# When the Language Must Choose: A Position on Gender Bias Evaluation for Arabic NLP

**Wajdi Zaghouani**
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

## Abstract

Standard English-centric gender bias evaluation provides incomplete coverage of Arabic because Arabic grammar obligates gender choices on verbs, adjectives, pronouns, and nouns. Translating occupation-noun tests or English rewriting benchmarks to Arabic ignores the grammatical forcing that is central to the problem. We argue evaluation for grammatical gender languages requires language-specific extensions starting from the morphology that forces a choice, naming who is affected by the default form chosen, and then assessing whether a system-internal fix helped. Drawing on Arabic Parallel Gender Corpus work, WinoMT, and gender rewriting tools, we offer a protocol for transparent evaluation: name the linguistic condition, including construction type, grammatical person, and number; name the default (the masculine-leaning BLEU gaps); separate linguistic gender from social gender; and do not flatten intersectionality. Arabic evaluation work should be built from Arabic grammar, not adapted from English methods.

## 1 Introduction

Gender bias in NLP is most developed as English-centered research. Savoldi et al. (2025) review 133 machine-translation gender-bias studies and flag the pervasiveness of English-centric evaluation setups and the lack of simple technical fixes across language pairs. In English, gender on most words is optional and external: pronouns like *they* or *one* can obscure or defer gender attribution to social context. Stanovsky et al. (2019) introduced WinoMT, which evaluates whether machine translation preserves the gender indicated by English pronouns in ambiguous syntactic contexts—cases where the referent gender is supplied by English morphology, not occupational stereotypes alone. This includes pro- and anti-stereotypical sentence pairs to test whether systems default to or resist stereotype-matching. Nangia et al. (2020) introduced CrowS-Pairs, which evaluates preferences between sentence pairs expressing differing degrees of stereotyping across multiple demographic categories. Both are influential benchmarks for English-centered bias evaluation.

This design choice works for English. It provides incomplete coverage for languages with obligatory gender morphology. Arabic requires gender agreement in many constructions referring to people, including predicate adjectives and certain second-person forms. When the input does not specify the relevant grammatical gender preference, producing these constructions requires a choice. Evaluation should distinguish this underspecification from failures to preserve gender information explicitly available in the input.

**What is new here.** The Arabic and broader NLP literature already discusses obligatory gender morphology, intersectionality, and post-editing effort (Blodgett et al., 2020; Savoldi et al., 2022). This paper's contribution is operationalization: translating those principles into an explicit evaluation checklist that (1) names the morphological cell being tested, (2) separates linguistic from social gender, and (3) distinguishes between metric scores and fairness verdicts. The Arabic Parallel Gender Corpus exists precisely because Arabic developers understood these requirements. We formalize what those tools demonstrate, providing a standardized approach for reporting gender evaluation results.

This position paper argues three things. First, English-centric benchmarks provide incomplete coverage of gender phenomena in grammatical gender languages (languages requiring obligatory gender agreement on predicates or pronouns) because they do not account for morphological choices the language requires. Second, Arabic NLP already has better instruments: the Arabic Parallel Gender Corpus, gender rewriting systems,



*Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 173–180
November 9, 2026 ©2026 Association for Computational Linguistics

and shared tasks that treat gender as a forced grammatical choice plus a user who bears the cost of the default. Third, evaluation should start from that insight: name which morphological cell was tested, name which form the system defaulted to, and distinguish gender evaluation work built from Arabic grammar from gender evaluation work translated from English methods.

## 2 What Arabic requires of a system

Three facts about Arabic matter for evaluation. None of them is new to Arabic NLP. All of them are easy to lose if the starting point is an English template.

**Gender is written into the sentence.** Changing who is being described is not a matter of swapping one token. Alhafni et al. (2022b) distinguish several rewrite types. A suffix change turns *'amiir* (prince) into *'amiira* (princess). A pattern change turns a masculine adjective into a different templatic shape. A lexical change replaces *akh* (brother) with *ukht* (sister). An enclitic change rewrites a pronoun attached to the word, so that 'your prince' marked for a masculine addressee is not the same string as 'your prince' marked for a feminine addressee. Habash (2010) is the standard description of this system for NLP. "Debiasing" cannot mean making two English pronouns equally likely. The system has to emit a fully inflected sentence.

