Whose wife is it anyway?
Assessing bias against same-gender relationships in machine translation

Ian Stewart
Pacific Northwest National Laboratory
ian.stewart@pnnl.gov

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

Machine translation often suffers from biased data and algorithms that can lead to unacceptable errors in system output. While bias in gender norms has been investigated, less is known about whether MT systems encode bias about social *relationships*, e.g., "the lawyer kissed her wife." We investigate the degree of bias against same-gender relationships in MT systems, using generated template sentences drawn from several noun-gender languages (e.g., Spanish) and comprised of popular occupation nouns. We find that three popular MT services consistently fail to accurately translate sentences concerning relationships between entities of the same gender. The error rate varies considerably based on the context, and same-gender sentences referencing high female-representation occupations are translated with lower accuracy. We provide this work as a case study in the evaluation of intrinsic bias in NLP systems with respect to social relationships.

Bias Statement

(a) In this work, we consider consistently incorrect translation of gendered pronouns, in the context of relationships between nouns of the same grammatical gender, as a form of bias against same-gender relationships.

(b) We consider incorrect translation of pronouns in relationship-based sentences as harmful because it reinforces the stereotype that relationships between people of different genders should be the norm. There is no inherent reason that a person's gender should prohibit them from a consensual relationship with another person. NLP systems that only recognize certain types of relationships (i.e. different-gender) impose a normative bias on their users. Incorrect machine translations of same-gender relationships may disenfranchise people for whom their relationship is especially important and should not be mischaracterized. Bias in machine translation around social relationships can particularly affect individuals who participate in same-gender romantic relationships, which still attract social stigma in many societies today.

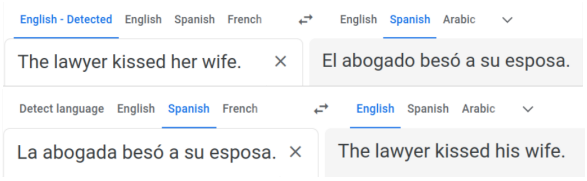


Figure 1: Example translation error of same-gender sentence between English and Spanish (Google Translate; accessed 1 November 2023).

1 Introduction

Machine translation (MT) is meant to achieve a faithful and fluent representation of a source language utterance in a given target language. While NLP research continues to improve the accuracy and robustness of MT systems (Lai et al., 2022; Liu et al., 2020), the full space of possible translation failures remains to be determined, particularly with respect to gender (Stanovsky et al., 2019). MT systems often generate masculine-gender words as the default for gendered languages (Savoldi et al., 2021), e.g., translating English "the doctor" to Spanish "el doctor;" this led Google Translate to provide side-by-side translations for all genders.

Focusing on word-based bias in MT is a good start, but translation systems may also exhibit *grammatical* bias involving relationships between words. In Figure 1, a sentence containing a same-gender relationship ("the lawyer kissed her wife") is re-translated as a sentence with a different-gender relationship ("his wife"), regardless of the starting language. This error seems to reveal the model's bias toward *fluent* translation at the cost of *faithfulness* (Feng et al., 2020), generating an output sentence with higher

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 365–375

likelihood in the target language ("his wife") but a possibly inaccurate meaning for the source language. Furthermore, this kind of grammatical error can only be brought to light by focusing on *relationships* between entities, an issue equally important as bias toward individual words like "doctor." Addressing bias in translation of relationships is important for such social groups as LGBTQ people, who often face discrimination for engaging in relationships with partners of the same gender (Poushter and Kent, 2020).

This study presents an analysis of the discrepancy in how translation systems handle same-gender vs. different-gender relationships, with a focus on languages with noun gender-marking. Our paper makes the following contributions:

- We generate a curated dataset of sentence templates on the topic of romantic relationships in prominent noun-gender languages (French, Italian, and Spanish). (§ 3.1).
- We test several leading MT models on this dataset, and we find a consistent bias against same-gender relationships when translating from a noun-gender language to English (§ 3.2).
- We assess possible correlates of bias using social factors and find that sentences referencing occupations with higher income have lower accuracy for same-gender relationships (§ 3.3).

