Can Emotions Signal Gender? Investigating Implicit Cues in Human and LLM Translations of Amazon Reviews

Shushen Manakhimova[1,2] **Ekaterina Lapshinova-Koltunski**[2]
[1]DFKI, Berlin, Germany
[2]University of Hildesheim, Germany
manakhimova@uni-hildesheim.de, lapshinovakoltun@uni-hildesheim.de

Abstract

This study investigates whether source-text emotion is associated with grammatical gender choices in English-Russian translation when the author's gender is not specified. Building on Popović and Lapshinova-Koltunski (2024), we analyse first-person grammatical gender in Amazon review translations produced by professional translators, translation students, and three LLMs: GPT-4, Llama, and Mistral. We assign emotion labels to the English source reviews and test whether these labels are associated with masculine or feminine forms in the Russian translations. Reviews labelled as expressing *love* are more likely to be translated with feminine grammatical gender by both human translator groups, with small-to-moderate effect sizes. This pattern is absent in GPT-4 and Mistral, but appears in Llama. Other frequent emotion labels do not show the same positive association. We therefore treat the findings as exploratory evidence that specific source-text emotions may be associated with gendered translation choices, while emphasizing the need for larger-scale validation across additional emotions, genres, and language pairs.

1 Introduction

This paper presents a study on gender and emotion in Amazon product review translations. We focus on first-person grammatical gender, following Popović and Lapshinova-Koltunski (2024), who showed that product category functions as an implicit gender cue in English-Russian and English-Croatian review translations, with masculine forms dominating across all translator types. Unlike English, Russian marks gender on past-tense verbs, adjectives, and passive participles, requiring translators to specify the reviewer's gender even when the source text provides no such information. What other contextual factors might drive these gendered choices remains an open question (Hackenbuchner et al., 2024).

Psychology research has documented systematic associations between specific emotions and gender stereotypes. For example, people seem to believe women experience and express the majority of emotions more than men, including sadness, fear, and love with anger and pride as the main exceptions (Rosenkrantz et al., 1968). These patterns hold across cultures (Brody and Hall, 2008). Large language models (LLMs) seem to replicate this stereotype pattern by consistently attributing anger to male personas and sadness to female ones when prompted with gendered scenarios (Plaza-del-Arco et al., 2024). Our central goal is to examine whether source-text emotion is associated with grammatical gender choices in translation when gender is otherwise under-determined.

To achieve this goal, we use emotion as an explanatory variable for gender bias, applying BERT-based emotion classifiers based on the GoEmotions taxonomy (Demszky et al., 2020) to the English source texts. We then analyse first-person gender in professional and student human translators, as well as in the three LLM translation outputs: GPT-4, Llama and Mistral to ensure a broader comparison across model families. Following previous work on this corpus (Popovic

and Lapshinova-Koltunski, 2024), our aim is not to mitigate gender bias directly, but to examine whether systematic patterns in translation data can be linked to gendered emotion stereotypes present in the source texts.

For our analysis, we formulate the following research questions (RQs):

RQ1: Are classifier-assigned emotion labels in the English source texts associated with grammatical gender choices in Russian translations, and are these associations specific to particular emotions?

RQ2: Do these patterns differ across human translator groups and LLMs?

Our study is correlational: we do not claim to identify the causes of gender bias, but rather to uncover patterns in translation data. The analysis focuses on binary gender (masculine/feminine) as required by Russian grammar, and we acknowledge this as a limitation.

The remainder of the paper is organised as follows: Section 2 reviews related work. Section 3 describes the corpus and annotation. Section 4 outlines the methodology. Results are presented in Section 5, followed by discussion and conclusion in Sections 6 and 7. We also discuss the limitations in Section 8.

2 Related Work

A widely reported pattern in MT is the masculine default: absent other information, masculine forms are used disproportionately across all translator types (Savoldi et al., 2021; Stanovsky et al., 2019). Savoldi et al. (2021) provide the foundational framework for the field, defining gender bias in MT as systematic and unfair discrimination that manifests principally as underrepresentation of feminine forms and stereotyped translation choices. Recent work also situates gender bias in MT within the broader shift toward LLMs (Vanmassenhove, 2024).

