Uncertainty Quantification for Evaluating Machine Translation Bias

Anonymous ACL submission

Abstract

In machine translation (MT), when the source sentence includes a lexeme whose gender is not overtly marked, but whose target-language equivalent requires gender specification, the model must infer the appropriate gender from the context and/or external knowledge. Studies have shown that MT models exhibit biased behaviour, relying on stereotypes even when they clash with contextual information. We posit that apart from confidently translating using the correct gender when it is evident from the input, models should also maintain uncertainty about the gender when it is ambiguous. Using recently proposed metrics of semantic uncertainty, we find that models with high translation and gender accuracy on unambiguous instances do not necessarily exhibit the expected level of uncertainty in ambiguous ones. Similarly, debiasing has independent effects on ambiguous and unambiguous translation instances.[1]

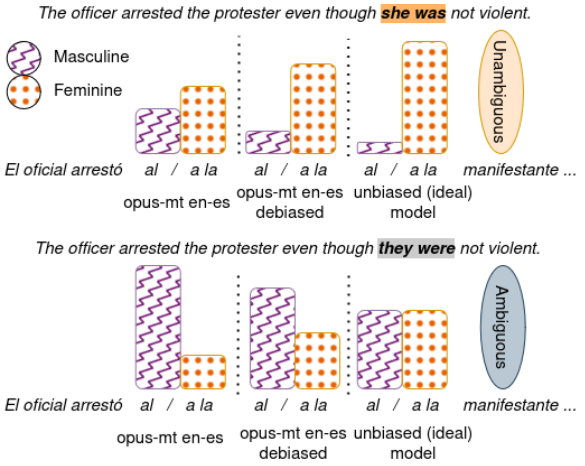

Figure 1: Probabilities for feminine and masculine determiners in a Spanish translation of a sentence containing a noun that is either feminine (referred to as 'she') or ambiguous ('they'), by two existing models and the ideal expected attribution of an unbiased model.

1 Introduction

Language is inherently ambiguous, and meaning is often resolved through context. However, not all ambiguity is resolvable (van Deemter, 1998). When humans process language, they draw on linguistic, cognitive, and social biases to arrive at an interpretation (Cairns, 1973). While linguistic biases ease cognitive processing, some can also have harmful effects, such as reinforcing existing social inequalities (Beukeboom, 2013). NLP models exhibit sensitivity to many of the human biases (Echterhoff et al., 2024), and even exaggerate them (Dhamala et al., 2021), as well as introduce additional ones (Tjuatja et al., 2024). When the input is unresolvably ambiguous, favoring a single output necessarily relies on biases. Therefore, a well-designed model should refrain from making a single prediction, instead requesting clarification or generating multiple alternative outputs.

Most studies on decoding with language models (LMs) for machine translation (MT) evaluate a single prediction, usually generated with beam search, per translation instance. Consequently, most studies on uncertainty quantification (UQ) use uncertainty to predict the quality of the translation recovered by beam search (Fomicheva et al., 2020; Cheng and Vlachos, 2024). Previous work on bias in MT has focused on LM performance against gold standard labels (Stanovsky et al., 2019), and previous work on ambiguity in MT has likewise focused on resolvable cases (Barua et al., 2024; Martelli et al., 2025). Thus, most work on uncertainty and ambiguity assumes that there is a single correct translation per instance. However, less attention has been given to ambiguous source sentences, where the choice of the LM cannot be guaranteed to be correct without additional context. This may be seen as a form of aleatoric uncertainty (Hora, 1996), uncertainty which is inherent in the data and irreducible. According to Baan

[1]The code will be available at https://anonymous.4open.science/r/uncertainty_bias_ambiguity-8A4C/

et al. (2024), the spread of probability mass in LMs represents both lack of confidence as well as variation in human generations. We adopt this line of thinking and examine whether LMs accurately and fairly represent the range of possible translations for ambiguous source sentences.

In this work, we leverage distribution-level uncertainty metrics to evaluate cases where the model *should not* be certain about its predictions due to ambiguity in the input. Figure 1 shows two versions of a sentence, with unambiguous (top) or ambiguous (bottom) gender of the noun '*protester*', and the different probabilities assigned to the Spanish translations of the determiner of this noun. An ideal model should assign a higher probability for the feminine determiner ('*la*') when the gender is disambiguated by the pronoun, and produce equal probabilities for masculine and feminine translations when the gender is ambiguous. However, state-of-the-art MT models, including debiased ones, tend to produce more uniform probability distributions for unambiguous inputs, and less uniform distributions for ambiguous ones. This indicates that the model probabilities are influenced by stereotypical associations between protesting and masculinity, causing the models to default to the masculine form even when no gender preference is warranted, and to select the feminine form with low confidence despite clear contextual cues.

To study this systematically, we focus on translating sentences from a language that does not mark gender in nouns and verbs (English) into languages that do (Spanish, French, Ukrainian, and Russian). We use the WINOMT dataset (Stanovsky et al., 2019), which includes stereotypical gender roles, and extend it with manual translations and automatic annotations of additional cognitive bias cues, such as implicit causality verbs. In some cases, the gender is resolvable from context, while in others it is not. We explore different variants of Semantic Uncertainty metrics (Cheng and Vlachos, 2024; Farquhar et al., 2024) to quantify the semantic diversity of translation samples, finding that these metrics effectively capture the variation in gender caused by bias triggers. We validate the metrics against the established gender accuracy metric and account for the effects of translation accuracy. Our main findings are: 1) stereotypes and linguistic biases influence gender translation, 2) the degree of bias corresponds to overall model translation accuracy in unambiguous cases, 3) the degree of bias corresponds to translation accuracy at the instance level in ambiguous cases, 4) debiasing effects vary depending on input ambiguity, translation accuracy, and target language.

2 Related Work

Researchers have addressed various biases in MT, including algorithmic bias (Vanmassenhove et al., 2021), as well as gender, number and formality biases (Měchura, 2022). Gender stereotypes are triggered not only by semantic content, but also by speech mannerisms (Dawkins et al., 2024) and person names (Saunders and Olsen, 2023). Existing solutions for gender translation in both unambiguous (Robinson et al., 2024) and ambiguous cases (Cho et al., 2019; Gonen and Webster, 2020; Vanmassenhove and Monti, 2021) rely on tools or human annotation of gender, limiting their generalisability to other types of ambiguities. For example, Cho et al. (2019) explore unresolvable ambiguity by generating multiple translations for a sentence with an ambiguous pronoun, however, it is restricted to specific sentence types.

In NLP and ML research, methods for distinguishing aleatoric and epistemic uncertainty have been proposed (Hou et al., 2024), however they do not distinguish between data randomness and data ambiguity. Some work has made the link between biases and uncertainty (Sicilia et al., 2024; Kuzucu et al., 2025), as well as between ambiguity and uncertainty (Kim, 2025; Cheng and Amiri, 2024), however, these papers construe uncertainty as a signal for poor performance and ambiguity as low quality inputs, which differs from our definition of ambiguity as an indispensable feature of language. Work in Question Answering has found that the best methods for detecting ambiguous inputs involve quantifying repetition within sampled model outputs (Cole et al., 2023), and using white-box metrics such as entropy (Yang et al., 2025).

UQ in MT has been used as a proxy for Quality Estimation (QE). For example, Fomicheva et al. (2020) use MT model uncertainty to estimate translation quality without references, while Glushkova et al. (2021) apply the same technique to the uncertainty of the QE models themselves. Other approaches use UQ to identify difficult instances and enhance training by applying curriculum learning (Zhou et al., 2020), semantic augmentation (Wei et al., 2020), balancing of multilingual training data (Wu et al., 2021) or test-time adaptation (Zhan et al., 2023). Wang et al. (2024b) examine zero-

shot translation and distinguish between model uncertainty and data uncertainty, however their focus with regard to data uncertainty is on noisy, low-quality training data rather than inherent ambiguity. Cognitive science research has demonstrated that entropy-based uncertainty metrics are a suitable measure of ambiguity in human translations (Bangalore et al., 2016), but this insight has yet to be applied in MT. To the best of our knowledge, no prior UQ-based approach in MT has explored ambiguity as a particular type of data uncertainty.

3 Method

We propose to quantify gender bias in Neural Machine Translation (NMT) models by characterising how gender is assigned to nouns across the predictive distribution. To do so, we base our methods on recently proposed UQ metrics which are founded on the classic Shannon entropy but take into account similarities between random Monte Carlo samples from the model. We first provide a brief overview of these UQ methods.

Let $\mathcal{Y}$ be a random variable whose value is drawn from the predictive distribution of an NMT model $p(y|x)$. Then entropy is defined as:

$$\mathcal{H}(\mathcal{Y}) = \mathop{\mathbb{E}}_{y\sim\mathcal{Y}}\left[I(y)\right],$$

where I is the *surprisal* of y. In the classic Shannon entropy, $I = -\log p(y)$, but the UQ methods we consider vary in their definition of surprisal.

