"You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases

Dirk Hovy **Federico Bianchi** **Tommaso Fornaciari**
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
{dirk.hovy, f.bianchi, fornaciari.tommaso}@unibocconi.it

Abstract

The main goal of machine translation has been to convey the correct content. Stylistic considerations have been at best secondary. We show that as a consequence, the output of three commercial machine translation systems (Bing, DeepL, Google) make demographically diverse samples from five languages "sound" older and more male than the original. Our findings suggest that translation models reflect demographic bias in the training data. These results open up interesting new research avenues in machine translation to take stylistic considerations into account.

1 Introduction

Translating *what* is being said is arguably the most important aspect of machine translation, and has been the main focus of all its efforts so far. However, *how* something is said also has an impact on how the final translation is perceived. Mirkin et al. (2015) have pointed out that demographic aspects of language do play a role in translation, and could help in personalization. As Vanmassenhove et al. (2018) have shown, gendered inflections like "Sono stanco/a" (*Italian* I am tired) are an important aspect of correct translations.

In many cases, capturing the style of a document is equally important as its content: translating a lover's greeting as "I am entirely pleased to see you" might be semantically correct, but seems out of place. Demographic factors (age, gender, etc.) all manifest in language, and therefore influence style: we do not expect a 6-year old to sound like an adult, and would not translate a person to seem differently gendered. However, in this paper, we show such a change is essentially what happens in machine translation: authors sound on average older and more male.

Prior work (Rabinovich et al., 2017) has shown that translation weakens the signal for gender prediction. We substantially extend this analysis in terms of languages, demographic factors, and types of models, controlling for demographically representative samples. We show the direction in which the predicted demographic factors differ in the translations, and find that there are consistent biases towards older and more male profiles. Our findings suggest a severe case of *overexposure* to writings from these demographics (Hovy and Spruit, 2016), which creates a self-reinforcing loop.

In this paper, we use demographically-representative author samples from five languages (Dutch, English, French, German, Italian), and translate them with three commercially available machine translation systems (Google, Bing, and DeepL). We compare the true demographics with the predicted demographics of each translation (as well as a control predictor trained on the same language). Without making any judgment on the translation of the content, we find a) that there are substantial discrepancies in the perceived demographics, and b) that translations tend to make the writers appear older and considerably more male than they are.

Contributions We empirically show how translations affect the demographic profile of a text. We release our data set at `https://github.com/MilaNLProc/translation_bias`. Our findings contribute to a growing literature on biases in NLP (see Shah et al. (2020) for a recent overview).

2 Data

We use the Trustpilot data set from Hovy et al. (2015), which provides reviews in different languages, and includes information about age and gender. We use only English, German, Italian, French, and Dutch reviews, based on two criteria: 1) availability of the language in translation mod-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1686–1690

els, and 2) sufficient data for representative samples (see below) in the corpus. For the English data, we use US reviews, rather than UK reviews, based on a general prevalence of this variety in translation engines.

2.1 Translation Data

For each language, we restrict ourselves to reviews written in the respective language (according to `langid` [1] (Lui and Baldwin, 2012)) that have both age and gender information. We use the CIA factbook[2] data on age pyramids to sample 200 each male and female. We use the age groups given on the factbook, i.e., 15–24, 25–54, 55–64, and 65+. Based on data sparsity in the Trustpilot data, we do not include the under-15 age group. This sampling procedure results in five test sets of about 400 instances each (the exact numbers vary slightly according to rounding and the proportions in the CIA factbook data), balanced for binary gender. The exception is Italian, where the original data is so heavily skewed towards male reviews that even with downsampling, we only achieve a 48:52 gender ratio.

We then translate all non-English test sets into English, and the English test set into all other languages, using three commercially available machine translation tools: Bing, DeepL, and Google Translate.

2.2 Profile Prediction Data

We use all instances that are not part of any test set to create training data for the respective age and gender classifiers (see next section). Since we want to compare across languages fairly, the training data sets need to be of comparable size. We are therefore bounded by the size of the smallest available subset (Italian). We sample about 2500 instances per gender, according to the respective age distributions. This sampling results in about 5000 instances per language (again, the exact number varies slightly based on the availability of samples for each group and rounding). We again subsample to approximate the actual age and gender distribution, since, according to Hovy et al. (2015), the data skews strongly male, while otherwise closely matching the official age distributions.

[1] https://github.com/saffsd/langid.py

[2] https://www.cia.gov/library/publications/the-world-factbook/

3 Methods

To assess the demographic profile of a text, we train separate age and gender classifiers for each language. These classifiers allow us to compare the predicted profiles in the original language with the predicted profiles of the translation, and compare both to the actual demographics of the test data.

We use simple Logistic Regression models with L_2 regularization over 2-6 character-grams, and regularization optimized via 3-fold cross-validation.[3] The numbers in Table 1 indicate that both age and gender can be inferred reasonably well across all of the languages. We use these classifiers in the following analyses.

	de	en	fr	it	nl
gender	0.65	0.62	0.64	0.62	0.66
age	0.52	0.53	0.45	0.52	0.49

Table 1: Macro-F1 for age and gender classifiers on each language.