Our study builds directly on Popović and Lapshinova-Koltunski (2024), who examine first-person grammatical gender in human and machine translations of English Amazon product reviews into Russian and Croatian, using the DiHuTra corpus. Their analysis confirms a strong masculine default across all translator types, but also shows that the masculine default is not absolute: product category functions as an implicit cue, with beauty reviews attracting substantially more feminine forms and sports reviews rendered almost exclusively in the masculine. Human translators show considerably more sensitivity to these cues than MT systems, which produce near-uniformly masculine output even in categories where feminine forms might be expected. This finding motivates the present study: if the product category can function as an implicit gender cue, what other features of the source text might do the same?

This broader role of context is also supported by work on gender decisions under ambiguity. In an English-German study, gender assignments by DeepL were compared with those of 22 human annotators for the role names presented both in isolation and in sentence context (Hackenbuchner et al., 2024). Adding context changed 44% of the human gender labels, compared to only 17% of the MT labels. This suggests that humans are more context-sensitive than MT systems. For MT, gender shifts mostly occurred when the context contained explicit cues, such as pronouns, female proper names, or an unambiguous referent.

The theoretical motivation for examining emotion as a gender cue rests on associations between specific emotions and gender stereotypes in psychology. Since at least the late 1960s, cultural gender stereotypes have reliably assigned emotional traits along gender lines (Rosenkrantz et al., 1968). A partial replication three decades later confirms that cross-gender agreement on these trait attributions remains high, though the number of items reaching stereotype consensus also decline (Nesbitt and Penn, 2000). Plant et al. (2000) show that people believe women experience and express the majority of emotions more intensely than men, with anger and pride as the main exceptions. Brody and Hall (2008) extend this picture across cultures: women are stereotypically linked to love, sadness, fear, and guilt; men to anger and pride. These associations are sufficiently robust and widespread that they can reasonably be expected to form part of the background knowledge of any translator working between the two languages in our study.

Recent NLP work suggests that these associations are also reflected in LLM behaviour. Plaza-del-Arco et al. (2024) prompted five state-of-the-art LLMs with identical scenarios while varying only the gender of the persona; all models consistently attributed anger to male personas and sad-

ness to female ones, closely mirroring the psychological literature. Related evidence comes from Weber et al. (2026), who examine whether demographic information about the author of an emotionally ambiguous text influences emotion interpretation. They find that human annotators are more likely to match the author-provided emotion label when that label aligns with gender stereotypes; their LLM experiments point in the same direction. Together, these studies suggest that associations between gender and emotion are not only documented in psychological research, but also appear in recent NLP settings involving both human judgments and LLM behaviour.

To our knowledge, the connection between emotion and gender has not yet been systematically explored in translation. This study addresses that gap by examining whether source-text emotion is associated with grammatical gender choices when the author's gender is not specified in the source text. In doing so, we connect work on gender bias in translation with research on gendered emotion stereotypes in psychology and NLP. By also considering product category, we assess whether emotion provides explanatory value beyond the domain-level associations already identified in prior work.

3 Data

Our study uses the DiHuTra corpus (Lapshinova-Koltunski et al., 2022), a publicly available parallel corpus of 196 English Amazon product reviews and their translations into multiple languages, collected in 2022[1]. The corpus is structured as 14 reviews in each of 14 product categories: beauty, books, CD/vinyl, cell phones, grocery and gourmet food, health care, home and kitchen, movies and TV, musical instruments, patio and garden, pet supplies, sports and outdoors, toys and games, and video games. The corpus contains reviews across a range of star ratings (1-2 = negative, 4-5 = positive); 3-star reviews were excluded to obtain a clear positive/negative split (Popovic and Lapshinova-Koltunski, 2024).

The human translations were produced by two groups: professional translators and translation students. Translators were provided with the review texts only — no star ratings or other metadata were shared, ensuring that any gender choices reflect responses to the text itself. They were also given minimal instructions and were not allowed to use any MT systems. Crucially, no instructions were given regarding how to handle grammatical gender in the target language, making the corpus suitable for studying gender choices as individual decisions. The present study extends the data to LLM-generated translations: GPT-4[2], Llama[3], and Mistral[4], each prompted only with "translate into Russian".