This study not only highlights latent bias in MT, it also addresses the need to assess complex social constructs as part of bias testing, including relationships. Diagnosing and addressing this kind of bias can ensure that the needs of minority groups are addressed in the evaluation of common NLP methods (Blodgett et al., 2020).

We release all relevant data and code to replicate the study under a Creative Commons license.[1]

2 Related Work

Traditionally, research in ML-related bias has focused on well-established social demographics that are protected by law such as gender, race, and religion (Field et al., 2021; Nadeem et al., 2021; Rudinger et al., 2018). While demographics are an important area of focus, many other facets of social identity can also be affected by bias (Hovy and Yang, 2021), especially social *relationships*: power dynamics (Prabhakaran et al., 2012), friendship (Krishnan and Eisenstein, 2015), and romance (Seraj et al., 2021). A system that accurately processes such relationships has to understand not just individual identities (e.g., "man" and "woman") but also the social norms around the interactions between individuals (why two adults choose to live together) (Bosselut et al., 2019; Choi et al., 2020).

While norms around social relationships vary widely between societies (Miller et al., 2017), it is reasonable to assume that NLP systems should treat romantic relationships as equally valid regardless of the demographics of the participants. Furthermore, relationships represent an important part of social identity for many people (Wang and Jurgens, 2021), including LGBTQ people whose self-image may be negatively impacted by stereotypes about their relationships (Park et al., 2021). To fill the gap in the space of relationship-related bias, this study offers a path forward in assessing bias against with same-gender relationships in NLP systems.

Translating from one language to another is an inherently noisy process (Yee et al., 2019), sometimes leading to systematic errors that reveal inherent bias. Machine translation systems have been extensively audited for bias in prior work, particularly with respect to gender (Bianchi et al., 2023; Savoldi et al., 2021; Stanovsky et al., 2019) and linguistic structure (Behnke et al., 2022; Murray and Chiang, 2018; Vanmassenhove et al., 2021). Methods for mitigating bias in machine translation range from retraining on a targeted clean datasets (Saunders and Byrne, 2020; Stafanovičs et al., 2020) to modifying the model training/inference behavior for improved fairness (Lee et al., 2023; Sharma et al., 2022). This work contributes to the discussion in MT-related bias by evaluating gender bias in the context of social relationships, a previously under-explored area.

3 Assessing Bias in Relationship Translation

3.1 Data Generation

This study evaluates the presence of bias for same-gender vs. different-gender relationships in machine translation. To our knowledge, prior work in MT has not developed a dataset specifically to handle relationships based on pairs of grammatical gender, although some prior work has included

[1] Available at https://github.com/ianbstewart/multilingual-same-gender-bias.

Word category	Examples	Count
Occupation	el abogado (M; "lawyer"); la abogada (F)	100
Relationship template	X besó a Y ("X kissed Y")	5
Relationship target	el novio (M; "boyfriend"); la novia (F; "girlfriend")	6
Sentence	El abogado besó a su novio. ("The lawyer kissed his boyfriend.")	3000

Table 1: Summary of relationship sentences, for a single source language.

relationships as part of their data in assessment of gender bias (Kocmi et al., 2020; Troles and Schmid, 2021). We therefore develop our own data using a set of fixed sample sentences as templates.

We generate sample sentences to test the ability of multilingual models to process human relationships. We begin with sentence templates that describe a range of activities in romantic relationships, where each template has a subject X and an object Y, e.g., "X met Y on a date,". We fill in the subject position of the templates with occupation nouns which have different male and female versions in the source languages, e.g., Spanish "panadero" ("baker," male) vs. "panadera" (female). The occupations are drawn from a prior study of gender bias (Gonen and Goldberg, 2019).

We fill the object position of the templates with relationship targets, e.g., boyfriend/girlfriend. This procedure generates example sentences such as "El autor conoció a su esposo en una cita" ("The author met his husband on a date"). For each language we generate up to 3000 sentences to match every combination of occupation, gender, template, and target, and a summary is shown in Table 1.[2] All English translations for the relevant words and templates are listed in Table 3.