For each non-English sample, we predict the age and gender of the author in both the original language and in each of the three English translations (Google, Bing, and DeepL). I.e., we use the respective language's classifier described above (e.g., a classifier trained on German to predict German test data), and the English classifier described above for the translations. E.g., we use the age and gender classifier trained on English data to predict the translations of the German test set.

For the English data, we first translate the texts into each of the other languages, using each of the three translation systems. Then we again predict the author demographics in the original English test set (using the classifier trained on English), as well as in each of the translated versions (using the classifier trained on the respective language). E.g., we create a German, French, Italian, and Dutch translation with each Google, Bing, and DeepL, and classify both the original English and the translation.

We can then compare the distribution of age groups and genders in the predictions with the actual distributions. If there is *classifier bias*, both

[3] We also experimented with a convolutional neural network with attention, as well as with BERT-based input representations, but did not see significantly better results, presumably due to the higher number of parameters in each case.

the predictions based on the original language and the predictions based on the translations should be skewed in the same direction. We can measure this difference by computing the Kullback-Leibler (KL) divergence of the predicted distribution from the true sample distribution. In order to see whether the predictions differ statistically significantly from the original, we use a use a χ^2 contingency test and report significance at $p <= 0.05$ and $p <= 0.01$.

If instead there is a *translation bias*, then the translated predictions should exhibit a stronger skew than the predictions based on the original language. By using both translations from and into English, we can further tease apart the direction of this effect.

4 Results

4.1 Gender

Translating *into* English Table 2 shows the results when translating *into* English. It shows for each language the test gender ratio, the predicted ratio from classifiers trained in the same language, as well as their KL divergence from the ratio in the test set, and the ratio predictions and KL divergence on predictions of an English classifier on the translations from three MT systems.

For most languages, there exists a male bias in predictions of the original language. The translated English versions create an even stronger skew. The notable exception is French, which most translation engines render in a demographically faithful manner. Dutch is slightly worse, followed by Italian (note, though, that the Italian data was so heavily imbalanced that we could not sample an even distribution for the test data). Somewhat surprisingly, the gender skew is strongest for German, swinging by as much as 15 percentage points.

Translating *from* English Table 3 shows the results when translating *from* English into the various languages. The format is the same as for Table 2.

Again we see large swings, normally exacerbating the balance towards men. However, translating into German with all systems produces estimates that are a lot more female than the original data. This result could be the inverse effect of what we observed above. Again, there is little change for French, though we also see some female bias in two MT systems.

4.2 Age

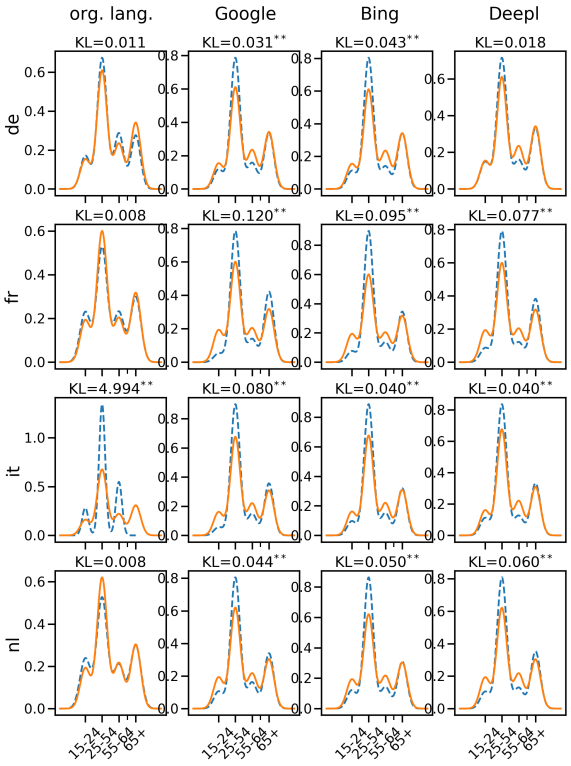

Figure 1: Density distribution and KL for age prediction in various languages and different systems in original and when translated **into** English. Solid yellow line = true distribution. * = predicted distribution differs significantly from gold distribution at $p <= 0.05$. ** = significant difference at $p <= 0.01$.

Figure 1 shows the kernel density plots for the four age groups in each language (rows) in the same language prediction, and in the English translation. In all cases, the distributions are reasonably close, but in all cases, the predictions overestimate the most prevalent class.

To delve a bit deeper into this age mismatch, we also split up the sample by decade (i.e., seven classes: 10s, 20s, etc., up to 70s+). Figure 2 shows the results. The caveat here is that the overall performance is lower, due to the higher number of classes. We also can not guarantee that the distribution still follows the true demographics, since we are subsampling within the larger classes given by the CIA factbook.