Gender annotation of all Russian translations was performed manually at both the segment and review levels, following Popović and Lapshinova-Koltunski (2024). The annotation identifies first-person gendered words in the Russian text, primarily past participles, adjectives, and passive participles.

Following their annotation scheme, each review is assigned a gender label according to the first-person gendered words it contains. If all first-person gendered words within a review share the same gender, the review is labelled m (masculine) or f (feminine) accordingly. If the review contains no first-person gendered words at all, it is labelled x. If the review contains a mixture of first-person genders, it is labelled mixed (covering sub-types such as m+f, m—f, m+m—f, and f+m—f).

Group	f	m	x	mixed
GPT-4	21	113	61	1
Llama	24	101	63	8
Mistral	23	113	59	1
Prof	36	88	71	1
Stud	27	91	78	0

Table 1: Distribution of gender labels across translator groups (n = 196). f = all first-person gendered words are feminine; m = all first-person gendered words are masculine; x = no first-person gendered words present; mixed = review contains both masculine and feminine markings

We restrict our quantitative analysis to reviews unambiguously labeled as masculine (**m**) or feminine (**f**). All mixed or ambiguous cases are excluded. Reviews where the Russian text requires no first-person gender marking are similarly excluded. This results in between 118 and 136 reviews per translator group available for statistical analysis, depending on the proportion of gendered constructions in each translation variant.

[1] https://github.com/katjakaterina/dihutra

[2] accessed in December 2024 through https://chat-ai.academiccloud.de/chat

[3] accessed in January 2025 through https://chat-ai.academiccloud.de/chat

[4] accessed in January 2025 through https://chat-ai.academiccloud.de/chat

4 Methodology

4.1 Emotions under analysis

Prior work on gender and emotion stereotypes links emotions such as *love*, *sadness*, *fear*, *guilt*, and *caring* with femininity, while *anger* and *pride* are more often associated with masculinity (Brody and Hall, 2008; Plant et al., 2000; Plaza-del-Arco et al., 2024). We use this literature as a theoretical motivation for examining whether emotional content may function as an implicit cue for grammatical gender choice in translation.

The stereotypically masculine emotions identified in the literature, such as *anger* and *pride*, do not appear among the top-1 or top-2 labels assigned by the GoEmotions classifier in our corpus and are therefore excluded from the analysis.

4.2 Emotion annotation

For emotion classification of the English source texts, we use `SamLowe/roberta-base-go_emotions`[5], a RoBERTa-base transformer fine-tuned on the GoEmotions dataset (Demszky et al., 2020). The model predicts 27 fine-grained emotion categories plus *neutral*, grouped into positive, negative, ambiguous, and neutral classes. We selected this model because its taxonomy is compatible with our research question and because it is publicly available and can be run locally, supporting reproducibility. The model outputs a probability distribution over all labels for each review.

In the main analysis, we retain the two highest-probability labels per review. This decision reflects the multi-label design of GoEmotions, where a text may express more than one emotion, and allows us to capture cases in which an emotion is a plausible secondary label rather than the single highest-ranked prediction. This is important for the present study because several reviews show small probability margins between the top-ranked labels.

Across the corpus, 22 of the 28 labels appear as top-1 predictions. The most frequent are *admiration* (30.6%), *neutral* (14.8%), *disappointment* (13.3%), *disapproval* (9.2%), *approval* (6.6%), and *love* (5.1%, $n = 10$). Among the theoretically relevant feminine-coded emotions discussed above, only *love* occurs often enough for quantitative analysis. It appears as the top-1 label in 10 reviews and as either the top-1 or top-2 label in 27 reviews.