3.2 Same-Gender Bias in Translation

We test the ability of publicly available MT models to *faithfully* translate text about same-gender relationships vs. different-gender relationships. While we cannot cover all available translation services, we focus on several of the most popular services available to developers: Google Cloud Translation, Amazon Translate, and Microsoft Azure AI Translator (Amazon, 2023; Google, 2023; Microsoft, 2023).

[2]Not every language has exactly 3000 sentences due to missing words in certain languages, e.g. we omit "analyst" in French because the translation "l'analyste" has an identical female/male form and is therefore ambiguous in translation.

We provide all generated sentences to the translation model and specify English as the target language. We count a translation as correct if the gender of the English possessive pronoun in the translated sentence matches the gender of the subject noun in the source language sentence. For the Spanish sentence "la abogada besó a su esposa," we count the translated English sentence as correct if it contains the pronoun "her" for "the lawyer kissed her wife."

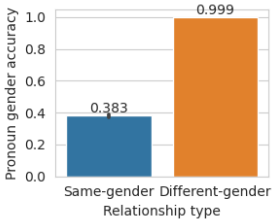
We show the aggregate results in Figure 2. All visualized differences are significant via McNemar's test ($p < 0.001$), where we test the difference in proportion correct vs. incorrect between the same-gender condition and the different-gender condition. In aggregate, the translation systems produce the correct subject gender at a lower rate for same-gender relationships than different-gender relationships (Figure 2a).

The accuracy is slightly better for female same-gender relationships than for male same-gender relationships (Figure 2b), which may indicate that the female-gender occupation words are inherently less ambiguous. Out of all the models, the Amazon MT model has the highest accuracy for same-gender relationships, but the gap between same-gender and different-gender relationships remains substantial with 51% accuracy for all same-gender relationship sentences versus 100% accuracy for different-gender relationship sentences (Figure 2c). Across all languages (Figure 2d), we see the best performance for Spanish, followed by French and Italian, which could indicate substantially different capabilities for the different languages, e.g. lower performance on Italian language in general.

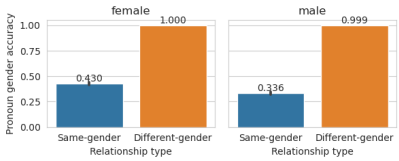
3.3 Assessing Social Correlates of Bias

The aggregate accuracy results reveal significant variation among different occupations (Figure 2e, 2f). Occupations with higher income tend to see a very low accuracy for same-gender translations (e.g. "judge," 15% accuracy), while occupations that may be more well-represented in popular media have higher accuracy for same-gender translations ("athlete," 66% accuracy), although the accuracy never reaches parity. This variation across occupations leads us to test the relative effect of different aspects of the occupations, to investigate social correlates of bias.

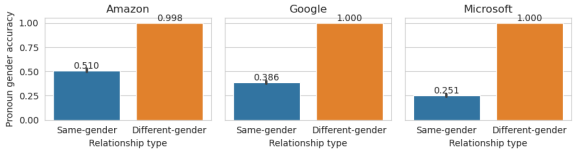
Prior work in NLP bias has found correlations



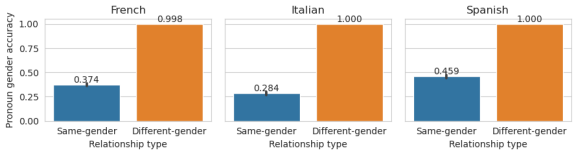
(a) Aggregate accuracy.



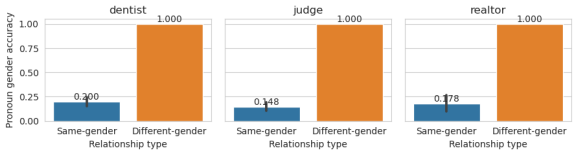
(b) Accuracy per-gender (subject).



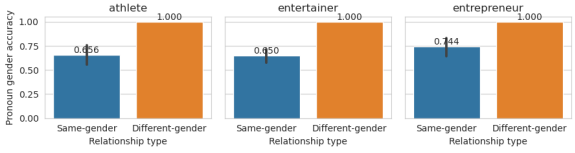
(c) Accuracy per-model.



