Gender bias in automatic translation

Beatrice Savoldi

UNIVERSITÀ
DI TRENTO
Dipartimento di
Lettere e Filosofia

Universität
Augsburg
University

International Research Ph.D.
“Forms of cultural exchange”

35° cycle

Ph.D. Thesis

Gender Bias in Automatic Translation

Supervisor:
Ermenegildo Bidese
Università degli Studi di Trento

Candidate:
Beatrice Savoldi

Supervisor:
Luisa Bentivogli
Fondazione Bruno Kessler

Co-supervisor:
Claudia Claridge
Universität Augsburg

Ph.D. program coordinator:
Fulvio Ferrari

Academic Year 2021-2022

Contents

Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BLEU	BiLingual Evaluation Understudy
DL	Deep Learning
GBET	Gender Bias Evaluation Test set
IAA	Inter Annotator Agreement
IWSLT	International Conference on Spoken Language Translation
ML	Machine Learning
MT	Machine Translation
MATTR	Moving Average Token/Type Ratio
NER	Named Entity Recognition
NMT	Neural Machine Translation
NLP	Natural Language Processing
NN	Neural Network
POS	Parts-of-Speech
RNN	Recurrent Neural Network
ST	Speech Translation
SOTA	State-of-the-art
TER	Translation Edit Rate
TTR	Token/Type Ratio

List of Figures

List of Tables

1

Introduction

> *The L in NLP is Language, language means people.*
>
> — Schnoebelen (2017)

1.1 Motivation

We are currently witnessing profound changes in interactions, communication, and tasks that rely on language. Natural Language Processing (NLP) has made tremendous progress in front of our eyes, with tools that automate language-related activities at scale and with unprecedented capabilities. Such advancements were made possible by the introduction of a new neural development paradigm powered by a vast amount of digitized speech and text resources, and have converged with the demands of our increasingly interconnected society. Indeed, language tools are increasingly integrated and blended into the fabric of our everyday life and activities.

In this context, the automatic translation of text and speech has gained traction and popularity, addressing the need for multilingual communication and mediation. The last decade has seen the longstanding task of text-to-text Machine Translation (MT) achieving remarkable improvements in a variety of morphological, lexical, and syntactic aspects, with the generation of highly fluent automatic translations across several language pairs (Bentivogli et al., 2016). MT is used by millions of people on a daily basis for a variety of purposes, whether for leisure, travel, accessing foreign content online, or work-related activities (Vieira et al., 2022), simply by accessing commercial systems available online on their phones. MT applications, also, govern online multilingual queries and e-commerce. Beside the translation of textual documents and webpages, technological efforts have also targeted the audiovisual content, which represents one of the most common means of communications online. Accordingly, emerging Speech Translation (ST) solutions that directly translate speech into text in another language are on the rise, with application scenarios that can include subtitling (Matusov et al., 2019), travel conversations (Takezawa et al., 1998), or even crisis response (Bansal et al., 2017).

Undoubtedly, automatic translation tools have facilitated navigating multilingual contexts, by providing accessible shortcuts for gathering, processing, and spreading information. As language technologies become more widely used and deployed on a large scale, however, their societal impact has sparked concern both within (Hovy and Spruit, 2016; Bender et al., 2021) and outside (Dastin, 2018) the research community. The enthusiasm that has marked this technological progress can no longer be decoupled from critical work, which examines how the success and benefits of NLP do not accrue to all users equally, and can negatively impact individuals and communities (Jin et al., 2021).

A significant concern for automatic translations is gender bias: systems have been found to propagate stereotypes and favor the generation of masculine forms, potentially offering a degraded quality of service for women (Sun et al., 2019a). While gender bias affects many NLP tools, its presence in automatic translation is complicated by cross-lingual issues, also depending on the fact that languages differ in the way in which they grammatically express and encode gender references. Indeed, the issue is not purely technical, inasmuch it compromises the quality of automatic translations, but rises larger concerns on the phenomenon of mis- and under-representation of socially disadvantaged groups. Also, the issue is not completely novel either: discussions on the appropriate representations of gendered groups and their linguistic visibility have long occupied studies in linguistics and the social sciences. Accordingly, the unifying research question behind this work is: what are the social, (cross-)linguistic, and technical factors that contribute to the emergence of gender bias and how do they interrelate with one another?

