous 1st person reference but NO 2nd person reference	حسناً. أنا هنا
N-B	The sentence has an ambiguous 2nd person reference but NO 1st person reference	الجو ممتع في بلدك
N-N	The sentence has no reference for 1st nor 2nd person	الجو ممتع

The **RewrittenSentence** column should have the Arabic sentence after converting the masculine references to feminine and the feminine references to masculine.
Here are the changes the sentences above should receive (the comment column does not exist in the annotations but it is just to let you understand the task).

Code	Arabic	RewrittenSentence	Comment
M-M	أنا صبور مثلك يا صديقي	أنا صبورة مثلك يا صديقي	1st person changes to go from M-M to F-M
M-M	أنا صبور مثلك يا صديقي	أنا صبور مثلك يا صديقتي	2nd person changes to go from M-M to M-F
F-M	أنا صبورة مثلك يا صديقي	أنا صبورة مثلك يا صديقتي	2nd person changes to go from F-M to F-F
M-B	أنا صبور مثلك	أنا صبورة مثلك	
F-B	مرحبا، أنا جارتك الجديدة	مرحبا، أنا جارك الجديدة	
B-M	أنا من تنتظره	أنا من تنتظريه	
B-F	أحبك يا صديقتي	أحبك يا صديقي	
B-B	أحبك انت		No change is needed. Leave **RewrittenSentence** empty
M-N	أنا قوي	أنا قوية	
F-N	أنا سعيدة جدا	أنا سعيد جدا	
N-M	! قف	! قفي	
N-F	انت تكتبين	انت تكتب	
B-N	حسناً. أنا هنا		No change is needed. Leave **RewrittenSentence** empty
N-B	الجو ممتع في بلدك		No change is needed. Leave **RewrittenSentence** empty
N-N	الجو ممتع		No change is needed. Leave **RewrittenSentence** empty

VERY IMPORTANT

What should we do with 3rd person references?

Nothing. This task is only about 1st and 2nd person references.
For example: هذا هو الرجل الشجاع is an N-N case.

How to handle dual and plural cases?

We do not expect the dual and plural to happen much, but when they do add a "!" to the label (e.g., M!, F! and B!). For example:
لقد سمعتكن تتكلمن عنها should be tagged as B-F!
لقد سمعتك تتكلمين عن أخلاقنا should be tagged as B!-F

How to handle proper nouns?

Handle proper nouns as ambiguous. Do not make any assumptions about proper nouns.
For examples: مرحبا، انا ماري should be tagged as B-N
However, مرحبا، انا ماري جارتك الجديدة should be tagged as F-B because of the word الجديدة

What do to with writing mistakes?

Do NOT correct any mistakes of any kind.
Do NOT correct even hamzas, Ta' marboota, Alif maqsoura or misplaced punctuation.
Do not add any diacritization. Notice that أحبك انت is a B-B case as diacritization is absent.
If the Arabic sentence has major mistakes (ex: نعم ، أوه ، طفلرضيع ، أعرف), just make the sentence type B-S and leave **RewrittenSentence** empty.

What if the English column does not match the translation in the Arabic column?

Ignore the English column, do not worry about it. Also, do not use the English column to judge the gender or any other information.
Just use the English column if you need to understand an Arabic sentence that is not clear for some reason.
For example: We know that the word ذهبت in ذهبت إلى هناك refers to 1S not 2MS because the English translation is "I went" not "You went".

Any hard cases we can encounter?

Yes, but this is infrequent.
For example: Sometimes you need to change the lemma itself for the gender conversion. Examples:

انا رجل → انا امرأة
أحبك يا أمي →أحبك يا أبي
نعم أنا هو → نعم أنا هي (note that although هي is 3rd person, but we need to change it in order to adjust the 1st person reference).

However, if you encounter any HARD case that you cannot figure out (e.g., انا حامل), just write the word HARD in the **RewrittenSentence** column.