(d) Accuracy per-language.



(e) Accuracy per-subject word, for relationship subjects with lowest same-gender accuracy.



(f) Accuracy per-subject word, for relationship subjects with highest same-gender accuracy.

Figure 2: Translation accuracy for relationship sentences, grouped by relationship type (same-gender vs. different-gender).

with language-external phenomena that relate to the perception of various social groups, such as immigrant populations and their representation in word embeddings (Garg et al., 2018). To that end, we conduct additional analysis of the bias using social variables that map to the different occupations mentioned in the example sentences:

- Income level (high-income occupations may be more equitable);
- Female representation (high female-representation occupations may be more equitable);
- Age representation (youth-oriented occupations may be more equitable).

We collect the occupation-related variables using statistics from the US Department of Labor and Bureau of Labor Statistics (BLS, 2023; DOL, 2023). We manually match each occupation to the corresponding official category: e.g., "boss" is mapped to "General and Operations Managers" (see Appendix B).

We run a logistic regression to predict whether a sentence was translated with the correct subject gender, limiting the analysis to same-gender sentences to isolate correlates of the bias. We add categorical variables for the subject gender, source language, MT model, and the relationship target. We also include the occupation-related variables mentioned above as scalar values, with the values Z-normalized for fair comparison of effect sizes. The regression can be represented with the following equation:

$$\begin{aligned}\text{Correct-Gender} \sim\ & \beta_1 * \text{Subject-Gender} + \\ & \beta_2 * \text{Language} + \beta_3 * \text{Model} + \\ & \beta_4 * \text{Relationship-Target} + \beta_5 * \text{Income} + \\ & \beta_6 * \text{Female-Representation} + \\ & \beta_7 * \text{Age} + \epsilon \end{aligned} \quad (1)$$

The regression results are shown in Table 2. The model replicates the trends observed from aggregate comparisons: lower likelihood of correct subject-gender prediction for sentences with a male-gender subject, sentences in Italian, and in cases where the Microsoft MT model was used. We also find that a lower likelihood of correct subject-gender prediction for occupations that had a higher income, a higher female representation, and higher age.

The negative correlation between female representation and accuracy is somewhat unexpected. The correlation may be related to the more general bias against occupations with traditionally higher female representation, e.g. "secretary" being associated with more traditionally "female" norms such as "her husband." As for the other occupation variables, the MT systems may have learned more social conservative norms associated with high-income occupations (e.g. dentist, lawyer) and higher-age occupations (farmer, judge).

	β	SE	Z	p
Intercept	1.3091	0.067	19.642	*
Subject gender (default female)				
Male	-0.5664	0.047	-12.024	*
Language (default French)				
Italian	-0.5329	0.062	-8.632	*
Spanish	0.5156	0.055	9.294	*
Model (default Amazon)				
Google	-0.7138	0.057	-12.598	*
Microsoft	-1.5303	0.060	-25.616	*
Relationship target (default fiancé(e))				
Boy/girlfriend	-0.3981	0.051	-7.823	*
Husband/wife	-2.9832	0.073	-41.020	*
Occupation variables				
Income	-0.1915	0.027	-6.993	*
Female representation	-0.3110	0.027	-11.516	*
Age	-0.1227	0.031	-3.930	*

Table 2: Logistic regression for correct pronoun prediction for same-gender sentences; positive coefficient means higher likelihood of correct pronoun prediction. d.f.=10, N=11070, LLR=3758 (p<0.001). * indicates $p < 0.001$.

4 Conclusion

In this study, we identified consistent bias against same-gender relationships in MT among several Romance languages. Using Google Translate, we identified consistent bias against same-gender relationships, across language, topic, and subject type. Upon further investigation, we found that occupations with higher income, higher female representation, and higher median age tend to exhibit higher rates of bias. Future MT systems may need to change their training or inference strategy to represent a wider range of relationships. Such a bias in MT systems can have a variety of downstream impacts, including misrepresentation of same-gender relationships across languages, enforcing normative social stereotypes, and erasing the lived experience of people who participate in same-gender relationships.