Semantic Entropy (SE; Farquhar et al., 2024) identifies semantic equivalences between elements and clusters them together according to a textual entailment model, mapping each y to a cluster c. In our implementation we use a multilingual mDeberta model (He et al., 2021) finetuned on the Natural Language Inference (NLI) task by Laurer et al. (2022). Then, surprisal is the negative log probability of an element being in c:

$$I_{\text{SE}}(y) = -\log \mathop{\mathbb{E}}_{y'\sim\mathcal{Y}} 1\left[y' \in c\right].$$

Similarity-sensitive Shannon Entropy (S3E; Ricotta and Szeidl, 2006; Cheng and Vlachos, 2024) sets the surprisal of y to the negative log of its expected similarity with all other outputs:

$$I_{\text{S3E}}(y) = -\log \mathbb{E}_{y'\sim\mathcal{Y}}\left[\mathcal{S}(y,y')\right],$$

where $\mathcal{S}$ is a similarity function satisfying $\mathcal{S}(y,y') \in [0,1]$ and $\mathcal{S}(y,y') = 1$ if $y = y'$. Following Cheng and Vlachos (2024), we use cosine similarity between sentence embeddings of y and y' generated by a multilingual E5 text embedding model (Wang et al., 2024a).

We also define Gender Entropy (GE), which is calculated like SE but clusters elements based on the gender class of the translated focus noun. To determine the gender class, we use Spacy[2] and pymorphy2 (Korobov, 2015) morphological parsers.

SE, S3E, and GE must be approximated with random sampling from $p(\cdot|x)$, which we perform with ϵ-sampling (Hewitt et al., 2022), drawing 128 samples per source sentence. Further details about these UQ methods are given in Appendix A.

Our gender bias metrics are based on surprisal and entropy given by these UQ methods. The first desideratum is that for source sentences with unambiguous gender, an unbiased model should have lower surprisal of a translation with correct gender inflection compared to an incorrect inflection. Therefore, unbiased models should minimise *relative surprisal*, defined as:

$$\Delta I = \frac{I(y_{\text{correct}}) - I(y_{\text{incorrect}})}{\frac{1}{2}(I(y_{\text{correct}}) + I(y_{\text{incorrect}})}.$$

The second desideratum is that the entropy of unbiased models should not be affected by the presence of bias cues. Thus, we define *normalised entropy*, which compares the $\mathcal{H}$ of a source sentence x to the average entropy across its *contrast set* $\mathcal{G}_x$. $\mathcal{G}_x$ is a group of minimally different sentences that are identical to x apart from the pronoun (e.g., '*she*', '*he*', '*they*'), including x itself. The three sentences in Table 1 comprise a $\mathcal{G}_x$. Formally:

$$\text{norm-}\mathcal{H}(x) = \frac{\mathcal{H}(\mathcal{Y}_x)}{\frac{1}{|\mathcal{G}_x|}\sum_{x'\in\mathcal{G}_x}\mathcal{H}(\mathcal{Y}_{x'})},$$

This formulation isolates variation in $\mathcal{H}(\mathcal{Y}_x)$ attributable specifically to gender by holding all other lexical, syntactic, and semantic content constant across the contrast set.

The third desideratum is that models should show higher uncertainty for an input which is ambiguous with regard to gender, as compared to an input which is unambiguous, disregarding all biases in the input. It should therefore minimise *relative entropy*, defined as:

$$\Delta\mathcal{H} = \frac{\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) - \mathcal{H}(\mathcal{Y}_{\text{ambiguous}})}{\frac{1}{2}\left(\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) + \mathcal{H}\left(\mathcal{Y}_{\text{ambiguous}}\right)\right)}.$$

[2] https://spacy.io/

Sentence	Pronoun	Stereotype	Subject	Recency	IC
The **mechanic** called to inform someone that *he* had completed the repair.	M	M	M	N	N
The **mechanic** called to inform someone that *she* had completed the repair.	F	M	F	N	N
The **mechanic** called to inform someone that *they* had completed the repair.	N	M	N	N	N

Table 1: WINOMT (Stanovsky et al., 2019) examples with additional annotations of bias cues.

4 Experimental setup

To test our proposed bias metrics, an MT dataset containing information about gender ambiguity and stereotypes in the source sentences is required. We use WINOMT (Stanovsky et al., 2019), which includes annotations on minimal pairs of 1,584 sentences with masculine, feminine or neutral pronouns referring back to stereotypical or anti-stereotypical gender roles. An example of three sentences from the dataset can be seen in Table 1. The gender of the focus noun '*mechanic*' is unambiguous in the first two sentences based on the contextual information (Pronoun M & F), but remains ambiguous in the third on account of the neutral (N) pronoun. The Stereotype (column 3) that mechanics are more often men than women either contradicts (row 1) or aligns with (row 2) the disambiguating context (column 2).

This dataset also contains other linguistic phenomena that were not explicitly annotated in its released version. Thus, we automatically annotate additional linguistic bias cues, namely subject, recency, implicit causality, and person names, using syntactic parses with Spacy[2]. We release the additional annotations to the public for reproducibility.

For the Subject bias, following the literature on human biases for coherence (Nieuwland and Van Berkum, 2006), we hypothesise that models may assume coreference between the subject of the main clause (often the focus noun) and the subject of the complement clause (often the pronoun). In the example in Table 1 '*the mechanic*' is the subject, therefore the Subject bias primes an interpretation in which the gender of the subject aligns with that of the subsequent pronoun (M in row 1, F in row 2). Furthermore, person names have been shown to have a strong effect on pronoun resolution (Saunders and Olsen, 2023). To assess the impact of person names on gender translation, we augment the dataset with common feminine and masculine names matching the gender of the pronoun, selected for their cross-linguistic familiarity (see Appendix B). As an example, when translating into French, the second sentence from Table 1 would read "*The mechanic Anne called to inform someone that she had completed the repair.*"

Recency bias elicits the attribution of the gender of the most recent noun phrase to the following pronoun (Gautam et al., 2024). However, in our example, the most recent noun phrase is '*someone*', so the focus noun is unaffected. The implicit causality (IC) bias induces an expectation in humans and LMs that when an IC verb is followed by a causal connective, the following pronoun will corefer with either the Subject or the Object of the IC verb, depending on the causal inference (Kementchedjhieva et al., 2021). For instance, a clause with the verb '*call*', when followed by an explanation starting with '*because*', is expected to be followed by a mention of the *caller* rather than the *callee*. We use an IC verb corpus (Garnham et al., 2021) for the annotation and find that about 10% of the dataset contains IC structures. The sentences in Table 1 do not include a causal complement, hence the IC bias does not apply.

Target Languages We selected target languages that express gender through morphological markers on nouns and adjectives, and sometimes verbs, namely Spanish, French, Ukrainian, and Russian. The target languages vary in their representation within NLP research. A relevant idiosyncrasy of the Russian language is that for some nouns describing professions, even if a feminine version exists, it may be considered derogatory in use (Komova, 2024). For example, 'врач' is the masculine term for '*doctor*', and the alternative feminine 'врачиха' is considered rude, which leads to 'врач' being used even when the doctor is known to be a woman. This results in constructions where masculine nouns are paired with feminine verb forms, or masculine markers are used throughout the sentence. In order to account for this, we also include a classification of professions which adhere to such (lack of) gender marking into WINOMT, using data from Komova (2024).

Human Translations WINOMT does not contain target translations, thus the accuracy of MT models cannot be directly evaluated. To overcome this, we hired professional translators to translate a set of 100 WINOMT sentences into French, Span-

ish, Ukrainian and Russian. Each sentence is translated twice, with the focus noun in feminine and masculine variants respectively. They also annotate the translations as Correct or Incorrect with regard to the gender translation in the given context. For instance, when translating the English sentence "*The farmer bought a book from the **writer** and paid her*" into French, where '*writer*' is the focus noun, the feminine '*l'auteure*' should be marked as Correct, while '*l'auteur*' would be Incorrect. In ambiguous cases, i.e. if the pronoun in the above sentence was '*they*', both gender translations would be Correct. Appendix C provides the translation guidelines and details, and Appendix D discusses the quality of human annotations. We release the translations and correctness annotations to the public to enable further research.

Models We experiment with two commonly used translation models, namely OPUS-MT (Tiedemann and Thottingal, 2020) and M2M100 (Fan et al., 2021). OPUS-MT models are NMT models trained on freely available parallel corpora. M2M100 is a many-to-many multilingual translation model, which directly translates between any pair of 100 languages. To examine how effective debiasing is regarding the three desiderata stated in Section 3, we apply the hard-debiasing method from Iluz et al. (2024) on the OPUS-MT models, which have been shown to lower bias scores on the WINOMT dataset (Stanovsky et al., 2019) while maintaining translation quality. The hard-debiasing method neutralises the biased words in the representation space, so that neutral words are not associated with a specific gender (Bolukbasi et al., 2016). We adopt the most effective debiasing approach from Iluz et al. (2024), which applies debiasing to one-token profession words on the encoder side. See Appendix E for the performance of all models.

5 Research questions

Does Semantic Uncertainty Capture Gender Bias? To validate the application of UQ metrics for bias evaluation, we compare their scores with the established gender accuracy metric. Gender accuracy uses the morphological parsers described in Section 3 to determine the focus noun gender in translations. As it relies on gold-standard references, it is applicable only to unambiguous items and unsuitable for cases with multiple valid gender realisations. We therefore limit this experiment to unambiguous items. We rank all models according to their ΔI scores and compare this ranking to that based on gender accuracy using Kendall's τ and Pearson's r. In order to establish whether a sampling-based metric is necessary, we also test a simple $\Delta LogProb$ value, which compares the Log Probabilities assigned to correct and incorrect instances, as an alternative to ΔI.