B. Fine-Grained Corpus Statistics

(a)

Sentences		Original Corpus			
		Label	**Reinflection Label**		
6,156	10.6%	BB			
601	1.0%	B!B			
148	0.3%	BB!			
20	0.03%	B!B!			
22,379	38.6%	BN			
457	0.8%	B!N			
5,329	9.2%	NB			
189	0.3%	NB!			
1,701	2.9%	NN			
254	0.4%	FB	MB		
403	0.7%	MB	FB		
2	0.003%	FB!	MB!		
7	0.01 %	MB!	FB!		
0	0%	F!B	M!B		
15	0.03%	M!B	F!B		
867	1.5%	FN	MN		
1,510	2.6%	MN	FN		
0	0%	F!N	M!N		
5	0.01%	M!N	F!N		
2,562	4.4%	BF	BM		
5,171	8.9%	BM	BF		
199	0.3%	B!F	B!M		
433	0.7%	B!M	B!F		
26	0.04%	BF!	BM!		
851	1.5%	BM!	BF!		
1	0.002%	B!F!	B!M!		
125	0.2%	B!M!	B!F!		
2,391	4.1%	NF	NM		
4,904	8.5%	NM	NF		
31	0.1%	NF!	NM!		
680	1.2%	NM!	NF!		
64	0.1%	FF	MF	FM	MM
110	0.2%	FM	MM	FF	MF
115	0.2%	MF	FF	MM	FM
242	0.4%	MM	FM	MF	FF
0	0%	F!F	M!F	F!M	M!M
0	0%	F!M	M!M	F!F	M!F
2	0.003%	M!F	F!F	M!M	F!M
10	0.02%	M!M	F!M	M!F	F!F
4	0.01%	FF!	MF!	FM!	MM!
25	0.04%	FM!	MM!	FF!	MF!
0	0%	MF!	FF!	MM!	FM!
43	0.1%	MM!	FM!	MF!	FF!
0	0%	F!F!	M!F!	F!M!	M!M!
0	0%	F!M!	M!M!	F!F!	M!F!
0	0%	M!F!	F!F!	M!M!	F!M!
3	0.01%	M!M!	F!M!	M!F!	F!F!
58,035					

(b)

Balanced Corpus						
Input	**Target$_{MM}$**	**Target$_{FM}$**	**Target$_{MF}$**	**Target$_{FF}$**	**Sentences**	
BB	BB	BB	BB	BB	6,156	7.7%
B!B	B!B	B!B	B!B	B!B	601	0.7%
BB!	BB!	BB!	BB!	BB!	148	0.2%
B!B!	B!B!	B!B!	B!B!	B!B!	20	0.02%
BN	BN	BN	BN	BN	22,379	27.9%
B!N	B!N	B!N	B!N	B!N	457	0.6%
NB	NB	NB	NB	NB	5,329	6.6%
NB!	NB!	NB!	NB!	NB!	189	0.2%
NN	NN	NN	NN	NN	1,701	2.1%
FB	MB	FB	MB	FB	657	0.8%
MB	MB	FB	MB	FB	657	0.8%
FB!	MB!	FB!	MB!	FB!	9	0.01%
MB!	MB!	FB!	MB!	FB!	9	0.01%
F!B	M!B	F!B	M!B	F!B	15	0.02%
M!B	M!B	F!B	M!B	F!B	15	0.02%
FN	MN	FN	MN	FN	2,377	3.0%
MN	MN	FN	MN	FN	2,377	3.0%
F!N	M!N	F!N	M!N	F!N	5	0.01%
M!N	M!N	F!N	M!N	F!N	5	0.01%
BF	BM	BM	BF	BF	7,733	9.6%
BM	BM	BM	BF	BF	7,733	9.6%
B!F	B!M	B!M	B!F	B!F	632	0.8%
B!M	B!M	B!M	B!F	B!F	632	0.8%
BF!	BM!	BM!	BF!	BF!	877	1.1%
BM!	BM!	BM!	BF!	BF!	877	1.1%
B!F!	B!M!	B!M!	B!F!	B!F!	126	0.2%
B!M!	B!M!	B!M!	B!F!	B!F!	126	0.2%
NF	NM	NM	NF	NF	7,295	9.1%
NM	NM	NM	NF	NF	7,295	9.1%
NF!	NM!	NM!	NF!	NF!	711	0.9%
NM!	NM!	NM!	NF!	NF!	711	0.9%
FF	MM	FM	MF	FF	531	0.7%
FM	MM	FM	MF	FF	531	0.7%
MF	MM	FM	MF	FF	531	0.7%
MM	MM	FM	MF	FF	531	0.7%
F!F	M!M	F!M	M!F	F!F	12	0.01%
F!M	M!M	F!M	M!F	F!F	12	0.01%
M!F	M!M	F!M	M!F	F!F	12	0.01%
M!M	M!M	F!M	M!F	F!F	12	0.01%
F!F	MM!	FM!	MF!	FF!	72	0.1%
FM!	MM!	FM!	MF!	FF!	72	0.1%
MF!	MM!	FM!	MF!	FF!	72	0.1%
MM!	MM!	FM!	MF!	FF!	72	0.1%
F!F!	M!M!	F!M!	M!F!	F!F!	3	0.004%
F!M!	M!M!	F!M!	M!F!	F!F!	3	0.004%
M!F!	M!M!	F!M!	M!F!	F!F!	3	0.004%
M!M!	M!M!	F!M!	M!F!	F!F!	3	0.004%
					80,326	

