FBK@IWSLT Test Suites Task: Gender Bias evaluation with MuST-SHE

Beatrice Savoldi, Marco Gaido, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler
{bsavoldi,mgaido,negri,bentivo}@fbk.eu

Abstract

This paper presents the FBK contribution to the IWSLT-2024 "Test suites" shared subtask, part of the Offline Speech Translation Task. Our contribution consists of the MuST-SHE^{IWSLT24} benchmark evaluation, designed to assess gender bias in speech translation. By focusing on the en-de language pair, we rely on a newly created test suite to investigate systems' ability to correctly translate feminine and masculine gender. Our results indicate that – under realistic conditions – current ST systems achieve reasonable and comparable performance in correctly translating both feminine and masculine forms when contextual gender information is available. For ambiguous references to the speaker, however, we attest a consistent preference towards masculine gender, thus calling for future endeavours on the topic. Towards this goal we make MuST-SHE^{IWSLT24} freely available at: https://mt.fbk.eu/must-she/

1 Introduction

In today's interconnected world, speech translation technology stands as a cornerstone of global communication, facilitating seamless interactions across linguistic barriers. Indeed, the last few years have seen notable advancements for the task of speech-to-text translation (ST), which has made strides in generic performance (Bentivogli et al., 2021; Anastasopoulos et al., 2021, 2022; Agarwal et al., 2023). Also, the emergence massively multilingual solutions has greatly expanded the language coverage of competitive "one-model-fits-all" speech models (Radford et al., 2022; Communication et al., 2023; Peng et al., 2024; Pratap et al., 2024).

Amid such advancements, there arise the increasing need to pair traditional overall quality assessments of ST with more fine-grained analyses by accounting for relevant aspects of translations. It is within this context that the IWSLT Test Suites shared task emerges, aiming to provide a dedicated evaluation framework for specific dimensions of the ST output, which are otherwise overlooked with generic test sets and holistic metrics.

In light of the above, our contribution is dedicated to the critical themes of gender bias in automatic translation (Costa-jussà, 2019; Savoldi, 2023; Vanmassenhove, 2024).[1] Given the large-scale deployment of ST, biased translations are not only relevant from a technical perspective, where gender-related errors negatively impact the accuracy of automatic translation. Rather, biased and non-inclusive systems can pose the concrete risk of under/misrepresenting gender minorities by overproducing masculine forms and reinforcing gendered stereotypes (Blodgett et al., 2020; Sun et al., 2019). Indeed, gendered linguistic expressions affect the representation and perception of individuals (Stahlberg et al., 2007; Corbett, 2013; Gygax et al., 2019), and are actively used as a tool to negotiate the social, personal, and political reality of gender (Hellinger and Motschenbacher, 2015). A such, models that systematically favor masculine over feminine forms fail to properly recognize women, can reduce feminine visibility, and offer an unequal service quality (Crawford, 2017).

This paper presents the FBK participation in the Test Suites shared task by conducting evaluations on the **MuST-SHE**^{IWSLT24} en-de dataset. It represents the newly created *speech-to-text* extension of the English→German *textual-only* portion of MuST-SHE (Savoldi et al., 2023), a multilingual gender bias benchmark (Bentivogli et al., 2020).

In the hereby presented evaluations, we obtained translations of our test suites by systems that are part of the Offline Speech Translation Task of the 21st International Conference on Spoken Language

[1] Its relevance is also attested by the creation of dedicated workshops on theme of gender bias and inclusivity, such as GeBNLP (Hardmeier et al., 2022) and GITT (Vanmassenhove et al., 2023).

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 113–119

Form		Category 1: *Ambiguous first-person references*	Speaker
Fem.	src	The other hat that I've worn in my work is as **an activist**...	She
	Ref_{De}	Der andere Hut, den ich bei meiner Arbeit getragen habe, ist **der**<den> **Aktivistin**<Aktivist>...	
Masc.	src	I mean, I'm a **journalist**.	He
	Ref_{De}	Ich meine, ich bin **Journalist** <Journalistin>.	
		Category 2: *Unambiguous references with gender cue in context*	
Fem.	src	A college classmate wrote me a couple weeks ago and she said ...	He
	Ref_{De}	**Eine**<Ein> **Kommilitonin**<Kommiliton> hat mir vor ein paar Wochen geschrieben und gesagt...	
Masc.	src	I decided to pay a visit to **the manager** [...] and he pointed ...	She
	Ref_{De}	Also entschied ich mich **den**<die>**Filialleiter**<Filialleiterin> zu besuchen [...]	