Future work should broaden the investigation of how relationships are processed in multilingual models, including coreference resolution (Emelin and Sennrich, 2021) and natural language inference (Rudinger et al., 2017), to provide a more complete picture into the representation of relationships with varying social composition. While our study does not address underlying issues facing LGBTQ people such as legal discrimination, it does provide a way forward to identify implicit bias in NLP systems. We hope that the study encourages AI researchers to take a broader view of "ethics" when it comes to the design and evaluation of such systems as machine translation, in order to include minority groups who may not be considered visible (Hutchinson et al., 2020).

Limitations We acknowledge that the study is limited to a sub-set of languages, due to the need for grammatical gender marked on NP and unmarked on possessive pronouns. While this analysis is not appropriate for all languages, it can be adapted to fit other situations, e.g. identifying the inferred possessive pronoun when translating from a language without explicit possession marking (e.g. translating "she met ø wife" from Norwegian; Lødrup 2010) to a language with explicit possession marking.

From a linguistic perspective, the study also only focuses on one direction of translation (gender-NP to no-gender-NP), even though the opposite direction (no-gender-NP to gender-NP) is known to exhibit gender bias (Stanovsky et al., 2019). Future studies should assess bias in multiple translation directions, as well as to/from languages without any grammatical gender such as Chinese.

The analysis of occupations (§ 3.3) uses statistics from the United States, which may not match the statistics of the countries in which the languages under study are spoken. We assume that the relative *ranking* of occupations by the social variables will not be significantly different between countries. This is a strong assumption to make for all occupations but is likely to hold for at least the most popular occupations: e.g., in many countries, a physician will earn more money than a nurse. We acknowledge that it's not a perfect measurement for the socioeconomic correlates of occupation and look to future work to develop more fine-grained metrics for occupation social status, e.g. relative female representation per-country per-occupation.

5 Ethical Considerations

This study addresses the ethical ramifications of machine translation with respect to a large but not necessarily visible population, namely people who participate in same-gender relationships. Although not all LGBTQ people engage in same-gender relationships, they represent a sizable proportion of the US population, around 5.6% by a recent estimate (Jones, 2021). People in same-gender relationships specifically have often

faced considerable legal and social opposition within the US (Avery et al., 2007; Soule, 2004), and part of that opposition extends to the technology that supports communication in everyday life.

As a caveat around relationships, we want to emphasize that our study does not cover all types of relationships where gender plays an important role. In particular, we focus on grammatical gender rather than social gender, which may be an ethical concern. To illustrate this point, consider a situation where a person referred to as "el abogado" (Sp. masculine) identifies as female, which is an ongoing debate among speakers of noun-gender languages (Burgen, 2020; Horvath et al., 2016; Lipovsky, 2014). In this case, a sentence with "el abogado" as subject noun and a masculine-gender target noun (e.g. "su novio") may in fact refer to a relationship between a female-gender person and a male-gender person. Having established this, we do not claim that MT systems are necessarily biased with respect to the social or psychological construct of gender, only the grammatical construct of gender (Alvanoudi, 2014). In addition, we acknowledge that not all relationships should be considered valid when testing MT systems, e.g. relationships with an imbalance in age or power which may be a sign of abuse (Volpe et al., 2013).

As a particularly notable concern, our analysis only focuses on the binary case of masculine and feminine grammatical gender. This decision naturally omits the wide range of gender-neutral and non-binary expression available even in languages with traditional masculine/feminine noun gender (Hord, 2016). We do not claim that gender should always be studied as a binary variable. For example, gender-neutral pronouns should be accurately handled in coreference resolution (Cao and Daumé III, 2020). Future work should investigate the treatment of gender-neutral language in relationship-focused text, considering the additional complications that MT systems must overcome when handling constructs such as gender-neutral pronouns.

In this analysis, we do not claim that the observed bias is malicious or even intentional, only that it is systematic and should be corrected. Engineers who build AI systems such as Google Translate are rarely aware of all possible downstream errors that their system can cause (Nushi et al., 2017). Our study should not be used to blame individuals but instead highlight the kinds of stress-testing that machine translation systems need before they are released for public use.