A concrete speaker-gender example is a nominal sentence. The English sentences *I am a teacher and I am happy* do not grammatically mark the speaker. Arabic does, on the noun and the adjective: masculine *'anaa mu'allim wa-'anaa sa'iid* versus feminine *'anaa mu'allima wa-'anaa sa'iida*. In a chat interface or in MT from English, the speaker's grammatical gender is often not specified in the source. The system still has to choose one of those forms. A system that defaults to the masculine predicate is not incomplete. It is making an obligatory choice that favors one group.

What the language does *not* do matters too. First-person perfect *kallamtu* 'I spoke' and first-person imperfect *'uriidu* 'I want' do not change with speaker gender in Modern Standard Arabic. It is a mistake to treat those verbs as speaker-gender tests. Addressee gender, by contrast, is marked on the object clitic: *kallamtuka* 'I spoke to you (m.)' versus *kallamtuki* 'I spoke to you (f.)'. A four-way contrast needs both a form that marks the speaker and a form that marks the addressee, for example a gendered predicate plus a second-person clitic.

Orthography makes evaluation harder rather than easier. Some gender distinctions are not visible in undiacritized Arabic. The verbs *kallamtuka* and *kallamtuki* differ only through diacritics on the final clitic; stripped of diacritics, both surface as *klmtk*. By contrast, normalization of the final letter in *mu'allima* does not make it identical to *mu'allim*; the feminine suffix remains orthographically distinct after undiacritization. Evaluation should document diacritic handling and orthographic normalization and identify which gender contrasts remain observable after preprocessing.

**Speaker and addressee are separate variables.** Arabic second person is gendered. A sentence can be masculine-to-masculine, feminine-to-feminine, or mixed (Alhafni et al., 2022a,b). The Arabic Parallel Gender Corpus makes this concrete. APGC v1.0 covered first-person singular only, with 12,238 sentences (Habash et al., 2019; Alhafni et al., 2022b). The balanced APGC v2.0 corpus added second-person targets, including singular, dual, and plural, and grew to 80,326 sentences and 596,799 words (over six times the original) with four parallel target versions: masculine speaker and addressee (MM), feminine speaker and masculine addressee (FM), masculine speaker and feminine addressee (MF), and feminine speaker and addressee (FF) (Alhafni et al., 2022a,b). A benchmark that only toggles a third-person occupation noun never sees this space.

These four targets are not equivalent. A system might handle the symmetric cases (MM, FF) but fail on the mixed ones (FM, MF), which are common in real use. Another might produce the same gender in every cell. An aggregate score may conceal which configurations account for these failures.

**Dual and plural are morphologically distinct.** Alhafni et al. (2022a) expanded APGC v2.0 to include second-person singular, dual, and plural targets. The four-way MM/FM/MF/FF grid above is singular; dual and plural add separate morphological choices. In Modern Standard Arabic, the dual marks exactly two persons and has distinct forms; plural feminine and plural masculine are a different rewrite type than the singular *kallamtuka/kallamtuki* pair (Habash, 2010). A score that averages singular clitics with duals does not report a dual result. A system can look balanced on sin-

gular feminine and still emit the masculine plural for a group explicitly identified as entirely feminine. This failure is invisible in a WinoMT occupation item. Moreover, Habash (2010) shows that dialects do not copy the MSA plural inventory, so a performance number computed on MSA will not show what happens when a system trained on MSA encounters dialectal plural morphology.

**The masculine form is the unmarked default.** In English-to-Arabic translation, outputs historically matched masculine references more often than feminine ones. On gender-specific APGC sentences, Google Translate (as measured in September 2021) scored 13.1 BLEU on the MM (masculine-masculine) target and 8.6 BLEU on the FF (feminine-feminine) target, a gap of 4.5 BLEU.[1] The largest gap appeared on sentences requiring both first- and second-person gender specification: 6.2 BLEU between MM and FF. Even when the English source contained no gender-specific noun, the MM-FF gap remained 4.6 BLEU, evidence of a system default to masculine forms when the source underspecifies gender.