Table 1: Textual portion of MuST-SHE (Savoldi et al., 2023), with annotated segments organized per category. For each gender-neutral word referring to a human entity in the English source sentence (SRC), the reference translation (REF) shows the corresponding gender-marked (Fem/Masc) forms, annotated with their wrong <gender-swapped> forms. The last column provides information about the speaker's gender.

Translation (IWSLT 2024). Specifically, we evaluated 13 systems for MuST-SHE^{IWSLT24} en-de.

2 MuST-SHE^{IWSLT24}

MuST-SHE^{IWSLT24} is a test suite designed to evaluate the ability of ST systems to correctly translate gender. It is composed of 200 segments that require the translation of – at least – one English gender-neutral word into the corresponding masculine or feminine target word(s) in German.[2] The test suite is created as an extension of MuST-SHE, a multilingual, natural benchmark built on TED talks data (Bentivogli et al., 2020). The original corpus comprises ∼3,000 (*audio, transcript, translation*) triplets annotated with qualitatively differentiated gender-related phenomena for thee language pairs: English→ French/Italian/Spanish. Recently, MuST-SHE was also extended to English→ German for the MT task – i.e. MuST-SHE^{WMT23} (Savoldi et al., 2023). However, since it only consists of a textual portion (*transcript, translation*), it does not allow for the evaluation of ST models.

Here, we introduce the expansion of **MuST-SHE English→ German for the ST task**, by incorporating the additional speech input portion so as to obtain (*audio, transcript, translation*) triplets.

2.1 Audio Portion Creation

To ensure conformity, the dataset audio portion was obtained by following the same automatic procedures used for MuST-SHE and other TED-based resources, as reported in (Cattoni et al., 2021). Accordingly, from the official TED website we downloaded the videos of the talks included in the textual portion of MuST-SHE English→ German. On this basis, *i)* audio tracks were extracted from the videos, and *ii*) an alignment procedure was applied to split talks into segments and generate aligned (*audio, transcript, translation*) triplets. Since this automatic procedure generates 90% of properly aligned triples on average (Cattoni et al., 2021), we performed qualitative checks. Two evaluators – both students proficient in the German language and with a background in Applied Linguistics[3] – reviewed all the extracted audios and corrected any audio-text misalignment.[4] Hence, we ensured the quality of all audio segments included in MuST-SHE^{IWSLT24}, and the exact alignment of each (*audio, transcript, translation*) triplet.

2.2 Dataset Features

MuST-SHE is designed to evaluate the translation of a source English neutral word into its corresponding target gender-marked one(s) in the context of human referents, e.g. en: ***the*** *good* ***friend***, de: ***der/die*** *gute* ***Freund/in***. To allow for fine-grained analyses, each segment in MuST-SHE is enriched with the following annotations:

· GENDER, which allows to distinguish results for Feminine (Fem) and Masculine (Masc) forms, thus revealing a potential gender gap.

· CATEGORY, which differentiates between **CAT1**

[2] See §5 for a discussion on the use of (binary) gender as a variable.

[3] Their work was carried out during an internship at FBK.

[4] We relied on the ELAN annotation tool: https://archive.mpi.nl/tla/elan.

– first-person references to be translated according to the speakers' linguistic expression of gender[5] (e.g. *I am a teacher*) – and **CAT2** – references to any participant, to be translated in agreement with gender information available in the sentence (e.g. *He/she* is a *teacher*). These categories allow analysing models' behaviour across unambiguous and ambiguous gender translation instances.[6]

· GENDER-SWAPPED WORDS, providing, for each target gender-marked word annotated in MuST-SHE reference translations, a corresponding wrong form swapped in the opposite gender (e.g. en: *she is* ***a friend***; de: *Sie ist* ***eine<ein> Freundin<Freund>***). As described in §3.2, such pairs of annotated target gender-marked words are a key feature of MuST-SHE, which enables gender-focused evaluations.

All above-mentioned dimensions are already provided with the textual portion of MuST-SHE English→ German, and are consequently also included in MuST-SHE^{IWSLT24}. In Table 1, we show examples of annotated (*transcript, translation*) segments from the corpus. Overall dataset statistics are provided in Table 2.