What Biases do Models Exhibit via Uncertainty? To evaluate model bias, we assess how bias cues influence the diversity of gender markers in translations. Specifically, we perform an analysis of variance (ANOVA) to examine the effect of bias cues (independent variables) on normalised entropy measures norm-$\mathcal{H}$ (S3E), norm-$\mathcal{H}$ (SE), and norm-$\mathcal{H}$ (GE) (dependent variables), using T-tests for significance. This analysis incorporates linguistic biases not previously explored in WINOMT.

What does Semantic Entropy Reveal about Bias with Ambiguity? To evaluate model bias in ambiguous settings, which has not yet been explored, we compare models using their $\Delta\mathcal{H}$ scores. To isolate the uncertainty caused by ambiguity from that resulting from poorer model performance, we analyse the relationship between the $\Delta\mathcal{H}$ scores and the translation quality measured by the COMET metric (Rei et al., 2022). Since the WINOMT dataset does not contain gold target translations, we use the 100 professionally annotated items described in Section 4, and the WMT test sets which contain the target languages (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024) for translation quality evaluation.

6 Results

This section presents the results of the bias evaluation using semantic uncertainty metrics.

Lang.	Model	Gender Acc	ΔLog prob	ΔI (S3E)	COMET
ES	OPUS-MT	67.95	0.00	-0.10	84.90
	deb-OPUS-MT	68.13	0.00	-0.13	84.86
	M2M100	70.77	0.00	-0.13	72.05
FR	OPUS-MT	64.27	0.01	-0.04	83.56
	deb-OPUS-MT	64.79	0.01	-0.08	83.55
	M2M100	61.66	0.01	-0.07	73.06
UK	OPUS-MT	45.34	0.00	-0.03	70.79
	deb-OPUS-MT	46.12	0.00	-0.03	70.79
	M2M100	47.76	0.00	-0.02	52.85
RU	OPUS-MT	48.57	0.00	0.00	79.37
	deb-OPUS-MT	48.42	0.00	-0.03	79.36
	M2M100	48.49	0.00	-0.03	58.62

Table 2: Gender Accuracy, Δ Log Probability, and ΔI (S3E) on Unambiguous instances, COMET scores on WMT test sets (see Appendix E for details).

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Pronoun				Default M		Ambiguity
			F	M	SF	SM	OF	OM	SF	SM	OF	OM	F	M	SF	SM	OF	OM	S	O	
es	OPUS-MT	**0.41**	**0.41**	-0.05	**0.25**	**-0.19**	**0.24**	**-0.33**	0.06	0.10	**0.13**	**0.17**	**0.38**	**-0.21**	**0.24**	**-0.31**	**0.29**	**-0.13**	N/A	N/A	**-0.18**
	deb-OPUS-MT	**-0.05**	**0.30**	**-0.11**	**0.27**	**-0.20**	**0.16**	**-0.38**	0.05	**0.14**	**0.08**	0.05	**0.49**	**-0.14**	**0.39**	**-0.24**	**0.14**	**-0.21**	N/A	N/A	**-0.10**
	M2M100	**0.14**	**0.33**	**-0.11**	**0.29**	**-0.42**	**0.28**	**-0.25**	**0.18**	**0.27**	**0.12**	0.07	**0.51**	-0.04	**0.52**	**-0.08**	0.04	**-0.35**	N/A	N/A	**-0.11**
fr	OPUS-MT	**0.54**	**0.42**	**0.16**	0.05	**-0.34**	0.06	**-0.24**	**0.43**	**0.45**	**0.26**	**0.21**	**0.16**	**-0.14**	**0.16**	**-0.17**	**0.20**	-0.04	N/A	N/A	**-0.29**
	deb-OPUS-MT	**0.19**	**0.22**	0.02	0.05	**-0.25**	**0.17**	**-0.11**	**0.23**	**0.32**	**0.23**	**0.13**	**0.24**	-0.03	**0.31**	**-0.08**	0.01	**-0.14**	N/A	N/A	**-0.12**
	M2M100	**-0.12**	0.11	**-0.23**	**0.50**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.70**	**0.31**	**0.47**	0.05	**-0.08**	**-0.40**	N/A	N/A	0.06
uk	OPUS-MT	**-0.27**	0.00	**-0.11**	**0.26**	-0.02	**0.19**	**0.13**	-0.11	**0.17**	0.03	**-0.14**	**0.27**	**0.21**	**0.22**	**0.13**	**-0.11**	**-0.23**	N/A	N/A	0.06
	deb-OPUS-MT	**-0.43**	-0.06	**-0.12**	**0.41**	0.00	**0.18**	0.07	**-0.24**	0.01	0.01	**-0.17**	**0.23**	**0.17**	**0.16**	**0.10**	**-0.11**	**-0.17**	N/A	N/A	**0.09**
	M2M100	**0.08**	-0.04	**-0.18**	**0.27**	0.02	**0.33**	**0.18**	**0.19**	**0.37**	-0.03	**-0.17**	**0.53**	**0.34**	**0.50**	**0.26**	**-0.30**	**-0.42**	N/A	N/A	**0.11**
ru	OPUS-MT	0.01	**-0.28**	**-0.41**	0.00	-0.12	0.08	**-0.10**	**-0.19**	**-0.20**	**-0.41**	**-0.39**	**0.07**	-0.03	**0.14**	0.03	**-0.10**	**-0.24**	**-0.40**	0.04	**0.35**
	deb-OPUS-MT	**0.23**	**-0.10**	**-0.17**	0.05	0.03	0.05	-0.03	**-0.10**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**-0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	0.00	**0.13**
	M2M100	**-0.74**	**-0.16**	**-0.21**	0.10	-0.04	**-0.12**	**-0.25**	**0.12**	**0.12**	**-0.20**	**-0.18**	**0.06**	-0.02	**0.08**	0.01	**-0.07**	**-0.11**	**-0.20**	**-0.22**	**0.18**

Table 3: ANOVA results: single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on norm-$\mathcal{H}$ (S3E). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

Model Rankings According to Semantic Surprisal and Gender Accuracy Correlate. The results of the first experiment are presented in Table 2. We find that while ΔLog prob does not correlate with the gender accuracy ranking, ΔI (S3E) shows a statistically significant negative correlation with gender accuracy (Kendall's $\tau = -0.58$, Spearman's $\rho = -0.78$; see Appendix K). We attribute the effectiveness of this metric in distinguishing Correct from Incorrect gender translations to its ability to capture gender information beyond noun morphology, including verb inflections and agreement, through its embedding representations. The flexibility of S3E enables it to encode nuances that GE does not. For example, when translating a sentence with a feminine pronoun into Russian, OPUS-MT generates sentences with a masculine noun and a verb that is either feminine or masculine (e.g., "Перевозчик поблагодарила (fem) / поблагодарил (masc)" "*The courier thanked*"). This is reflected in a higher $\mathcal{H}$ score by S3E (0.65) than GE (0.00), as the variation in verb inflections is only captured by S3E.

The strong negative correlation between ΔI (S3E) and gender accuracy thus validates the core component of our proposed metric for evaluating bias in machine translation.

In addition, the ranking of models by their overall performance, as indicated by COMET scores, is partially aligned with the rankings based on Gender Accuracy and ΔI for unambiguous instances. This suggests that in these instances, better performing models tend to be less biased (all rankings are listed in Appendix J).

Semantic Entropy Scores Vary With Regard to Bias Cues. The results of the second experiment presented in Table 3 show that most bias cues in the data have a significant effect on the variance of norm-$\mathcal{H}$ (S3E), indicating that the tested models exhibit various social and linguistic biases.[3] The results corroborate previous findings. The high absolute coefficient values in the Names column indicate that person names have an effect on gender translation even when a disambiguating pronoun is present. This is in line with Saunders and Olsen (2023), who have shown that both pronouns and names induce gender bias and are often not sufficient for full disambiguation. Secondly, the fact that some Russian nouns have a default masculine grammatical gender regardless of the context (Komova, 2024) is reflected in significant decrease in gender diversity for sentences containing such nouns (negative coefficients in the Default M columns indicating lower norm-$\mathcal{H}$). Thirdly, we observe that masculine biases generally reduce norm-$\mathcal{H}$ (negative coefficients in the M columns), while feminine biases tend to increase it (positive coefficients in the F columns). This suggests a general default toward masculine translations in models, with outputs becoming more similar under masculine biases and more varied under feminine ones. This finding aligns with Kuzucu et al.

[3] norm-$\mathcal{H}$ (S3E) shows the strongest sensitivity to bias cues compared to norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE). We also experiment with unnormalised $\mathcal{H}$ scores, the results of which are less comparable across metrics, bias types, and models. The full results are presented in Appendix F. All trends observed for norm-$\mathcal{H}$ (S3E) are also present in norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE), as well as for unnormalised $\mathcal{H}$.

(2025), who show that model uncertainty is typically higher for minority groups.

Translation Accuracy Affects the Bias–Entropy Relationship Differently Across Levels of Analysis. Having validated the S3E metric in the first experiment, we investigate the results of $\Delta\mathcal{H}$ (S3E) as a bias metric. Table 4 illustrates that some models (OPUS-MT-UK, deb-OPUS-MT-UK, M2M100-UK, OPUS-MT-RU, M2M100-RU) exhibit the desired negative $\Delta\mathcal{H}$. Surprisingly, this result suggests that when it comes to ambiguous instances, contrary to the unambiguous cases, the models which perform better on translation accuracy (namely models for Spanish and French) are not generally less gender biased (model rankings according to all metrics are available in Appendix J). This finding mirrors the results in the Ambiguity column in Table 3, where norm-$\mathcal{H}$ increases for Ukrainian and Russian (positive coefficients), but not for Spanish and French. Higher norm-$\mathcal{H}$ for Ambiguous items (or negative $\Delta\mathcal{H}$) is expected for an unbiased model.