WinoMT evaluates occupational referent gender under ambiguous syntactic contexts, using pro- and anti-stereotypical pairs. The reported systems exhibit gender-biased translation errors, with failures particularly evident in anti-stereotypical cases despite explicit pronoun and coreference cues. For Arabic, feminine inflection often appears as *tāʾ marbūṭa*, so the test is concrete. Habash et al. (2019) showed that a gender-aware wrapper on top of a gender-blind MT system raised first-person feminine BLEU by 8 percent relative and masculine BLEU by 5.3 percent relative. APGC tests whether first- and second-person morphology match a stated user. Both find a masculine default but they do not find the same error. A system could get WinoMT closer to parity by hedging occupations and still address every user as a man. Another system could rewrite first-person predicates correctly but still translate *doctor* as masculine by default. Using only one number hides that split.

All these are published measurements and motivate our proposed evaluation framework.

## 3 Instruments and what they measure

A paper that reports one row and calls it "gender bias in Arabic" has measured a slice. As shown in Table 1, embedding-based association measures (Bolukbasi et al., 2016; Caliskan et al., 2017) quantify gender associations in vector space. Nadeem et al. (2021) introduce StereoSet, which evaluates stereotypical associations and language-modeling ability using contexts with stereotypical, anti-stereotypical, and unrelated alternatives. WinoMT sees occupation plus pronoun, not first-person predicates. APGC sees I and you morphology, not nationality or ability. ArGAN sees evaluation across gender, ability, and nationality, not whether *saʿiid* matched the speaker. Each instrument is valid for its row; the field errs by treating a single row as the full coverage and using one number as a fairness verdict. An evaluation framework should name which cell is being tested and report disaggregated results alongside any aggregate score.

Alrajhi et al. (2024) quantify gender bias in Arabic pretrained language models (AraBERT, CAMeLBERT, MARBERT, and others) using an Arabic template-based benchmark for evaluating gender bias, complementing the morphological and user-centered focus of APGC.

## 4 Where the English pipeline breaks

The English pipeline assumes you can hold gender constant while you swap a word. Arabic breaks that assumption in three ways that the literature already identifies.

First, linguistic gender and social gender are not the same thing (Larson, 2017; Cao and Daumé III, 2020; Devinney et al., 2022). Arabic grammatical gender is a two-way agreement system. Social gender is not. A method that only equalizes masculine and feminine *agreement* can still erase people who do not use those two forms, and it can still stereotype the people who do (Dev et al., 2021). Surveys of NLP bias papers keep finding that this distinction is stated and then ignored in the method (Stanczak and Augenstein, 2021; Devinney et al., 2022).

Second, a lower gap on a template is not the same as less harm. Blodgett et al. (2020) argued that "bias" papers often fail to say who is harmed and how. Savoldi et al. (2024) measured post-editing behavior for machine translation into Italian, Spanish, and German, finding greater technical and temporal effort for feminine outputs, with associated financial costs that existing bias measures did not adequately reflect. This motivates

[1] BLEU is computed per-configuration target; this is not an aggregate score.

| Inst. | Unit | Person | Missing | Ctrl |
|---|---|---|---|---|
| Embed. | vec. offset | n/a | morph. | no |
| StereoSet | context + alt. | varies | I/you | no |
| CrowS-Pairs | pair | varies | I/you | no |
| WinoMT | occ. | 3rd | S/A | no |
| MuST-SHE | MT | 1st+3rd | 2nd | no |
| APGC v2 | rewrite | 1st+2nd | occ. | post |
| ArGAN | LLM | mixed | speaker | no |

Table 1: Published gender instruments. Inst. = Instrument. Unit: what the instrument measures (vector space, sentence context, pronoun cues, morphological agreement, or generation task). Person: which grammatical persons are tested. Missing: phenomena not covered (morphology, speaker/addressee distinctions, occupations, or intersectional axes). Ctrl = Control (whether post-editing is measured). Abbreviations: Embed. = Embeddings; vec. = vector; alt. = alternatives; occ. = occupational phenomena; morph. = morphology; S/A = speaker/addressee; 2nd = 2nd person; speaker = speaker morphology (ArGAN covers gender, ability, and nationality but not speaker/addressee morphology matching).

an equivalent user study for Arabic. The APGC BLEU gaps above are one automatic trace of potential correction effort, but they are not a substitute for measuring post-editing behavior or asking users what the wrong form cost them.