	CAT1	CAT2
Fem.	23 *(35)*	77 *(121)*
Masc.	23 *(38)*	77 *(155)*
Tot.	200 *(349)*	

Table 2: MuST-SHE^{IWSLT24} statistics: number of sentences and (*gender-marked target words*).

3 Experimental Settings

3.1 Models

The test suite evaluation is carried out on the systems that were submitted to the IWSLT Offline Speech Translation tasks. Overall, four different participants – i.e. HW-TSC, CMU, NYA, and KIT – submitted a total of 13 models. Of those, six models were presented as primary system submission, while the other 7 models are additional, contrastive models. All systems contributions are built upon *cascade* architectures, which resolve the ST task as pipelined ASR+MT solutions.

Since the participants (with the only exception of NYA) segmented the sentences before generating the outputs, we isolated the predicted translation for each reference sentence by means of the mWERSegmenter tool (Matusov et al., 2005). This procedure mirrors what is done in the standard evaluation of the offline task (Agarwal et al., 2023).

3.2 Evaluation

Following the original MuST-SHE evaluation protocol described in Gaido et al. (2020), MuST-SHE^{IWSLT24} evaluation allows to focus on the gender realization of the target gender-marked forms, which are annotated in the reference translations together with their *wrong*, gender-swapped form (see Table 1). The evaluation is carried out in two steps, and by matching the annotated (*correct*/*wrong*) gender-marked words against the ST output. Accordingly, we first calculate the **Term Coverage** as the proportion of gender-marked words annotated in the MuST-SHE references (either in the correct or wrong form) that are actually generated by the system, on which the accuracy of gender realization is therefore *measurable*. Then, we define **Gender Accuracy** as the proportion of correct gender realizations among the words on which it is *measurable*. This evaluation method[7] has several advantages. On one side, *term coverage* unveils the precise amount of words on which systems' gender realization is measurable. On the other, *gender accuracy* directly informs about systems' performance on gender translation and related gender bias: scores below 50% indicate that the system produces the wrong gender more often than the correct one, thus signalling a particularly strong biased behaviour.

4 Results

In Table 3 we present the MuST-SHE^{IWSLT24} results of the 13 IWSLT Offline ST cascade models. Starting from **coverage scores** (All-Cov), all models achieve overall positive results, which range from ∼70% (`HW-TSC_CONSTRAINED-wLLM.primary`) to 74.79% (`HW-TSC_CONSTRAINED.primary`). Hence, these models produce a good amount of

[5] Speaker's gender information is provided for each segment. Note that gender has been labeled based on the personal pronouns the speakers used to describe themselves in their publicly available personal TED section.

[6] For *direct* ST solutions that directly translate from the audio input without intermediate textual representations, CAT1 can also reveal whether such models leverage speakers' voice as an unwanted cue to translate gender. See Gaido et al. (2020).

[7] The evaluation script is publicly available at: https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/mustshe_gender_accuracy.py.

Model	All-Cov	All-Acc	F-Acc	M-Acc	1F-Acc	1M-Acc	2F-Acc	2M-Acc
HW-TSC_CONSTRAINED.primary	**74.79**	**82.99**	**84.44**	81.70	**68.18**	85.71	**87.61**	80.80
HW-TSC_UNCONSTRAINED.primary	73.93	82.52	82.96	82.12	65.22	85.71	86.61	81.30
HW-TSC_UNCONSTRAINED.contrastive	75.07	81.72	81.16	82.24	56.52	85.71	86.09	81.45
CMU_mbr_ensemble_all_50+50+50.primary	73.07	81.36	80.00	**82.73**	50.00	80.00	87.50	**83.33**
CMU_beam_5.contrastive	74.21	80.56	79.58	81.51	52.00	76.00	85.47	82.64
CMU_mbr_50.contrastive	73.93	80.21	80.14	80.28	55.17	70.83	86.61	82.20
NYA.contrastive3	72.21	79.72	77.37	81.94	39.13	**86.96**	85.09	80.99
HW-TSC_CONSTRAINED-wLLM.primary	70.49	79.70	78.63	80.71	45.45	79.17	85.32	81.03
NYA.contrastive1	72.49	79.64	77.54	81.69	39.13	**86.96**	85.22	80.67
NYA.primary	72.49	79.64	77.54	81.69	39.13	**86.96**	85.22	80.67
NYA.contrastive2	73.35	79.51	78.99	80.00	45.83	76.00	85.96	80.83
KIT.primary	71.92	77.70	78.03	77.40	43.48	65.38	85.32	80.00
KIT.contrastive1	71.92	77.42	78.20	76.71	40.91	65.38	85.59	79.17
standard dev.	±.1.3	±.1.6	±.2.1	±.1.8	±.9.4	±.7.8	±.0.8	±.1.0