Table 7: Fine-grained sentence-level statistics of the original corpus (a) and the balanced corpus (b) with its five versions.

(a)

Words		Original Corpus	
		Label	Reinflection Label
395,658	93.5%	B	
1,511	0.4%	1F	1M
2,678	0.6%	1M	1F
0	0%	1F!	1M!
38	0.01%	1M!	1F!
6,756	1.6%	2F	2M
14,004	3.3%	2M	2F
88	0.02%	2F!	2M!
2,521	0.6%	2M!	2F!
423,254			

(b)

Balanced Corpus					Words	
Input	Target$_{MM}$	Target$_{FM}$	Target$_{MF}$	Target$_{FF}$		
B	B	B	B	B	538,733	90.3%
1F	1M	1F	1M	1F	4,868	0.8%
1M	1M	1F	1M	1F	4,868	0.8%
1F!	1M!	1F!	1M!	1F!	55	0.01%
1M!	1M!	1F!	1M!	1F!	55	0.01%
2F	2M	2M	2F	2F	21,406	3.6%
2M	2M	2M	2F	2F	21,406	3.6%
2F!	2M!	2M!	2F!	2F!	2,704	0.5%
2M!	2M!	2M!	2F!	2F!	2,704	0.5%
					596,799	

Table 8: Fine-grained word-level statistics of the original corpus (a) and the balanced corpus (b) with its five versions.

C. Gender Specific English Nouns

Below is the list of English gender-specific noun lemmas annotated in our dataset. We discard misspelled lemmas in this list.

Masculine Nouns actor, boy, boyfriend, bro, brother, businessman, chairman, chap, chauffeur, congressman, cornerman, cowboy, dad, daddy, doorman, emperor, father, footman, foreman, freedman, gentleman, godfather, grandad, granddaddy, grandfather, grandpa, grandson, guy, hangman, henchman, highwayman, homeboy, imam, landlord, lawman, lord, lordship, male, man, mate, milkman, nephew, oldman, papa, policeman, pop, praetor, priest, prince, prophet, salesman, samurai, sir, son, stepbrother, uncle, waiter, wizard.

Feminine Nouns actress, ballerina, bride, businesswoman, cow, daughter, gal, girl, girlfriend, girlie, goddess, godmother, granddaughter, grandma, grandmother, housewife, lady, lesbian, ma'am, madam, mama, mom, momma, mommy, mother, mum, nana, nanny, niece, patroness, prima, queen, schoolgirl, sister, sorceress, stepmom, suffragette, supergirl, tsarina, waitress, widow, wife, wingwoman, witch, woman.