Third, translating an English benchmark does not produce an Arabic one. WinoMT is useful as a stress test, and it does include Arabic (Stanovsky et al., 2019). It still starts from English occupation sentences. It does not ask whether duals were handled, whether second person was marked, whether dialectal agreement was treated as if it were Modern Standard Arabic (Habash, 2010), or whether nationality and disability shift the stereotype (Aly et al., 2025). APGC exists because that gap was obvious to people building Arabic systems (Habash et al., 2019; Alhafni et al., 2022a).

Arabic dialects make the last point sharper. Habash (2010) treats dialectal Arabic as a family of varieties with their own morphology, not as noisy MSA. Some varieties keep gender distinctions that others reduce in the plural. Egyptian and Levantine varieties both differ from MSA, and they differ from each other. A system trained and tested on MSA subtitles, then deployed on dialectal chat, is not "multilingual Arabic." It is one variety standing in for many. Reporting an English-style occupation gap on that system does not tell us what happened to a speaker of a different variety, or to a speaker who code-switches.

## 5 What already works, and what it is not

The Arabic literature already has a better object than pronoun swapping: *gender rewriting*. Given a sentence and a target gender for the speaker and the addressee, produce the sentence that matches those preferences (Habash et al., 2019; Alhafni et al., 2020, 2022b). The task is concrete. It has public data. It can sit on top of a gender-blind MT system as a post-editor, which is how Habash et al. (2019) and Alhafni et al. frame it.

Alhafni et al. (2022b) combine rule-based and neural rewriting and report 88.42 $M^2$ $F_{0.5}$ on a blind APGC test set. They also demonstrate a 3.05-point absolute improvement in $M^2$ $F_{0.5}$ on the first-person-only task. They further show the intended use: post-edit a commercial MT output so that first- and second-person morphology match the user's stated preferences. The WANLP 2022 shared task put the same problem in front of five teams from four countries, using APGC v2.1 (Alhafni et al., 2022c). Alhafni et al. (2023) then wrapped the rewriter in a user-facing tool that accepts Arabic or English and lets people set first- and second-person gender.

That work is easy to misread as "the mitigation." It is not. Rewriting handles a case when users can state a grammatical preference. It does not decide whether the system should have asked. The cited MSA systems do not establish performance across dialects with different agreement patterns. It does not address intersectional stereotypes about nationality or disability that emerge once the sentence is fluent. And it still works inside a two-

gender morphology. Cao and Daumé III (2020) and Dev et al. (2021) are clear about the cost of that limit for trans and non-binary people. Alhafni et al. (2023) explicitly acknowledge that limiting the choice to Arabic grammatical genders excludes non-binary and gender-inclusive expressions. Arabic NLP has not solved that problem by building a better rewriter.

There is another temptation: treating BLEU against MM and FF as if it were itself a fairness metric. It is not. Blodgett et al. (2021) argue that contrastive-pair benchmarks often lack a clear measurement model, and Delobelle et al. (2024) show that a bias metric is actionable only if intended use and reliability are stated. BLEU rises when output overlaps a reference. If the system already produces masculine morphology, the MM reference will match more n-grams. The 4.6 BLEU gap on English sentences classified as gender-ambiguous under the study's noun-based annotation shows a default exists. It is not a complete account of harm, and shrinking it by making all output less fluent would not be progress. That is why Savoldi et al. (2024) measured users' post-editing effort, and why Arabic evaluation should do the same.

The rewriter and the corpus are valuable contributions. However, they should not be treated as the end of the argument. They demonstrate that the right unit for evaluation is a grammatical choice the language forces, together with the user who bears the consequences of the system's choice.

## 6 Evaluation protocol

The protocol is a checklist. This section presents it as a formal framework for systematic reporting.

Naming what each paper tested, where the system defaulted, and what remained unmeasured turns the protocol from a series of principles into a method. A paper can measure a slice correctly without claiming to measure the whole.

## 7 Protocol, applied to published work

This section demonstrates the checklist as a method by applying it to published Arabic gender-bias research.