We additionally inspect the star-rating distribution of the 27 love-labelled reviews. Most are positive: 18 are 5-star reviews and 4 are 4-star reviews. The remaining 5 have low ratings, but manual inspection shows that they still contain affectionate language directed at specific product features despite overall dissatisfaction. For example, one 1-star review states: *"I really wish this was not the case because I love the design, just be aware before you buy!"* This suggests that the *love* label captures genuine positive affect in the source text rather than merely reflecting overall review polarity.

As a plausibility check, we also re-annotate the reviews with GPT-4o (zero-shot, temperature 0), following recommendations that LLM-based annotation should be validated and interpreted on a task-specific basis (Pangakis et al., 2023). To reduce label ambiguity in this supplementary check, we prompt GPT-4o with a reduced 9-label scheme covering the most frequent emotion categories in our data: *admiration*, *approval*, *disapproval*, *disappointment*, *annoyance*, *joy*, *love*, *neutral*, and *other*. Since the classifier and GPT-4o use different label spaces, the resulting figures are not treated as inter-annotator agreement, but only as a secondary plausibility check.

Exact top-1 agreement between RoBERTa and GPT-4o is 32.0% (62/196), while soft agreement, where the RoBERTa top-1 label appears within GPT-4o's top-2 predictions, reaches 47.9% (93/196). Agreement at the broader polarity level is higher, at 63.4% (123/196). This suggests that the two systems diverge on fine-grained emotion labels but show greater convergence on the general affective direction of the reviews. The statistical analysis is based on the `SamLowe/roberta-base-go_emotions` classifier output.

4.3 Analysis

To examine the association between source-text emotion and grammatical gender in translation, we cross-tabulate the binary emotion flag (*love* present vs. absent) with the binary gender label (**m** vs. **f**) for each translator group separately. The emotion flag is set to positive if *love* appears as either the top-1 or top-2 label assigned by the classifier, yielding 27 reviews with a *love* signal out of 196

[5] `https://huggingface.co/SamLowe/roberta-base-go_emotions`

	N	+love		–love		OR	p	V
		f	m	f	m			
GPT-4	134	5	15	16	98	2.04	.313	.079
Llama	125	8	9	16	92	5.11	.005**	.251
Mistral	136	6	15	17	98	2.31	.201	.106
Prof.	124	9	7	27	81	3.86	.017*	.204
Stud.	118	9	7	18	84	6.00	.002**	.285

Table 2: Association between *love* and feminine grammatical gender. N = retained masculine/feminine cases; OR = odds ratio; p = two-tailed Fisher's Exact Test; V = Cramér's V. * $p < .05$; ** $p < .01$.

total.

For each 2×2 contingency table, we apply Fisher's Exact Test rather than Pearson's chi-square, because at least one expected cell count in each love/gender table falls below 5, the standard threshold for applying chi-square tests (Agresti, 1990). Fisher's Exact Test is therefore the appropriate alternative in this situation, providing exact p-values without relying on large sample approximations.

To complement significance testing with a measure of association strength, we additionally compute Cramér's V for each table. Cramér's V ranges from 0 (no association) to 1 (perfect association). By convention, values around 0.1 indicate a weak effect, 0.3 a moderate effect, and 0.5 or above a strong effect (Cohen, 1988).

We extend the same procedure to all eight emotions with sufficient combined frequency across top-1 and top-2 labels ($n \geq 10$): *admiration*, *approval*, *annoyance*, *disappointment*, *disapproval*, *joy*, *love*, and *neutral*. For each emotion, the appropriate test (Fisher's Exact or chi-square) is selected automatically based on the minimum expected cell count. This comparative analysis is used to determine whether any observed association is specific to *love* or reflects a broader relationship between emotion labels and grammatical gender.

5 Results

5.1 Love and Grammatical Gender

Table 2 presents the contingency counts, odds ratios, Fisher's Exact Test p-values, and Cramér's V effect sizes for the association between *love* and feminine grammatical gender across all five translator groups.

Across both human translator groups, *love* is significantly associated with feminine grammatical gender. Professional translators show a moderate effect ($V = .204$, $p = .017$), with 56.2% of love-labeled reviews receiving feminine gender compared to 25.0% of non-love reviews. Student translators show a comparable pattern with a slightly stronger effect ($V = .285$, $p = .002$), with the same 56.2% feminine rate for love reviews against 17.6% for non-love reviews. These findings extend the results reported by Popović and Lapshinova-Koltunski (2024).