Lang.	Model	Unamb	Amb	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09
	deb-OPUS-MT	0.97	0.89	0.08
	M2M100	1.79	1.45	0.19
FR	OPUS-MT	1.79	1.43	0.20
	deb-OPUS-MT	1.21	1.08	0.11
	M2M100	3.22	2.78	0.14
UK	OPUS-MT	1.96	2.16	-0.10
	deb-OPUS-MT	1.98	2.15	-0.09
	M2M100	2.05	2.28	-0.11
RU	OPUS-MT	1.56	1.68	-0.08
	deb-OPUS-MT	1.05	0.97	0.08
	M2M100	1.83	2.29	-0.25

Table 4: Unambiguous and Ambiguous $\mathcal{H}$ (S3E)

In contrast, we observe the expected effect of debiasing on the Spanish and French models (lower $\Delta\mathcal{H}$ for deb-OPUS-MT in Table 4), suggesting that models which perform better overall are more susceptible to debiasing. The impact of debiasing is also reflected in Table 3, as the effects are mostly smaller (lower absolute values of coefficients) in debiased models compared to their non-debiased counterparts across languages, confirming that debiasing is at least partially effective.

In Figure 2, results are grouped by COMET score bins for a more fine-grained analysis at the instance level. For the models with negative $\Delta\mathcal{H}$ (S3E) scores (Ukrainian and Russian), $\Delta\mathcal{H}$ is typically most pronounced for the highest-accuracy translations (e.g. ambiguous scores for M2M100-RU in Bin 3 are substantially higher than B1). Although debiasing does not reduce the overall $\Delta\mathcal{H}$ score for Ukrainian (see Table 4), it results in the largest improvement in the highest-quality translations: in bin B3, ambiguous $\mathcal{H}$ scores for deb-OPUS-MT-UK are notably higher than those of the original model. This improvement is further supported by a substantial 8.41% drop in masculine focus noun inflections for Ukrainian, compared to 0.88–2.49% for other languages. We hypothesise that this is due to the limited training data in Ukrainian, which may lead to a less stable model that performs worse overall but is more responsive to debiasing in high-quality outputs. The relationship between translation accuracy and bias under ambiguity appears to differ depending on the level of analysis: *between* models, higher accuracy does *not* imply lower bias on ambiguous instances, whereas *within* models, higher-accuracy instances tend to show *lower* bias.

Qualitative Analysis A qualitative analysis of the example in Table 1, presented with corresponding $\mathcal{H}$ values in Table 5, corroborates the quantitative findings in Figure 2. When translating the sentence in row 1, across all target languages, OPUS-MT models consistently produce only the masculine variants of the focus noun ('*El mecánico*', '*Le mécanicien*', 'Механик' and 'Механік'). In the anti-stereotypical case (row 2), all languages except Russian include both masculine and feminine forms ('*La mecánica*', '*La mécanicienne*' and 'Механіка'), indicating that these models are sensitive to the masculine stereotype even when the referent in the context is clearly feminine. For Russian, the models fail to generate any feminine constructions, even when the context is unambiguously feminine. This difference is evident in the $\mathcal{H}$ (S3E) scores, which are higher in row 2 than row 1 for the first three languages. Moreover, in the ambiguous case (row 3), all OPUS-MT models produce only masculine nouns, regardless of language. Consequently, $\mathcal{H}$ (S3E) scores are generally higher for the unambiguous cases (mean of rows 1 and 2) than for the ambiguous case (row 3), except for Russian, where $\mathcal{H}$ remains low across all conditions. These observations are consistent with expectations for biased models: they default to stereotypical gender realisations when the pronoun is ambiguous and sometimes even when the context clearly suggests an anti-stereotypical interpretation.

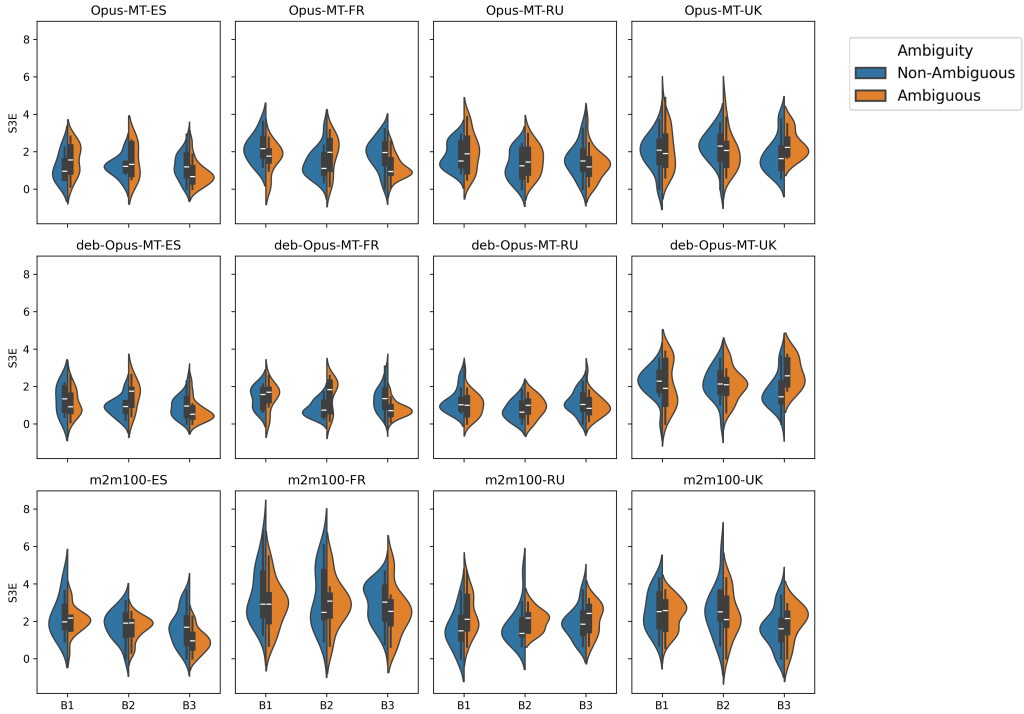

Figure 2: Violin plots of binned COMET scores and $\mathcal{H}$ (S3E) on ambiguous and unambiguous inputs. Low, medium and high COMET scores from left to right, evaluated with human translations, multi-reference for ambiguous items.

Sentence	**ES** → deb		**FR** → deb		**UK** → deb		**RU** → deb	
The **mechanic** called to inform someone that *he* had completed the repair.	0.75	0.82	0.00	0.00	2.41	1.57	0.33	0.31
The **mechanic** called to inform someone that *she* had completed the repair.	1.64	1.85	0.53	0.56	2.75	2.05	0.00	0.00
The **mechanic** called to inform someone that *they* had completed the repair.	0.74	0.87	0.00	0.00	2.38	2.03	0.37	0.37

Table 5: WINOMT examples with $\mathcal{H}$ (S3E) values, for OPUS-MT (left) and deb-OPUS-MT (right) models.

The qualitative analysis also reveals interesting effects of debiasing. In the anti-stereotypical case in row 2, debiasing increases the number of feminine constructions generated in Spanish (from 43/128 to 55/128, corresponding to a slight increase in $\mathcal{H}$, as feminine forms remain a minority) and Ukrainian (from 72/128 to 128/128, reflected in a decrease in $\mathcal{H}$). No notable changes are observed for French or Russian, consistent with stable $\mathcal{H}$ (S3E) scores. When it comes to the ambiguous pronoun (row 3), the debiased models continue to generate only masculine variants of '*mechanic*' in Spanish, French and Russian, with $\mathcal{H}$ remaining largely unchanged. In contrast, all debiased model outputs in Ukrainian include a feminine translation of the noun, corresponding to a decrease in $\mathcal{H}$ from OPUS-MT to debiased OPUS-MT in row 3. This pattern illustrates that when debiasing leads to overgeneration of feminine morphology in ambiguous contexts, our proposed metric flags this as increased bias (positive $\Delta\mathcal{H}$), indicating that such changes are not deemed as improvements.

7 Conclusion

In this work, we apply distribution-level UQ to evaluate bias in MT models. This method complements gender accuracy, particularly where gender accuracy is inapplicable. Specifically, it captures the more subtle manifestations of gender bias that arise when models show a preference for one gender in ambiguous contexts. Our overall contribution is the novel use of UQ as a bias metric in MT, which 1) does not rely on gender references, 2) is general and captures multiple types of bias, 3) is validated by the established metric of gender accuracy, and 4) provides new insights into biased behavior in ambiguous contexts, a setting not previously studied. Future work will extend the proposed bias evaluation method to tasks beyond translation.

Limitations

This study is limited to Romance and Slavic languages, not including many other language families which mark gender and express stereotypes in diverse ways. While we tried to account for language differences by including different names for different target languages, accounting for specific masculine-only nouns in Russian, debiasing with language-specific vocabularies, etc., some linguistic idiosyncracies are still not accounted for, such as the fact that profession stereotypes are defined in English and may apply differently in different regions. Finally, our work is limited to two grammatical genders, and treats 'they' as a neutral pronoun that may refer to any gender, however we do not study the interpretation of the pronoun as referring to non-binary people specifically. Further directions include applying UQ to ambiguity detection, which could enable more gender-inclusive translations through morphological doubling, where both masculine and feminine morphemes are included for gender neutrality. Future work should address these directions.