**Habash et al. (2019): Automatic reinflection.** Cell tested: gender identification and reinflection on automatically identified predicates and nominatives in MSA text. Default named: systems failed more often on feminine forms; gender-aware wrapper improved both masculine and feminine BLEU. Not tested: second-person marking, dual forms, addressee gender, dialect. Verdict: valid unit test for gender identification on MSA; not a first-person or second-person user study.

| Protocol Step | Report in Paper |
|---|---|
| 1. Cell tested | Which morphological choice? (person, number, predicate type, construction) |
| 2. Unit description | What is the test unit? (occupation, 1st person, 2nd person clitic, plural verb, etc.) |
| 3. System default | Where does masculine/MM leaning appear? (BLEU gap, frequency, pattern) |
| 4. Configurations tested | Which cells were actually measured? (singular, dual, plural; MM/FM/MF/FF) |
| 5. Configurations omitted | What was not tested? (nationality, ability, dialect, non-binary forms) |
| 6. Verdict | Is this the right unit for your claim? Does the unit match the conclusion? |

Table 2: Evaluation Checklist: apply this framework to any paper claiming gender evaluation results for Arabic or other grammatical gender languages.

**WinoMT on Arabic (Stanovsky et al., 2019).** Cell tested: third-person occupation noun and gender information supplied through English pronouns and coreference relations under ambiguous syntactic setup. Default named: systems exhibit gender-biased translation errors, including failures in anti-stereotypical contexts despite explicit pronoun-based gender cues. Not tested: first-person predicates, second-person clitics, dual, plural, or user override. Verdict: valid unit test for occupation and pronoun; not an Arabic speaker/addressee gender result.

**APGC commercial MT (Alhafni et al., 2022a; Google Translate, September 2021).** Cell tested: first- and second-person rewrite against MM/FM/MF/FF. Default named: 4.6 BLEU gap MM>FF on English sentences classified as gender-ambiguous under the study's noun-based annotation; gaps widened to 6.2 BLEU when both first and second person were gender-specific. Not tested: occupation stereotypes, nationality or ability, dialect. Verdict: this is the right object for speaker/addressee morphology. The BLEU gap is evidence of a system default, not a fairness metric. An aggregate score may conceal which

configurations cause failures.

ArGAN (Aly et al., 2025) tests gender, ability, and nationality in LLM completions; it does not replace APGC but complements it by evaluating biases across gender, ability, and nationality, which APGC does not jointly address.

## 8 Objections

We address four objections and reader concerns about the proposed protocol.

**Just translate WinoMT into Arabic.** WinoMT and APGC measure different phenomena. A system that accurately preserves occupational referent gender on WinoMT may still default to masculine first- and second-person morphology. Conversely, a system with correct gender-rewriting may still translate occupations stereotypically. Using only one metric obscures these independent failures. This holds for other language pairs: Savoldi et al. (2022) show that occupation accuracy misses agreement chains in French, Italian, and Spanish speech translation.

**Gender rewriting is already the fix.** Rewriting tools like those of Alhafni et al. (2023) handle stated user preferences but address only a subset of the problem: they assume users can specify their preferences, work only for MSA, and operate within a binary morphological system. Alhafni et al. (2023) explicitly acknowledge that a two-gender interface excludes non-binary and gender-inclusive expressions. They should be complemented by broader evaluation that captures the full range of gendered choices the language forces.

**MSA is enough; dialects are future work.** Systems trained on MSA subtitles and deployed on dialectal text stand in for many varieties simultaneously. A single metric computed on MSA does not reveal what happens to speakers of different varieties with different gender agreement systems.

**Open-ended models make the English template more useful, not less.** The opposite applies. Open-ended generation may require gender-marked forms even when the prompt and conversational context do not specify the relevant grammatical preference. The protocol still applies: name which choice was measured, name the system's default, and distinguish behavior across different configurations.

## 9 Beyond Arabic: grammatical gender languages

The same logic applies across languages with obligatory gender on predicates or pronouns. Savoldi et al. (2022) show that occupation-noun accuracy misses agreement chains in French, Italian, and Spanish speech translation. Bentivogli et al. (2020) introduce MuST-SHE, a naturally translated dataset with first- and third-person gender phenomena in those languages, but it still does not measure an Arabic I/you grid. Hebrew is a Semitic counterpart with related clause morphology, offering comparable methodological concerns. The point is methodological: obligatory gender on predicates is not unique to Arabic. An evaluation framework built from the grammar that forces a choice applies to the grammar type, not one language.

## 10 Conclusion

In Arabic, gender is not a feature to swap. It is a choice the sentence makes whether or not the source marked it. The systems we have default that choice in predictable ways. Arabic researchers have already built the tools (corpora, rewriters, shared tasks, user interfaces) that treat the choice as the object. Gender-bias research should follow that lead. Measure the forced choice. Say who pays for the default. Distinguish gender evaluation work built from Arabic grammar from gender evaluation work translated from English methods.