Table 3: MuST-SHE^{IWSLT24} results for en-de. Systems are ranked based on overall Gender Accuracy (All-Acc). Primary model submissions in violet color.

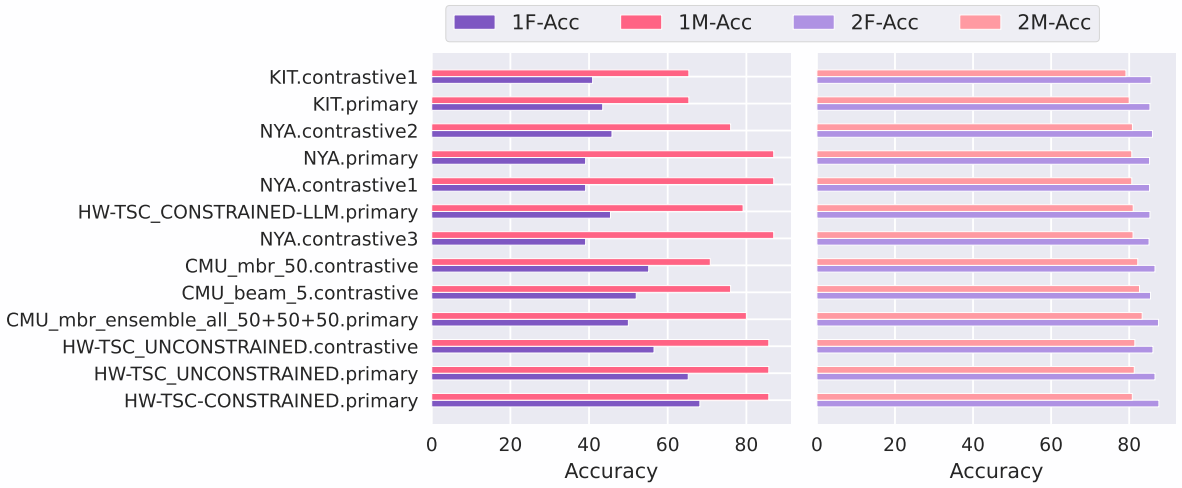

Figure 1: MuST-SHE^{IWSLT24} accuracy results across categories 1 and 2 per each gender (F/M).

gender-marked words that can be evaluated with regards to the accuracy of their gender realization.

Moving onto the **overall accuracy scores** (All-Acc), we can see that – while there is still room for improvement – all of the evaluated ST systems achieve reasonable results, by being able to correctly translate gender with an accuracy of at least 77.42% (KIT.contrastive1) up to 84.44% for HW-TSC_CONSTRAINED.primary. Similar accuracy ranges are attested also by disaggregating results across feminine (F-ACC) and masculine (M-Acc) genders. Interestingly, such results show that none of the models exhibit perfectly equal performance across both genders. Still, the divide is fairly limited, with *i)* a comparable number of ST systems achieving slightly higher results on either the feminine or masculine set of MuST-SHE, and *ii)* little variation in scores across the 13 models, as attested in terms of standard deviation. If we go more fine-grained into disaggregated results, however, we unveil a higher degree of variation.

In Figure 1, we report results across categories for masculine (1M and 2M) and feminine gender realizations (1F and 2F). On the one hand, for unambiguous gender translation from CAT2, systems are slightly better in performing feminine gender translation. Instead, results on CAT1 unveil a wide gender gap, where feminine accuracy is consistently lower compared to its masculine counterpart. In fact, most models tend to generate the correct feminine form in less than 50% of the cases, namely below random chance. The ST model HW-TSC_CONSTRAINED-wLLM.primary, which overall emerges as the best system for gender translation, still remains at 68.18%.