References

Angeliki Alvanoudi. 2014. *Grammatical gender in interaction: Cultural and cognitive aspects*. Brill.

Amazon. 2023. Amazon Translate.

Alison Avery, Justin Chase, Linda Johansson, Samantha Litvak, Darrel Montero, and Michael Wydra. 2007. America's changing attitudes toward homosexuality, civil unions, and same-gender marriage: 1977–2004. *Social work*, 52(1):71–79.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487.

Federico Bianchi, Tommaso Fornaciari, Dirk Hovy, and Debora Nozza. 2023. Gender and age bias in commercial machine translation. In *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 159–184. Springer.

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.

BLS. 2023. Labor force statistics from the current population survey.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

S Burgen. 2020. Masculine, feminist or neutral? The language battle that has split Spain. *The Guardian*.

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.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, pages 1514–1525.

DOL. 2023. Employment and earnings by occupation.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532.

Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 59–66.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. 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 1905–1925.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. 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 609–614.

Google. 2023. Google Translate.

Levi CR Hord. 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics*, 3(1).

Lisa K Horvath, Elisa F Merkel, Anne Maass, and Sabine Sczesny. 2016. Does gender-fair language pay off? the social perception of professions from a cross-linguistic perspective. *Frontiers in psychology*, 6:2018.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Jeffrey M Jones. 2021. LGBT identification rises to 5.6% in latest US estimate. *Gallup News*, 24.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender Coreference and Bias Evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364.

Vinodh Krishnan and Jacob Eisenstein. 2015. "You're Mr. Lebowski, I'm the Dude": Inducing Address Term Formality in Signed Social Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2022. Improving both domain robustness and domain adaptability in machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

Caroline Lipovsky. 2014. Gender-specification and occupational nouns: has linguistic change occurred in job advertisements since the French feminisation reforms? *Gender & Language*, 8(3).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Helge Lødrup. 2010. Implicit possessives and reflexive binding in norwegian. *Transactions of the Philological Society*, 108(2):89–109.

Microsoft. 2023. Azure AI Translator.

Joan G Miller, Hiroko Akiyama, and Shagufa Kapadia. 2017. Cultural variation in communal versus exchange norms: Implications for social support. *Journal of Personality and Social Psychology*, 113(1):81.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490.

Jacob Poushter and Nicholas Kent. 2020. The global divide on homosexuality persists. *Pew Research Center*, 25.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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.

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Sarah Seraj, Kate G Blackburn, and James W Pennebaker. 2021. Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences*, 118(7).

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sarah A Soule. 2004. Going to the chapel? Same-sex marriage bans in the United States, 1973–2000. *Social problems*, 51(4):453–477.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541.

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

Ellen M Volpe, Thomas L Hardie, Catherine Cerulli, Marilyn S Sommers, and Dianne Morrison-Beedy. 2013. What's age got to do with it? Partner age difference, power, intimate partner violence, and sexual risk in urban adolescents. *Journal of interpersonal violence*, 28(10):2068–2087.

Sky Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9918–9938.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. 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 5696–5701, Hong Kong, China. Association for Computational Linguistics.

A Template Data

We list the English translations of all words and phrases used to construct the translation sentences (§ 3.1) in Table 3. To save space we omit the target language translations of all words and phrases, but this data will be made available on the public repository after publication.

B Occupation Metadata

The occupations used in the sample data for the regression analysis (§ 3.3) were manually mapped to categories via statistics from the US Department

	Words/Phrases
Source noun (occupations)	analyst; artist; athlete; author; baker; banker; barber; boss; carpenter; coach; consultant; cop; counselor; custodian; dancer; dentist; director; doctor; editor; electrician; engineer; entertainer; entrepreneur; farmer; firefighter; journalist; judge; laborer; landlord; lawyer; librarian; mechanic; nanny; nurse; painter; pharmacist; photographer; plumber; president; professor; psychologist; realtor; scientist; secretary; senator; singer; student; surgeon; teacher; writer
Sentence template	X met PRON Y on a date.; X kissed PRON Y.; X married PRON Y.; X lived with PRON Y.; X and PRON Y have a child.
Target noun (relationship terms)	fiancé(e); girlfriend/boyfriend; wife/husband

Table 3: All occupations, relationship templates, and relationship targets used to generate the data for the study.

of Labor and Bureau of Labor Statistics. We list the occupation metadata in Tables 4 and 5. Empty cells indicate missing data not included in the regression.