However, the results still strongly suggest that the observed mismatch is driven predominantly by overprediction of the 50s decade. Because this decade often contributed strongly to the most frequent age category (25–54), predictions did not differ as much from gold in the previous test. It

from	gold F:M split	org. lang F:M split	org. lang KL	Google F:M split	Google KL	Bing F:M split	Bing KL	DeepL F:M split	DeepL KL
de	50 : 50	48 : 52	0.001	37 : 63**	0.034	35 : 65**	0.045	35 : 65**	0.045
fr	50 : 50	47 : 53	0.002	49 : 51	0.000	48 : 52	0.001	49 : 51	0.000
it	48 : 52	47 : 53	0.000	37 : 63**	0.026	43 : 57	0.006	36 : 64**	0.033
nl	50 : 50	49 : 51	0.000	47 : 53	0.001	47 : 53	0.002	44 : 56	0.007
avg			0.000		0.015		0.013		0.021

Table 2: Gender split (%) and KL divergence from gold for each language when translated **into** English. ** = split differs significantly from gold split at $p <= 0.01$.

gold F:M split	English F:M split	English KL	to	Google F:M split	Google KL	Bing F:M split	Bing KL	DeepL F:M split	DeepL KL
50 : 50	49 : 51	0.000	de	59 : 41*	0.015	58 : 42*	0.013	58 : 42*	0.011
			fr	49 : 51	0.000	52 : 48	0.001	54 : 46	0.003
			it	45 : 55	0.004	44 : 56	0.007	41 : 59*	0.016
			nl	40 : 60**	0.020	43 : 57*	0.010	40 : 60**	0.019
avg					0.010		0.008		0.012

Table 3: Gender split (%) and KL divergence from gold for each language when translated **from** English. * = split differs significantly from gold split at $p <= 0.05$. ** = significant difference at $p <= 0.01$.

also explains the situation of the Italian predictor.

In essence, English translations of all these languages, irrespective of the MT system, sound much older than they are.

4.3 Discrepancies between MT Systems

All three tested commercial MT systems are close together in terms of performance. However, they also seem to show the same systematic translation biases. The most likely reason is the use of biased training data. The fact that translations into English are perceived as older and more male than translations into other languages could indicate that there is a larger collection of unevenly selected data in English than for other languages.

5 Related Work

The work by Rabinovich et al. (2017) is most similar to ours, in that they investigated the effect of translation on gender. However, it differs in a few key points: they show that translation weakens the predictive power, but do not investigate the direction of false predictions. We show that there is a definitive bias. In addition, we extend the analysis to include age. We also use various commercially available MT tools, rather than research systems.

Recent research has suggested that machine translation systems reflect cultural and societal biases (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019), though mostly focusing on data selection and embeddings as sources.

Work by Mirkin et al. (2015); Mirkin and Meunier (2015) has set the stage for considering the impact of demographic variation (Hovy et al., 2015) and its integration in MT more general.

There is a growing literature on various types of bias in NLP. For a recent overview, see Shah et al. (2020).

6 Conclusion

We test what demographic profiles author attribute tools predict for the translations from various commercially available machine translation tools. We find that independent of the MT system and the translation quality, the predicted demographics differ systematically when translating into English. On average, translations make the author seem substantially older and more male. Translating from English into any of the other languages shows more mixed results, but similar tendencies.

Acknowledgments

The authors would like to thank Pietro Lesci, Serena Pugliese, and Debora Nozza, as well as the anonymous reviewers, for their kind suggestions. The authors are members of the Bocconi Institute

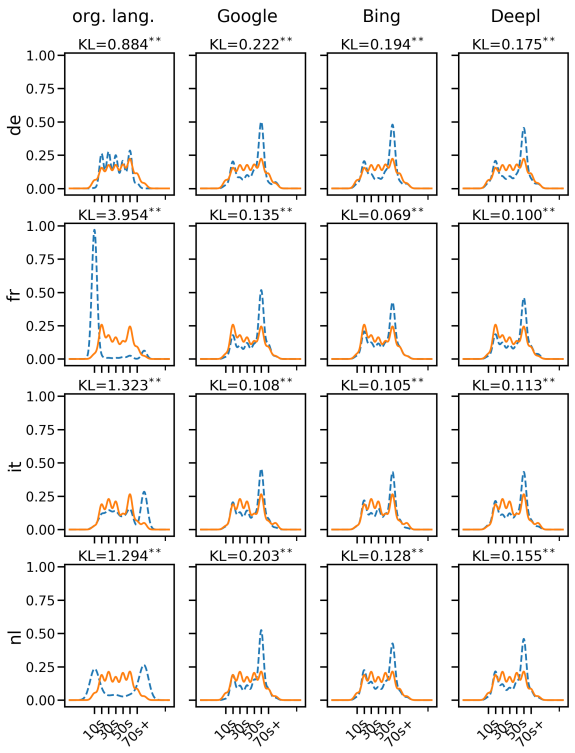

Figure 2: Density distribution and KL for decade prediction in various languages and different systems in original and when translated **into** English. Solid yellow line = true distribution. * = predicted distribution differs significantly from gold distribution at $p <= 0.05$. ** = significant difference at $p <= 0.01$.

for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit.

References

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461. International World Wide Web Conferences Steering Committee.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA, USA. Association for Computational Linguistics.

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

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.