The novelty of this work is not in naming these problems. It is in standardizing the operational response. That standardization means: (1) a checklist reviewers can apply, (2) separate reporting of linguistic gender (MM/FM/MF/FF cells) from social gender claims, and (3) reporting results separately for morphological configurations alongside any aggregate score. The contribution is a systematic reporting protocol rather than a new linguistic insight.

The same logic applies beyond Arabic. Similar evaluation concerns arise in other grammatical gender languages. Adding more languages to an English evaluation template is not the same as developing evaluation starting from those languages' grammars. Habash, Alhafni, Bouamor, and the community around APGC have demonstrated what the latter approach produces. This paper operationalizes that approach through an explicit checklist. It should be treated as a methodological con-

tribution, not an area-specific case study.

## Limitations

This is a position paper. It does not offer new Arabic annotations or system scores. Numbers we cite from APGC and rewriting papers are those authors' published results; we do not re-run the systems. We focus on Modern Standard Arabic more than dialects; agreement and gender inventories differ across varieties. We do not propose non-binary Arabic morphology. We also do not report a new human study of post-edit labor; claims about harm rest on Savoldi et al. (2024) and the published APGC BLEU gaps, not on a new user sample.

## Ethical considerations

Asking users to state gender preferences can help people being misgendered by MT. It can also pressure people to declare a gender they would rather not give a system. In some contexts that declaration is not safe. Preference widgets are not a substitute for asking whether the application should assign gender at all.

More explicit gender identification can amplify harm: Sorokovikova et al. (2025) find negligible persona effects on factual tasks, but clear bias when models give salary-negotiation advice, demonstrating that a system's gender awareness does not by itself reduce bias. The protocol names the default and distinguishes the technical measurement from fairness claims; it does not argue that gender awareness reduces harm.

Gender rewriting tools can be used to police as well as include. A tool for rewriting first- and second-person morphology on command helps speakers who want the right form. The same tool can enforce a binary, out someone whose written gender does not match institutional records, or "correct" dialectal speech into MSA. Alhafni et al. (2023) explicitly note that a two-gender interface excludes other expressions. Technical success is not ethical success.

Building APGC required professional linguists and guidelines against heterocentric assumptions. Replicating that labor without paying annotators, or without naming who is in the data, repeats a different harm. Papers using Arabic speakers only as a test set for an English method should say so.

## Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Research, Development and Innovation (QRDI) Council.

## References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 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.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022a. The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022b. User-Centric Gender Rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States.

Bashar Alhafni, Nizar Habash, Houda Bouamor, Ossama Obeid, Sultan Alrowili, Daliyah Alzeer, Khawlah M. Alshanqiti, Ahmed ElBakry, Muhammad ElNokrashy, Mohamed Gabr, Abderrahmane Issam, Abdelrahim Qaddoumi, K. Vijay-Shanker, and Mahmoud Zyate. 2022c. The Shared Task on Gender Rewriting. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 98–107, Abu Dhabi, United Arab Emirates. https://aclanthology.org/2022.wanlp-1.10

Bashar Alhafni, Ossama Obeid, and Nizar Habash. 2023. The User-Aware Arabic Gender Rewriter. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 3–11, Tampere, Finland.

Wafa Alrajhi, Hend S. Al-Khalifa, and Abdulmalik S. Al-Salman. 2024. Quantifying Gender Bias in Arabic Pre-Trained Language Models. *IEEE Access*. DOI: 10.1109/ACCESS.2024.3404668.

Ranwa Aly, Yara Allam, Rana Gaber, and Christine Basta. 2025. ArGAN: Arabic Gender, Ability, and Nationality Dataset for Evaluating Biases in Large Language Models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 256–267, Vienna, Austria.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of Gender in NLP Bias Research. In *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency*, pages 2083–2102.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool.

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.

Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5356–5370.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-Centered Study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A Decade of Gender Bias in Machine Translation. *Patterns*, 6(6):101257.

Aleksandra Sorokovikova, Pavel Chizhov, Iuliia Eremenko, and Ivan P. Yamshchikov. 2025. Surface Fairness, Deep Bias: A Comparative Study of Bias in Language Models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 206–227, Vienna, Austria.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint arXiv:2112.14168*.

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.