To conclude, our results show that – when confronted with ambiguous source sentences – current ST models tend to favour the generation of mas-

culine forms in the German target language. We acknowledge that the phenomena subject to our analysis (gender bias) are not currently accounted for in the design of ST systems, which are rather designed with the goal of optimizing overall translation quality. Towards the creation of fairer ST technology, however, we hope that our evaluation will raise awareness in the community, and encourage the development of capable models, which can equally accommodate feminine and masculine language.

5 Conclusion

This paper summarizes the results of our IWSLT-2024 Test Suites evaluation, which focused on gender bias in translation. To this aim, we have introduced the *speech* expansion of the en-de MuST-SHE test set. Overall, results on MuST-SHE^{IWSLT24} show that the evaluated ST systems are reasonably good at translating gender under realistic conditions, achieving comparable results across feminine and masculine gender translation. Also, all models are quite robust, and show a similar behaviour for translation of unambiguous gender phenomena, where they can rely on contextual gender information. However, for ambiguous cases where the input sentence does not inform about the gender form to be used in translation, we confirm a strong skew where all systems favour masculine generation almost by default. This finding calls for further research endeavours and evaluation initiatives to counter gender bias in ST and measure future advances.

Limitations

The main limitation of this work concerns the limited size of data points (i.e. gender-marked words) available for evaluation. As such, even in the case of gender performance parity, the dataset does not allow to make conclusive statements about the *absence* of bias in the assessed models. Despite its restricted size, however, MuST-SHE^{IWSLT24} provides a first glimpse into understanding and monitoring en-de systems' behaviour with respect to gender bias and translation.

Ethics Statement

The use of gender as a variable in this paper warrants some reflections. Namely, when working on the evaluation of speaker-related gender translation for MuST-SHE (i.e. Category 1) we solely focus on the rendering of their reported linguistic gender expressions. No assumptions about speakers' self determined identity (GLAAD, 2007) – which cannot be directly mapped from pronoun usage (Cao and Daumé III, 2020; Ackerman, 2019) – has been made.

Also, in our diagnosis of gender bias we only account for feminine and masculine linguistic forms, which are those traditionally in use and the only represented in the used data. However, we stress that – by working on binary forms – we do not imply or impose a binary vision on the extra-linguistic reality of gender, which is rather a spectrum (D'Ignazio and Klein, 2020). Also, we acknowledge the current challenges faced for grammatical gender languages like German in fully implementing neutral language (Paolucci et al., 2023), and support the rise of both non-binary language (Shroy, 2016; Gabriel et al., 2018; Conrod, 2020) and translation technologies (Lauscher et al., 2023; Gromann et al., 2023).

Acknowledgements

The work presented in this paper is funded by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People) and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. Also, we would like to thank the 2022 FBK internship students Sabrina Raus and Abess Benissmail from the University of Bolzano: the creation of MuST-SHE^{IWSLT24} was made possible by their work.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a Journal of General linguistics*, 4(1).

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde,

Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2873–2887, Online. Association for Computational Linguistics.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Roldano Cattoni, Mattia A. Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation.

Kirby Conrod. 2020. Pronouns and gender in language. *The Oxford Handbook of Language and Sexuality*.

Greville G. Corbett. 2013. *The Expression of Gender*. De Gruyter.

Marta R. Costa-jussà. 2019. An analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1:495–496.

Kate Crawford. 2017. The Trouble with Bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, California.

Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism*. MIT Press, London, UK.

Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. 2018. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

GLAAD. 2007. Media Reference Guide - Transgender.

Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and Katharina Bühn. 2023. Participatory research as a path to community-informed, gender-fair machine translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 49–59, Tampere, Finland. European Association for Machine Translation.

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10:1604.

Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors. 2022. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Seattle, Washington.

Marlis Hellinger and Heiko Motschenbacher. 2015. *Gender Across Languages. The Linguistic Representation of Women and Men*, volume IV. John Benjamins, Amsterdam, the Netherlands.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv:2212.04356 [cs, eess].

Beatrice Savoldi. 2023. Gender bias in automatic translation. *Università degli studi di Trento*.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Alyx J. Shroy. 2016. Innovations in gender-neutral French: Language practices of nonbinary French speakers on Twitter. *Ms., University of California, Davis*.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. *Social communication*, pages 163–187.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *arXiv preprint arXiv:2401.10016*.

Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors. 2023. *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation, Tampere, Finland.