Ethics Statement

The models used in this study, like all ML models, can be biased as well as make mistakes, including in gender attribution. Our contribution aims to specifically tackle masculine and feminine gender stereotypes via more stringent evaluation metrics, in order to avoid the perpetuation of gender bias.

Acknowledgments

References

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2016. *Syntactic Variance and Priming Effects in Translation*, pages 211–238. Springer International Publishing, Cham.

Josh Barua, Sanjay Subramanian, Kayo Yin, and Alane Suhr. 2024. Using language models to disambiguate lexical choices in translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4837–4848, Miami, Florida, USA. Association for Computational Linguistics.

Camiel J Beukeboom. 2013. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, pages 313–330.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Helen S Cairns. 1973. Effects of bias on processing and reprocessing of lexically ambiguous sentences. *Journal of Experimental Psychology*, 97(3):337.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. 2012. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein.

2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Alan Garnham, Svenja Vorthmann, and Karolina Kaplanova. 2021. Implicit consequentiality bias in english: A corpus of 300+ verbs. *Behavior Research Methods*, 53:1530–1550.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2024. *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics, Miami, Florida, USA.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2):217–223. Treatment of Aleatory and Epistemic Uncertainty.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bar Iluz, Yanai Elazar, Asaf Yehudai, and Gabriel Stanovsky. 2024. Applying intrinsic debiasing on downstream tasks: Challenges and considerations for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14914–14921, Miami, Florida, USA. Association for Computational Linguistics.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Hazel H. Kim. 2025. How ambiguous are the rationales for natural language reasoning? a simple approach to handling rationale uncertainty. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10047–10053, Abu Dhabi, UAE. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors. 2023. *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.

Liliana Komova. 2024. *Gender, Language and Perception: Linguistic Inclusivity in Russian*. Ph.D. thesis, Università Ca'Foscari Venezia.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. 2025. Uncertainty as a fairness measure. *Journal of Artificial Intelligence Research*, 81.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J.A. Van Berkum. 2006. Individual differences and contextual bias in pronoun resolution: Evidence from erps. *Brain Research*, 1118(1):155–167.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: Bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243.

Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. MiTTenS: A dataset for evaluating gender mistranslation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4115–4124, Miami, Florida, USA. Association for Computational Linguistics.

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93, Tampere, Finland. European Association for Machine Translation.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2024. Accounting for sycophancy in language model uncertainty estimation. *arXiv preprint arXiv:2410.14746*.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Kees van Deemter. 1998. Ambiguity and idiosyncratic interpretation. *Journal of Semantics*, 15(1):5–36.

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

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R. Lyu. 2024b. Understanding and mitigating the uncertainty in zero-shot translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:4894–4904.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2724–2735, Online. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–820, Toronto, Canada. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

A Further UQ discussion

Farquhar et al. (2024) does not define a per-element surprisal; the original definition computes Shannon entropy over clusters. $\mathcal{Y}$ are mapped to clusters $\mathcal{C}$, and SE is:

$$\mathcal{H}_{\text{SE}}(\mathcal{C}) = -\mathop{\mathbb{E}}_{c\sim\mathcal{C}} \log p(c|x).$$

Surprisal is thus defined for a cluster instead of an element, but it is easy to show that our per-element surprisal obtains equivalent entropy as the original definition.

Cheng and Vlachos (2024) introduce a hyperparameter α which is applied as an exponent to the similarity function. This is used to rescale $\mathcal{S}$ for more favorable performance on benchmarks. We tune α for the highest correlation between S3E and the entropy of the gender labels assigned to the nouns in question by the morphological parser. This way we aim for $\mathcal{H}$ (S3E) to capture gender variation, with higher values of $\mathcal{H}$ (S3E) indicating more diversity in the gender morphemes.

We also experiment with alternative similarity metrics to cosine similarity, such as Euclidean Distance, Chebyshev Distance, Manhattan Distance, Minkowski Distance, etc. but find no notable differences and use the simplest option of Cosine Similarity in the paper.

B Annotation of Names

Table 6 presents the names used for expanding the WINOMT dataset to include common names for masculine and feminine genders in French, Spanish, Ukrainian and Russian.

	Spanish	French	Ukrainian	Russian
F	Carla	Anne	Anna	Anna
M	Gabriel	Victor	Ivan	Ivan

Table 6: The names used for extending the WINOMT dataset for different target languages.

C Translation Guidelines and Translator Recruitment Details

Professional translators were hired and paid £0.2 per word as requested. The translation guidelines are as follows:

Please translate the following sentences into [French/Spanish/Ukrainian/Russian]. Please provide two translations for each sentence, with different gender inflections for the highlighted noun —one masculine and one feminine. Please keep the pronoun true to the original to the extent that it is possible. Please use the translations that appear natural to you as a native speaker rather than using forms that are prescribed[4]. Please note whether both translations are acceptable in terms of the gender of the noun matching the contextual information, and if they are not, which one is the correct one:

- The **developer** argued with the designer because she did not like the design.
- The undergraduate presented the **administrator** with a petition that they had organized.

D Human Annotation Quality

The annotations of Correct and Incorrect sentences by professional translators are compared to the gold standard gender annotations in the dataset, and across all four translators range between 93.17 and 93.27 in Cohen's κ scores, confirming that

[4]This instruction was added due to inquiries by translators about the use of formal or dominant dialects as well as language policy changes influencing how they might translate some words.

Model	OPUS-MT				deb-OPUS-MT				M2M100			
Dataset	es	fr	uk	ru	es	fr	uk	ru	es	fr	uk	ru
newstest2012	84.52	82.21	–	–	84.47	82.22	–	–	71.25	71.97	–	–
newstest2013	85.28	83.45	–	–	85.24	83.44	–	–	72.84	72.57	–	–
newstest2014	–	85.01	–	87.44	–	85.00	–	87.44	–	74.63	–	72.29
wmttest2023	–	–	74.58	79.02	–	–	74.58	79.02	–	–	56.49	55.93
wmttest2024	–	–	66.99	71.64	–	–	67.00	71.63	–	–	49.20	47.63
mean	84.90	83.56	70.79	79.37	84.86	83.55	70.79	79.36	72.05	73.06	52.85	58.62

Table 7: COMET scores on WMT test sets for the models used.

apart from some linguistic idiosyncrasies of each language (e.g. 'victim' in Spanish is always feminine and so regardless of the contextualising pronoun will take the same form), the annotators agree on which sentences should be correctly translated in which gender.

E Overall Model Performance

Table 7 presents the performance of the models used in this study in terms of the COMET metric[5] (Rei et al., 2022) on WMT datasets which contain the target languages (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024). The models are run on a single NVIDIA TU102 GPU.