Occupation	BOLS Categories	DOL Category	Median income	% Female	Median age
analyst	Management analysts; Budget analysts; Credit analysts; Financial and investment analysts; Computer systems analysts; Information security analysts; Software quality assurance analysts and testers; Software quality assurance analysts and testers	Budget analysts; Computer systems analysts; Credit analysts; Financial and investment analysts; Information security analysts; Management analysts; Market research analysts and marketing specialists; News analysts, reporters, and journalists; Operations research analysts; Software quality assurance analysts and testers	84776	41.09	
artist	Artists and related workers	Artists and related workers	49032	38.20	43.30
author	Writers and authors	Writers and authors	61189	55.50	44.80
baker	Bakers	Bakers	29241	57.40	41.70
banker	Financial managers; Business and financial operations occupations; Financial and investment analysts; Personal financial advisors; Financial examiners; Other financial specialists; Financial clerks, all other	Financial and investment analysts; Financial clerks, all other; Financial examiners; Financial managers	83174	49.67	
barber	Barbers	Barbers	29283	21.20	40.80
boss	General and operations managers; Advertising and promotions managers; Marketing managers; Sales managers; Public relations and fundraising managers; Administrative services managers; Facilities managers; Computer and information systems managers; Financial managers; Compensation and benefits managers; Human resources managers; Training and development managers; Industrial production managers; Purchasing managers; Transportation, storage, and distribution managers; Construction managers; Education and childcare administrators; Architectural and engineering managers; Food service managers; Funeral home managers; Entertainment and recreation managers; Lodging managers; Medical and health services managers; Natural sciences managers; Postmasters and mail superintendents; Property, real estate, and community association managers; Social and community service managers; Emergency management directors; Personal service managers, all other; Managers, all other;	Computer and information systems managers; Construction managers; Entertainment and recreation managers; Facilities managers; Financial managers; Food service managers; General and operations managers; Human resources managers; Industrial production managers; Lodging managers; Managers, all other; Marketing managers; Medical and health services managers; Natural sciences managers; Public relations and fundraising managers; Purchasing managers; Sales managers; Social and community service managers; Training and development managers; Transportation, storage, and distribution managers	77496	42.43	
carpenter	Carpenters	Carpenters	40759	1.90	40.80
coach	Coaches and scouts	Coaches and scouts	47895	31.60	34.60
cop	Police officers	Police officers	67927	14.80	40.50
counselor	Credit counselors and loan officers; Substance abuse and behavioral disorder counselors; Educational, guidance, and career counselors and advisors; Mental health counselors; Rehabilitation counselors; Counselors, all other	Substance abuse and behavioral disorder counselors; Counselors, all other; Credit counselors and loan officers; Educational, guidance, and career counselors and advisors; Mental health counselors; Rehabilitation counselors	54882	61.34	
custodian	Building and grounds cleaning and maintenance occupations	Janitors and building cleaners			46.40
dentist	Dentists	Dentists	152233	32.00	46.60
director	Producers and directors; Music directors and composers; Emergency management directors; Directors, religious activities and education	Directors, religious activities and education; Producers and directors	65662	43.54	
editor	Editors	Editors	62494	53.90	45.40
electrician	Electricians	Electricians	52959	1.80	41.40
engineer	Aerospace engineers; Agricultural engineers; Bioengineers and biomedical engineers; Chemical engineers; Civil engineers; Computer hardware engineers; Electrical and electronics engineers; Environmental engineers; Industrial engineers, including health and safety; Marine engineers and naval architects; Materials engineers; Mechanical engineers; Mining and geological engineers, including mining safety engineers; Nuclear engineers; Petroleum engineers; Engineers, all other\n	Aerospace engineers; Chemical engineers; Civil engineers; Electrical and electronics engineers; Engineers, all other; Environmental engineers; Industrial engineers, including health and safety; Materials engineers; Mechanical engineers	93763	13.49	
entertainer	Entertainers and performers, sports and related workers, all other	Other entertainment attendants and related workers			23.80
farmer	Farmers, ranchers, and other agricultural managers	Farmers, ranchers, and other agricultural managers	42498	12.10	56.00
firefighter	Firefighters	Firefighters	71600	3.50	39.70
journalist	News analysts, reporters, and journalists	News analysts, reporters, and journalists	61427	46.30	34.90
judge	Judges, magistrates, and other judicial workers	Judges, magistrates, and other judicial workers	105383	49.30	53.10
laborer	Construction laborers; Laborers and freight, stock, and material movers, hand	Laborers and freight, stock, and material movers, hand	33850	11.79	35.00
landlord	Property, real estate, and community association managers	Property, real estate, and community association managers	56061	52.40	48.70
lawyer	Lawyers	Lawyers	131501	37.50	46.50
librarian	Librarians and media collections specialists	Librarians and media collections specialists	54259	81.80	49.90
mechanic	Automotive service technicians and mechanics; Bus and truck mechanics and diesel engine specialists; Heavy vehicle and mobile equipment service technicians and mechanics; Small engine mechanics; Miscellaneous vehicle and mobile equipment mechanics, installers, and repairers	Aircraft mechanics and service technicians; Automotive service technicians and mechanics; Industrial and refractory machinery mechanics	40814	2.00	
nanny	Childcare workers	Childcare workers	23064	94.70	37.70