The pattern is more nuanced across LLMs. GPT-4 and Mistral show no significant association between *love* and feminine gender ($p = .313$ and $p = .201$ respectively), with weak effect sizes ($V = .079$ and $V = .106$). Llama, however, shows a significant and moderately strong association ($V = .251$, $p = .005$), with 47.1% of *love*-labeled reviews receiving feminine gender compared to 14.8% of non-*love*-reviews – a pattern comparable in magnitude to the human translator groups. This finding was not anticipated and is discussed further in Section 6 below.

The comparison across the frequent emotion labels shows that no emotion other than *love* has a significant positive association with feminine gender. Two significant negative associations are observed: *neutral* is associated with fewer feminine translations in professional translators ($p = .020$, $V = .208$), and *disapproval* is associated with fewer feminine translations in both Llama ($p = .044$, $V = .168$) and Mistral ($p = .008$, $V = .200$). Since these associations are not consistent across translator groups and are not directly predicted by the gender-emotion literature reviewed above, we interpret them cautiously. The clearest positive association with feminine gender therefore remains specific to *love* in the present dataset.

As noted in Section 3 above, the previous study (Popovic and Lapshinova-Koltunski, 2024) identifies product category as a strong predictor of feminine gender in the corpus under analysis. To assess whether this factor might confound our primary finding, we conduct a descriptive inspection of the distribution of *love*-labeled reviews across the 14 product categories in the corpus, and of the gender choices made by human translators within those reviews. Of the 27 *love*-labeled reviews, only 5 fall within the two product categories identified by Popović and Lapshinova-Koltunski (2024) as most strongly associated with feminine gender in translation (beauty: $n = 1$; home & kitchen: $n = 4$). The remaining 22 reviews span ten categories not typically associated with feminine grammatical gender, including cell phones & accessories,

video games, CD/vinyl, and sports & outdoors. Among the 9 *love*-labeled reviews that receive feminine gender from professional translators, 6 came from categories outside the feminine-coded set (books, cd & vinyl, cell phones & accessories, grocery & gourmet food, toys & games). The same pattern holds for student translators, whose 9 feminine choices for *love*-labeled reviews also included 6 from non-feminine-coded categories. This descriptive pattern suggests that the feminine gender choices associated with *love*-labeled reviews are not concentrated in the product categories that would be expected to attract feminine forms on category grounds alone.

6 Discussion

Our results offer partial support for the assumption that source-text emotion might be associated with grammatical gender choices in translation. The emotion *love* is consistently associated with a higher rate of feminine grammatical gender in both human translator groups, with moderate effect sizes and p-values well below the conventional threshold. This pattern is consistent with well-documented gendered emotion stereotypes in psychology and NLP research.

The behaviour of the three LLMs presents a more complex pattern. With respect to the main *love*-feminine association, GPT-4 and Mistral show no significant effect, with weak effect sizes. This is consistent with the broader finding by Popović and Lapshinova-Koltunski (2024) that machine-generated translations exhibit stronger masculine bias than human translations, defaulting to masculine forms regardless of contextual cues. One plausible interpretation is that these models treat grammatical gender as a largely context-free decision, relying on the statistically dominant masculine form in Russian rather than on implicit contextual cues such as emotion. In this sense, GPT-4 and Mistral may produce less variation than human translators in under-specified first-person contexts.

Llama stands apart from the other two LLMs, showing a significant and moderately strong *love*-feminine association comparable in magnitude to the human translator groups. This was not predicted and we are cautious about over-interpreting it. The result is based on a small number of reviews ($n = 17$ *love*-labeled reviews after filtering), and replication on a larger dataset is necessary before any strong conclusions can be drawn. We can only speculate about possible explanations: differences in training data composition, instruction tuning. We flag this as a priority for future investigation.