[5] https://huggingface.co/Unbabel/wmt22-comet-da

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	SF	SM	OF	OM	SF	SM	OF	OM	F	M	SF	SM	OF	OM	S	O	
										s3e											
	Opus-MT	**0.41**	**0.41**	-0.05	**0.25**	**-0.19**	**0.24**	**-0.33**	0.06	0.10	**0.13**	**0.17**	**0.38**	**-0.21**	**0.24**	**-0.31**	**0.29**	**-0.13**	N/A	N/A	**-0.18**
es	deb-Opus-MT	**-0.05**	**0.30**	**-0.11**	**0.27**	**-0.20**	**0.16**	**-0.38**	0.05	**0.14**	**0.08**	0.05	**0.49**	**-0.14**	**0.39**	**-0.24**	**0.14**	**-0.21**	N/A	N/A	**-0.10**
	m2m100	**0.14**	**0.33**	**-0.11**	**0.29**	**-0.42**	**0.28**	**-0.25**	**0.18**	**0.27**	**0.12**	0.07	**0.51**	-0.04	**0.52**	**-0.08**	0.04	**-0.35**	N/A	N/A	**-0.11**
	Opus-MT	**0.54**	**0.42**	**0.16**	0.05	**-0.34**	0.06	**-0.24**	**0.43**	**0.45**	**0.26**	**0.21**	**0.16**	**-0.14**	**0.16**	**-0.17**	**0.20**	-0.04	N/A	N/A	**-0.29**
fr	deb-Opus-MT	**0.19**	**0.22**	0.02	0.05	**-0.25**	**0.17**	**-0.11**	**0.23**	**0.32**	**0.23**	**0.13**	**0.24**	-0.03	**0.31**	**-0.08**	0.01	**-0.14**	N/A	N/A	**-0.12**
	m2m100	**-0.12**	0.11	**-0.23**	**0.50**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.70**	**0.31**	**0.47**	0.05	**-0.08**	**-0.40**	N/A	N/A	0.06
	Opus-MT	**-0.27**	0.00	**-0.11**	**0.26**	-0.02	**0.19**	**0.13**	-0.11	**0.17**	0.03	**-0.14**	**0.27**	**0.21**	**0.22**	**0.13**	**-0.11**	**-0.23**	N/A	N/A	0.06
uk	deb-Opus-MT	**-0.43**	-0.06	**-0.12**	**0.41**	0.00	**0.18**	0.07	**-0.24**	0.01	0.01	**-0.17**	**0.23**	**0.17**	**0.16**	**0.10**	**-0.11**	**-0.17**	N/A	N/A	**0.09**
	m2m100	**0.08**	-0.04	**-0.18**	**0.27**	0.02	**0.33**	**0.18**	**0.19**	**0.37**	-0.03	**-0.17**	**0.53**	**0.34**	**0.50**	**0.26**	**-0.30**	**-0.42**	N/A	N/A	**0.11**
	Opus-MT	0.01	**-0.28**	**-0.41**	0.00	-0.12	0.08	**-0.10**	**-0.19**	**-0.20**	**-0.41**	**-0.39**	**0.07**	-0.03	**0.14**	0.03	**-0.10**	**-0.24**	**-0.40**	0.04	**0.35**
ru	deb-Opus-MT	**0.23**	**-0.10**	**-0.17**	0.05	0.03	0.05	-0.03	**-0.10**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**-0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	0.00	**0.13**
	m2m100	**-0.74**	**-0.16**	**-0.21**	0.10	-0.04	**-0.12**	**-0.25**	**0.12**	**0.12**	**-0.20**	**-0.18**	**0.06**	-0.02	**0.08**	0.01	**-0.07**	**-0.11**	**-0.20**	**-0.22**	**0.18**
										se											
	Opus-MT	**-1.64**	**0.19**	-0.08	0.13	-0.19	0.09	**-0.29**	-0.01	0.02	0.06	0.07	**0.28**	**-0.12**	**0.17**	**-0.22**	**0.15**	**-0.08**	N/A	N/A	-0.05
es	deb-Opus-MT	**-0.06**	**0.27**	**0.13**	-0.01	-0.07	0.00	**-0.19**	0.06	**0.10**	**0.16**	**0.12**	-0.01	**-0.18**	-0.01	**-0.19**	**0.20**	**0.08**	N/A	N/A	**-0.20**
	m2m100	**-1.71**	**0.23**	-0.04	0.20	**-0.29**	**0.17**	**-0.16**	0.10	**0.18**	**0.15**	0.08	**0.28**	-0.06	**0.25**	**-0.09**	**0.07**	**-0.16**	N/A	N/A	-0.10
	Opus-MT	**-1.59**	**0.19**	0.00	0.09	**-0.25**	0.05	**-0.17**	**0.23**	**0.25**	**0.15**	**0.11**	**0.11**	**-0.12**	**0.10**	**-0.15**	**0.14**	-0.04	N/A	N/A	-0.09
fr	deb-Opus-MT	-0.01	**0.24**	**0.08**	0.11	**-0.17**	0.04	**-0.10**	**0.23**	**0.20**	**0.19**	**0.18**	**-0.06**	**-0.22**	0.01	**-0.15**	**0.18**	0.03	N/A	N/A	**-0.16**
	m2m100	**-0.15**	**0.29**	**0.16**	0.03	-0.05	0.02	**-0.14**	**0.27**	**0.24**	**0.27**	**0.30**	0.01	**-0.14**	0.01	**-0.15**	**0.19**	**0.06**	N/A	N/A	**-0.22**
	Opus-MT	**-1.54**	0.00	**-0.12**	**0.21**	-0.05	0.10	0.02	-0.09	0.05	0.01	-0.08	**0.13**	0.04	**0.10**	0.02	0.00	**-0.14**	N/A	N/A	0.06
uk	deb-Opus-MT	**0.09**	**0.12**	**-0.06**	**0.13**	-0.07	**0.08**	0.01	-0.02	**0.08**	**0.15**	**0.09**	**0.06**	**-0.11**	**-0.07**	**-0.10**	**0.13**	**0.06**	N/A	N/A	**0.09**
	m2m100	**-1.72**	-0.04	**-0.14**	0.20	0.03	**0.15**	0.08	0.05	**0.16**	-0.02	**-0.11**	**0.22**	**0.10**	**0.18**	**0.09**	**-0.09**	**-0.19**	N/A	N/A	0.09
	Opus-MT	**-1.50**	**-0.17**	**-0.25**	**0.21**	0.07	0.00	-0.02	**-0.13**	**-0.18**	**-0.25**	**-0.28**	0.03	-0.03	**0.10**	0.00	**-0.06**	**-0.14**	**-0.32**	-0.02	**0.21**
ru	deb-Opus-MT	**-0.06**	0.04	0.01	0.11	0.05	0.00	-0.06	**-0.22**	**-0.13**	**0.07**	-0.01	**-0.11**	**-0.15**	**-0.05**	**-0.11**	**0.09**	**0.07**	**-0.24**	**-0.12**	0.03
	m2m100	**-1.57**	-0.09	**-0.14**	**0.11**	0.10	0.04	**-0.18**	-0.01	0.00	**-0.10**	**-0.10**	0.00	-0.06	0.03	-0.01	-0.01	**-0.07**	**-0.12**	**-0.18**	0.11
										ge											
	Opus-MT	**0.02**	**0.30**	**0.04**	**-0.17**	**0.16**	**-0.10**	**-0.14**	**0.11**	**0.15**	**0.17**	**0.12**	**0.19**	**-0.12**	**0.19**	**-0.20**	**0.15**	**-0.06**	N/A	N/A	**0.17**
es	deb-Opus-MT	**-0.04**	**0.27**	**0.06**	**0.16**	**-0.10**	**0.15**	**-0.12**	**0.09**	**0.17**	**0.17**	**0.09**	**0.21**	**-0.09**	**0.24**	**-0.16**	**0.09**	**-0.06**	N/A	N/A	**-0.16**
	m2m100	**-0.10**	**0.16**	0.00	**0.08**	**-0.10**	**0.09**	**-0.08**	**0.05**	**0.07**	**0.10**	**0.09**	**0.09**	**-0.08**	**0.08**	**-0.16**	**0.12**	-0.01	N/A	N/A	**-0.08**
	Opus-MT	0.01	**0.20**	**0.05**	**0.12**	**-0.06**	**0.07**	**-0.11**	**0.09**	**0.13**	**0.12**	**0.08**	**0.09**	**-0.10**	**0.06**	**-0.19**	**0.15**	**0.03**	N/A	N/A	**-0.13**
fr	deb-Opus-MT	**-0.05**	**0.18**	**0.04**	**-0.11**	**0.08**	**0.08**	**-0.11**	**0.05**	**0.12**	**0.12**	**0.06**	**0.11**	**-0.09**	**0.10**	**-0.17**	**0.11**	0.01	N/A	N/A	**-0.11**
	m2m100	**-0.02**	**0.19**	0.02	**0.12**	**-0.10**	**0.09**	**-0.12**	**0.08**	**0.10**	**0.11**	**0.07**	**0.08**	**-0.11**	**0.04**	**-0.19**	**0.17**	**0.02**	N/A	N/A	**-0.11**
	Opus-MT	**-0.04**	**0.05**	**-0.04**	**0.07**	-0.03	0.03	**-0.07**	**0.06**	**0.06**	0.02	0.02	0.02	**-0.06**	**-0.11**	**-0.15**	**0.16**	**0.06**	N/A	N/A	0.01
uk	deb-Opus-MT	**-0.03**	**0.05**	0.02	**0.07**	-0.03	0.00	**-0.06**	**0.06**	**0.07**	**0.03**	**-0.03**	**-0.02**	**-0.05**	**-0.16**	**-0.17**	**0.17**	**0.14**	N/A	N/A	**0.03**
	m2m100	**-0.04**	**0.06**	**-0.04**	**0.09**	-0.05	**0.05**	**-0.06**	**0.07**	**0.08**	0.01	0.01	**0.03**	**-0.07**	**-0.07**	**-0.18**	**0.16**	**0.06**	N/A	N/A	0.01
	Opus-MT	-0.01	0.00	**-0.03**	-0.02	-0.01	0.00	**-0.04**	0.03	**0.03**	**-0.04**	**-0.06**	-0.01	**-0.04**	**-0.11**	**-0.13**	**0.12**	**0.08**	**-0.04**	0.00	-0.02
ru	deb-Opus-MT	**-0.02**	-0.01	-0.02	**-0.05**	**-0.05**	-0.01	-0.02	0.01	0.02	**-0.04**	**-0.05**	**-0.03**	**-0.04**	**-0.15**	**-0.15**	**0.13**	**0.12**	**-0.05**	-0.01	-0.02
	m2m100	**0.02**	**0.07**	0.02	-0.01	**-0.06**	-0.01	**-0.07**	**0.09**	**0.07**	0.02	**0.04**	-0.01	**-0.06**	**-0.05**	**-0.11**	**0.10**	**0.06**	**-0.04**	**-0.03**	**0.04**

Table 8: ANOVA results: single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on norm-$\mathcal{H}$ (S3E), norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

F ANOVA Results

Table 8 presents the ANOVA results for S3E, SE and GE metrics. Table 9 presents the ANOVA results without normalising the $\mathcal{H}$ values.