Table 4: Occupations and associated metadata for regression (part 1).

Occupation	BOLS Categories	DOL Category	Median income	% Female	Median age
nurse	Registered nurses	Registered nurses	69754	86.70	43.10
painter	Painters and paperhangers	Painters and paperhangers	33965	7.40	41.50
pharmacist	Pharmacists	Pharmacists	122473	54.60	41.40
photographer	Photographers	Photographers	44026	41.00	39.60
plumber	Plumbers, pipefitters, and steamfitters	Plumbers, pipefitters, and steamfitters	50451	1.40	40.60
president					
professor	Postsecondary teachers	Postsecondary teachers	72172	47.60	49.40
psychologist	Clinical and counseling psychologists; School psychologists; Other psychologists	Other psychologists	85411	68.30	48.60
realtor	Real estate brokers and sales agents	Real estate brokers and sales agents	61192	51.50	49.10
scientist	Life, physical, and social science occupations; Agricultural and food scientists; Biological scientists; Conservation scientists and foresters; Medical scientists; Life scientists, all other; Astronomers and physicists; Atmospheric and space scientists; Chemists and materials scientists; Environmental scientists and specialists, including health; Geoscientists and hydrologists, except geographers; Physical scientists, all other; Economists	Agricultural and food scientists; Biological scientists; Chemists and materials scientists; Computer and information research scientists; Conservation scientists and foresters; Environmental scientists and specialists, including health; Geoscientists and hydrologists, except geographers; Medical scientists; Miscellaneous social scientists and related workers; Physical scientists, all other	80335	43.84	
secretary	Executive secretaries and executive administrative assistants; Legal secretaries and administrative assistants; Medical secretaries and administrative assistants; Secretaries and administrative assistants, except legal, medical, and executive	Secretaries and administrative assistants, except legal, medical, and executive	42282	94.00	48.50
singer	Musicians and singers	Musicians and singers	42121	20.90	44.20
teacher	Preschool and kindergarten teachers; Elementary and middle school teachers; Secondary school teachers; Special education teachers; Tutors; Other teachers and instructors	Preschool and kindergarten teachers; Secondary school teachers; Special education teachers; Elementary and middle school teachers; Other teachers and instructors	50141	75.26	
writer	Technical writers; Writers and authors	Writers and authors; Technical writers	65267	55.69	

Table 5: Occupations and associated metadata for regression (part 2).