7 Conclusion

Addressing RQ1, we find that the association between classifier-assigned emotion labels and grammatical gender is not general across all frequent emotion labels. The clearest positive association concerns *love*: love-labelled English source reviews are translated with feminine grammatical gender more often in both human translator groups, while other frequent emotion labels do not show the same positive pattern.

Addressing RQ2, this pattern differs across translator groups. The association is statistically significant and consistent in translations produced by both professional translators and translation students. Among the three LLMs, only Llama shows a comparable *love*–feminine association, while GPT-4 and Mistral do not.

Within the given limitations, the findings suggest that one theoretically feminine-coded emotion, *love*, may be associated with gendered translation choices in this dataset. The study therefore does not show that source-text emotion in general predicts grammatical gender, but rather that specific emotion labels can be productively examined as possible contextual cues. Future work should replicate these findings on larger corpora and additional language pairs, investigate why different LLMs respond differently to emotional cues, and include manual emotion annotation to validate automated classifier output.

8 Limitations

Several limitations should be noted when interpreting our results.

First, emotion labels are produced by an automated classifier and cannot be interpreted as ground-truth emotion. Fine-grained emotion classification remains a challenging task, with state-of-the-art performance on GoEmotions reaching only a moderate macro-F1 of 0.45 (Lowe, 2022). Class imbalance and inconsistent annotation of rare emotions in the underlying dataset are known issues (Demszky et al., 2020). Although we include a GPT-4o re-annotation as a plausibility check, this is not a substitute for independent human validation. A stronger design would include multiple hu-

man annotators and inter-annotator agreement.

Second, this study examines translation outputs only. We have no access to translator decision logs, think-aloud data, or post-hoc interviews. We therefore cannot determine whether the patterns reflect conscious choices, implicit bias, lexical cues, or something else entirely. This is a standard constraint of corpus-based research.

Third, the study is restricted to a single language pair (English–Russian), a single domain (Amazon product reviews), and an extremely small corpus. Generalisation to other genres, language pairs, and translator populations requires further study. The small number of emotion-labelled cases, especially for individual emotion categories, also means that statistically significant patterns should be interpreted cautiously.

Finally, the analysis is correlational. We identify statistical associations between source-text emotion and grammatical gender choices in translation, but do not claim causal relationships or attempt to model translator decision processes directly.

Taken together, these limitations position the present work as exploratory rather than confirmatory. Our contribution is a methodological proof-of-concept demonstrating that emotion classification can be productively combined with grammatical gender annotation to surface patterns in translation behaviour. The findings justify more rigorous follow-up with larger corpora, human validation, and richer translator metadata.

References

Agresti, Alan. 1990. *Categorical data analysis*. A Wiley-Interscience publication. Wiley, New York [u.a.].

Brody, Leslie R. and Judith A. Hall. 2008. Gender and emotion in context. In Lewis, Michael, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of Emotions*, pages 395–408. The Guilford Press, New York, 3 edition.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

Hackenbuchner, Janiça, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with MT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Lapshinova-Koltunski, Ekaterina, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 337–338, Ghent, Belgium, June. European Association for Machine Translation.

Lowe, Sam. 2022. roberta-base-go_emotions. `https://huggingface.co/SamLowe/roberta-base-go_emotions`.

Nesbitt, Muriel N. and Nolan E. Penn. 2000. Gender stereotypes after thirty years: A replication of rosenkrantz, et al. (1968). *Psychological Reports*, 87:493 – 511.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *ArXiv*, abs/2306.00176.

Plant, Ashby, Janet Hyde, Dacher Keltner, and Patricia Devine. 2000. The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24:81 – 92, 03.

Plaza-del-Arco, Flor Miriam, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand, August. Association for Computational Linguistics.

Popovic, Maja and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Rosenkrantz, Paul, Susan Vogel, Helen Bee, Inge Broverman, and Donald Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32:287–295, 06.

Savoldi, Beatrice, 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.

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

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. *ArXiv*, abs/2401.10016.

Weber, Sabine, Lynn Greschner, and Roman Klinger. 2026. Less is more? the role of demographic author information in emotion classification of ambiguous text. In Piperidis, Stelios, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8147–8161, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).