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	S F	S M	O F	O M	S F	S M	O F	O M	F	M	S F	S M	O F	O M	S	O	
										S3E											
	OPUS-MT	**12.12**	**47.0**	**83.53**	-2.45	**20.67**	-2.44	**27.76**	**38.98**	**49.12**	**44.16**	**37.6**	0.23	**31.36**	**7.95**	**42.89**	**-22.44**	**14.64**	N/A	N/A	**-65.29**
es	deb-OPUS-MT	**-2.35**	**36.52**	**79.64**	**-58.1**	-7.06	**18.24**	**-8.44**	**31.09**	**33.94**	**42.22**	**36.85**	**32.48**	-1.86	**34.19**	**6.06**	**48.72**	**-29.24**	N/A	N/A	**14.06**
	M2M100	**0.19**	**0.54**	0.02	**-0.28**	**0.38**	**-0.27**	**0.37**	**-0.26**	**0.18**	**0.25**	**0.25**	**0.17**	**0.31**	**-0.25**	**0.31**	**-0.29**	**0.31**	N/A	N/A	**-0.17**
	OPUS-MT	**0.12**	**0.41**	**0.06**	**0.32**	**-0.15**	**0.23**	**-0.17**	**0.18**	**0.25**	**0.23**	**0.14**	**0.26**	**-0.16**	**0.32**	**-0.19**	**0.16**	**-0.14**	N/A	N/A	**-0.24**
fr	deb-OPUS-MT	**-0.02**	**0.26**	**0.06**	**-0.16**	**0.23**	**-0.07**	**0.15**	**-0.1**	**0.12**	**0.21**	**0.18**	**0.08**	**0.21**	**-0.08**	**0.25**	**-0.12**	**0.05**	N/A	N/A	**-0.08**
	M2M100	**-0.12**	0.11	**-0.23**	0.06	**0.5**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.7**	**0.31**	**0.47**	0.05	**-0.08**	N/A	N/A	**-0.4**
	OPUS-MT	**0.12**	**0.22**	0.03	**0.16**	**-0.07**	**0.09**	**-0.11**	**0.15**	**0.18**	**0.14**	**0.1**	**0.09**	**-0.1**	**0.05**	**-0.1**	**0.16**	**-0.04**	**-0.09**	**-0.05**	**-0.12**
uk	deb-OPUS-MT	0.83	**-10.2**	**24.93**	**-7.38**	**-19.32**	**18.37**	**-14.82**	**27.31**	**5.26**	**9.67**	**8.87**	**5.29**	**-15.13**	**23.85**	**-10.15**	**44.14**	**-26.98**	1.7	-2.21	-0.34
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**15.38**	**4.07**	**52.95**	**-27.98**	**17.38**	**-24.19**	**20.45**	**11.79**	**12.51**	**12.13**	**13.6**	**-24.94**	**22.11**	**-28.52**	**41.31**	**-18.56**	**23.26**	-0.79	-1.99	**-28.53**
ru	deb-OPUS-MT	**0.23**	**-0.1**	**-0.17**	**0.13**	0.05	0.03	0.05	-0.03	**-0.1**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	-0.0
	M2M100	**0.08**	**-0.07**	**-0.05**	**0.06**	**-0.02**	-0.01	**-0.02**	-0.0	**-0.06**	**-0.07**	**-0.05**	**-0.03**	**-0.01**	0.0	**-0.02**	0.0	**-0.01**	-0.0	0.01	**0.03**
										SE											
	OPUS-MT	**12.12**	**47.0**	**83.53**	-2.45	**20.67**	-2.44	**27.76**	**38.98**	**49.12**	**44.16**	**37.6**	0.23	**31.36**	**7.95**	**42.89**	**-22.44**	**14.64**	N/A	N/A	**-65.29**
es	deb-OPUS-MT	**-0.08**	**-0.06**	**-0.05**	**0.05**	-0.01	0.0	**-0.01**	-0.0	**-0.04**	**-0.05**	**-0.04**	**-0.03**	**-0.01**	0.0	**-0.01**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	M2M100	**0.08**	**-0.07**	**-0.06**	**0.07**	-0.01	-0.0	**-0.02**	**-0.01**	**-0.05**	**-0.05**	**-0.04**	**-0.04**	**-0.01**	0.0	**-0.02**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	OPUS-MT	**16.01**	**47.22**	**80.76**	-0.14	**19.01**	**-8.69**	**26.55**	**31.41**	**35.76**	**33.4**	**31.98**	-1.93	**29.87**	2.67	**35.97**	**-15.81**	**17.83**	N/A	N/A	**-64.01**
fr	deb-OPUS-MT	**-0.07**	**-0.05**	**-0.04**	**0.04**	-0.0	**0.02**	-0.01	0.0	**-0.03**	**-0.04**	**-0.03**	**-0.03**	**-0.01**	**0.01**	**-0.01**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	M2M100	**0.14**	**0.38**	**0.04**	**-0.21**	**0.35**	**-0.15**	**0.23**	**-0.17**	**0.15**	**0.2**	**0.19**	**0.13**	**0.23**	**-0.16**	**0.24**	**-0.19**	**0.19**	N/A	N/A	**-0.12**
	OPUS-MT	**-0.01**	**-0.01**	**-0.01**	**-0.03**	-0.02	0.0	0.0	**-0.03**	**-0.03**	**-0.01**	-0.01	-0.0	0.0	**-0.01**	0.01	-0.0	-0.0	**0.02**	**0.03**	**0.01**
uk	deb-OPUS-MT	0.0	**0.09**	0.02	**-0.05**	0.01	-0.06	**0.06**	-0.04	**0.12**	**0.16**	**0.08**	**0.04**	**0.04**	**-0.02**	**0.08**	0.02	0.01	**-0.06**	**-0.1**	**-0.03**
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**15.38**	**4.07**	**52.95**	**-27.98**	**17.38**	**-24.19**	**20.45**	**11.79**	**12.51**	**12.13**	**13.6**	**-24.94**	**22.11**	**-28.52**	**41.31**	**-18.56**	**23.26**	-0.79	-1.99	**-28.53**
ru	deb-OPUS-MT	1.34	-3.62	**46.99**	**-21.71**	**-33.31**	**16.73**	**-27.33**	**19.72**	**9.4**	**7.79**	**6.6**	**9.66**	**-28.51**	**22.12**	**-24.71**	**50.74**	**-26.37**	**15.86**	-0.28	-0.68
	M2M100	**20.01**	**18.76**	**50.38**	**-34.59**	**-11.85**	5.91	**-16.63**	**13.63**	**14.94**	**15.4**	**14.21**	**15.2**	**-17.46**	**14.68**	**-15.4**	**30.42**	**-10.35**	**16.48**	-0.74	-3.23
										GE											
	OPUS-MT	**-0.0**	**-0.02**	**-0.01**	-0.01	0.01	**-0.01**	0.0	**-0.02**	**-0.02**	**-0.01**	**-0.01**	**-0.01**	**0.01**	-0.0	**0.01**	**-0.01**	-0.0	N/A	N/A	**0.02**
es	deb-OPUS-MT	**-2.35**	**36.52**	**79.64**	**-58.1**	-7.06	**18.24**	**-8.44**	**31.09**	**33.94**	**42.22**	**36.85**	**32.48**	-1.86	**34.19**	**6.06**	**48.72**	**-29.24**	N/A	N/A	**14.06**
	M2M100	**0.19**	**0.54**	0.02	**-0.28**	**0.38**	**-0.27**	**0.37**	**-0.26**	**0.18**	**0.25**	**0.25**	**0.17**	**0.31**	**-0.25**	**0.31**	**-0.29**	**0.31**	N/A	N/A	**-0.17**
	OPUS-MT	**0.06**	**-0.05**	**-0.04**	0.01	**0.02**	**-0.01**	0.01	**-0.04**	**-0.04**	**-0.03**	**-0.02**	**-0.0**	**0.01**	**-0.01**	**0.01**	**-0.02**	**-0.01**	N/A	N/A	**0.04**
fr	deb-OPUS-MT	**-0.05**	**0.18**	**0.04**	**-0.11**	**0.08**	**-0.11**	**0.08**	**-0.11**	**0.05**	**0.12**	**0.12**	**0.06**	**0.11**	**-0.09**	**0.1**	**-0.17**	**0.11**	N/A	N/A	0.01
	M2M100	**0.14**	**0.38**	**0.04**	**-0.21**	**0.35**	**-0.15**	**0.23**	**-0.17**	**0.15**	**0.2**	**0.19**	**0.13**	**0.23**	**-0.16**	**0.24**	**-0.19**	**0.19**	N/A	N/A	**-0.12**
	OPUS-MT	**0.12**	**0.22**	0.03	**0.16**	**-0.07**	**0.09**	**-0.11**	**0.15**	**0.18**	**0.14**	**0.1**	**0.09**	**-0.1**	**0.05**	**-0.1**	**0.16**	**-0.04**	**-0.09**	**-0.05**	**-0.12**
uk	deb-OPUS-MT	0.0	**0.09**	0.02	**-0.05**	0.01	-0.06	**0.06**	-0.04	**0.12**	**0.16**	**0.08**	**0.04**	**0.04**	**-0.02**	**0.08**	0.02	0.01	**-0.06**	**-0.1**	**-0.03**
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**0.09**	**0.17**	**0.05**	0.01	**-0.11**	**0.05**	**-0.08**	**0.08**	**0.13**	**0.11**	**0.06**	**0.04**	**-0.08**	-0.02	**-0.06**	**0.14**	-0.01	**-0.08**	**-0.03**	**-0.11**
ru	deb-OPUS-MT	**-0.02**	-0.01	-0.02	0.02	**-0.05**	**-0.05**	-0.01	-0.02	0.01	0.02	**-0.04**	**-0.05**	**-0.03**	**-0.04**	**-0.15**	**-0.15**	**0.13**	**0.12**	**-0.05**	-0.01
	M2M100	**0.15**	**0.29**	**0.08**	**-0.18**	**0.13**	**-0.1**	**0.09**	**-0.12**	**0.16**	**0.21**	**0.15**	**0.11**	**0.07**	**-0.13**	**0.08**	**-0.05**	**0.15**	**-0.08**	**-0.1**	-0.02

Table 9: ANOVA results (no normalisation): single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on $\mathcal{H}$ (S3E), $\mathcal{H}$ (SE) and $\mathcal{H}$ (GE). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

G Gender Accuracy

Table 10 presents more fine-grained results than Table 2 with regard to gender accuracy, namely splitting the results by subset of the dataset. The results in the Ambiguous column are not meaningfully interpretable, as a single ground truth label of gender cannot capture the true desired behavior of the model, especially when the gold label for ambiguous cases is mostly 'neutral', and neutral is not commonly used as grammatical gender for animate objects in the languages used in this study. The case of Russian, where the performance increases on the Ambiguous subset actually reflects the model choosing the masculine forms, which are tagged as 'neutral' by the morphological parser due to the masculine form often being the default choice for both genders, as discussed in Section 4.

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	55.20	67.95	52.10	67.95	33.96
	deb-OPUS-MT	55.69	68.13	52.95	68.13	34.39
	M2M100	55.68	70.77	51.17	70.77	32.40
fr	OPUS-MT	52.05	64.27	46.55	64.27	37.25
	deb-OPUS-MT	52.98	64.79	48.10	64.79	37.75
	M2M100	50.95	61.66	47.57	61.66	34.84
uk	OPUS-MT	38.65	45.34	34.20	45.34	33.75
	deb-OPUS-MT	38.95	46.12	34.15	46.12	33.74
	M2M100	40.97	47.76	36.81	47.76	35.20
ru	OPUS-MT	39.50	48.57	33.27	48.57	33.24
	deb-OPUS-MT	39.50	48.42	33.38	48.42	33.33
	M2M100	41.01	48.49	36.81	48.49	33.81

Table 10: Comparison of Gender Accuracy Overall, in Pro-/Anti-Stereotypical and Ambiguous Cases Across Models, on WINOMT.

H Quality Estimation with Human Translations

Table 11 presents the results of the models used in this study on the 100 human-annotated instances. For unambiguous cases we use a single reference, whereas for ambiguous cases, we calculate performance by taking the maximum COMET score of both acceptable translations.

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	81.35	85.80	83.37	84.55	75.14
	deb-OPUS-MT	81.31	85.62	83.43	84.49	75.14
	M2M100	79.56	84.44	81.29	82.82	73.25
fr	OPUS-MT	77.63	82.23	81.16	81.66	70.47
	deb-OPUS-MT	77.69	82.41	80.88	81.60	70.73
	M2M100	76.24	80.69	78.65	79.61	70.25
uk	OPUS-MT	80.56	85.57	82.73	84.10	73.69
	deb-OPUS-MT	80.19	84.85	82.13	83.45	73.86
	M2M100	81.27	85.89	84.09	84.96	74.10
ru	OPUS-MT	82.53	86.20	84.99	85.58	76.86
	deb-OPUS-MT	82.76	86.46	85.27	85.85	77.02
	M2M100	81.71	86.39	84.06	85.21	75.21

Table 11: Comparison of COMET Scores Overall, in Pro-/Anti-Stereotypical and Ambiguous Cases Across Models, on the 100 manually translated sentences.

Language	Model	LogProb (Correct)	LogProb (Incorrect)	S3E I (Correct)	S3E I (Incorrect)	SE I (Correct)	SE I (Incorrect)	GE I (Correct)	GE I (Incorrect)
ES	OPUS-MT	-149.7	-149.78	7.83	8.88	0.3	0.35	0.33	0.3
	deb-OPUS-MT	-149.19	-149.01	8.08	9.16	0.29	0.31	0.35	0.33
	M2M100	-226.61	-227.29	23.61	26.03	0.41	0.4	0.42	0.43
FR	OPUS-MT	-197.1	-195.11	9.18	9.89	0.73	0.72	0.24	0.29
	deb-OPUS-MT	-196.98	-195.09	9.18	9.85	0.48	0.52	0.26	0.33
	M2M100	-283.91	-281.71	186.42	194.52	0.49	0.4	0.43	0.48
UK	OPUS-MT	-161.98	-161.14	147.6	152.15	0.6	0.54	0.22	0.22
	OPUS-MT-debiased	-161.46	-160.68	150.52	153.76	0.49	0.47	0.23	0.23
	M2M100	-241.0	-241.49	204.72	211.9	0.28	0.24	0.23	0.25
RU	OPUS-MT	-170.72	-170.9	32.14	33.15	0.38	0.43	0.08	0.19
	deb-OPUS-MT	-170.58	-170.75	32.31	33.27	0.32	0.37	0.06	0.17
	M2M100	-220.25	-220.78	218.11	219.06	0.45	0.4	0.16	0.3

Table 12: Log Probability and Surprisal Measures across Models and Languages

COMET Unambiguous	**COMET All**	**Gender Acc**	**Delta S**	**Delta H**
deb-OPUS-MT-RU	OPUS-MT-ES	M2M100-ES	M2M100-ES	M2M100-RU
OPUS-MT-RU	deb-OPUS-MT-ES	deb-OPUS-MT-ES	deb-OPUS-MT-ES	M2M100-UK
M2M100-RU	OPUS-MT-FR	OPUS-MT-ES	OPUS-MT-ES	OPUS-MT-UK
M2M100-UK	deb-OPUS-MT-FR	deb-OPUS-MT-FR	deb-OPUS-MT-FR	deb-OPUS-MT-UK
OPUS-MT-ES	OPUS-MT-RU	OPUS-MT-FR	M2M100-FR	OPUS-MT-RU
deb-OPUS-MT-ES	deb-OPUS-MT-RU	M2M100-FR	OPUS-MT-FR	deb-OPUS-MT-RU
OPUS-MT-UK	M2M100-FR	OPUS-MT-RU	deb-OPUS-MT-RU	deb-OPUS-MT-ES
deb-OPUS-MT-UK	M2M100-ES	M2M100-RU	M2M100-RU	OPUS-MT-ES
OPUS-MT-ES	OPUS-MT-UK	deb-OPUS-MT-RU	deb-OPUS-MT-UK	deb-OPUS-MT-FR
OPUS-MT-FR	deb-OPUS-MT-UK	M2M100-UK	OPUS-MT-UK	M2M100-FR
deb-OPUS-MT-FR	M2M100-RU	deb-OPUS-MT-UK	M2M100-UK	M2M100-ES
M2M100-FR	M2M100-UK	OPUS-MT-UK	OPUS-MT-RU	OPUS-MT-FR

Table 13: Model rankings across five evaluation metrics.

I Log Probability and Surprisal Scores

Table 12 presents the Log probability and surprisal scores, as well as their relative differences between the Correct and Incorrect translations of the Unambiguous instances in WINOMT.

J Rankings by Different Metrics

Table 13 presents the rankings of models according to various metrics employed in this study.

K Rank Correlation

Table 14 presents the correlation scores between ΔI and Δ Log Probabilies on the one hand, and gender accuracy scores on the other.

Correlation	**Metric**	**Statistic**	**p-value**
Spearman	ΔI (S3E)	**-0.78**	**0.00**
	ΔI (SE)	-0.37	0.24
	ΔI (GE)	0.27	**0.00**
	ΔLog prob	0.11	0.73
Kendall	ΔI (S3E)	**-0.58**	**0.01**
	ΔI (SE)	-0.27	0.25
	ΔI (GE)	0.23	**0.00**
	ΔLog prob	0.09	0.74

Table 14: Spearman and Kendall correlations between ΔI under different uncertainty metrics and Log Probabilities on the one hand, and gender accuracy on the other. Statistically significant correlations ($p < 0.05$) are in **bold**.

Language	Model	S3E			SE			GE		
		Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09	0.33	0.22	0.33	0.21	0.13	0.38
	deb-OPUS-MT	0.97	0.89	0.08	0.41	0.25	0.39	0.24	0.16	0.33
	M2M100	1.79	1.45	0.19	0.46	0.17	0.63	0.25	0.09	0.64
FR	OPUS-MT	1.79	1.43	0.20	0.57	0.40	0.30	0.23	0.08	0.65
	deb-OPUS-MT	1.21	1.08	0.11	0.64	0.50	0.22	0.23	0.09	0.61
	M2M100	3.22	2.78	0.14	0.56	0.29	0.48	0.25	0.17	0.32
UK	OPUS-MT	1.96	2.16	-0.10	0.40	0.39	0.03	0.20	0.14	0.30
	deb-OPUS-MT	1.98	2.15	-0.09	0.44	0.41	0.07	0.22	0.15	0.32
	M2M100	2.05	2.28	-0.11	0.37	0.41	-0.11	0.18	0.16	0.11
RU	OPUS-MT	1.56	1.68	-0.08	0.38	0.32	0.16	0.12	0.03	0.75
	deb-OPUS-MT	1.05	0.97	0.08	0.43	0.42	0.02	0.12	0.06	0.50
	M2M100	1.83	2.29	-0.25	0.50	0.29	0.42	0.15	0.10	0.33

Table 15: $\mathcal{H}$ scores across models and languages, with relative differences ($\Delta\mathcal{H}$) between unambiguous and ambiguous conditions.

L Entropy Scores

Table 15 presents the $\mathcal{H}$ scores and their relative differences between the unambiguous and ambiguous settings for different UQ metrics used in this study.