When I started this PhD in 2019, research in gender bias on automatic translation was in its early stages, characterized by the first studies on the topic in MT. This thesis starts from these early contributions and expands to the investigation of gender bias in both MT and ST systems. Ultimately, the goal of my work is to contribute to this pressing area of research with an interdisciplinary perspective, to raise awareness of bias, improve the understanding of the phenomenon, and investigate best practices and methods to unveil and mitigate it in translation systems.

1.2 Contributions

This thesis addresses gender bias in automatic translation. As just outlined above, this work is intended to foster the understanding of gender bias by means of an interdisciplinary perspective, as well as inform approaches to evaluate and mitigate bias in automatic translation also by means of new resources and methods. The main contributions of the thesis can be categorized as follows.

Theoretical contributions. The relatively new area of research on gender bias in MT has grown rapidly. The issue, however, has been addressed by disparate studies, often within a strictly technical focus. To bring cohesion and advance the field, the thesis offers:

- A unified framework that critically reviews the conceptualization, sources, and effects of bias in MT in light of insights from related disciplines, so to systematize the state-of-research on the topic by identifying blind spots and challenges (Savoldi et al., 2021).

Resource and methodological contributions. To foster the evaluation and gender-aware development of MT and ST systems, the thesis presents several resources:

- MuST-SHE: a multilingual, gender-sensitive, natural benchmark for three language pairs (English → French/Italian/Spanish). By design, MuST-SHE represents qualitatively differentiated and balanced gender-related phenomena, which have been manually annotated in the corpus. By means of such annotations, MuST-SHE implements a new evaluation protocol, which conciliates *i)* pinpointed analyses on the correct translation of gendered words at several levels of granularity, and *ii)* and global assessments of overall translation quality, which are made informative in terms of bias inasmuch as they isolate systems' ability to translate gender from other factors that might affect overall performance (Bentivogli, Savoldi, et al., 2020; Savoldi and Bentivogli, 2021).

- MuST-SHE extensions: the enrichment of MuST-SHE with two additional manual annotation layers on Parts-of-Speech (POS) and agreement chains to pinpoint the effect of bias across linguistic categories and morphosyntactic phenomena (Savoldi et al., 2022b).

- MuST-Speakers: the enrichment of the English→French/Italian/Spanish portion of MuST-C – one of the largest and most widely employed ST datasets consisting of TED talks data –

with speakers' gender information. Such gender information is based on speakers' personal pronouns, which were manually retrieved from their publicly available TED bios. This resource serves the development of gender-enhanced systems (Gaido, Savoldi, et al., 2020).

- The Gender-balanced Validation Set: a validation set made available for English → French/Italian/Spanish, which features a balanced number of masculine and feminine speakers. This resource is used at training time to discourage rewarding models' biased behavior (Gaido, Savoldi, et al., 2020).

Empirical contributions. In light of theoretical insights and with the newly introduced methods and resources, this thesis investigates:

- The effect of audio cues on gender bias in ST technologies. A systematic comparison between the traditional *cascade* approach and the emerging *direct* paradigm showed that direct systems leverage speakers' voice to translate gender (Bentivogli, Savoldi, et al., 2020).

- Mitigating techniques based on data-centric approaches to improve gender translation in direct ST, also offering a solution to override speakers' voice as a gender cue (Gaido, Savoldi, et al., 2020).

- Algorithmic choices that can concur to the exacerbation of bias. A comparison among five word segmentation techniques applied to direct ST systems unveiled that state-of-the-art segmentation methods come at the cost of higher bias (Gaido, Savoldi, et al., 2021).

- The interplay among ST systems' design choices, overall performance, and gender bias, with a focus on linguistic specificity. A multifaceted evaluation conducted at several levels of granularity showed that not all POS are equally impacted by gender bias, and that larger models do not grant an advantage for feminine gender translation (Savoldi et al., 2022b).

- The emergence of gender capabilities in direct ST. A dynamic evaluation over the whole course of systems' training demonstrates that standard training procedures aimed at maximizing overall performance are suboptimal for feminine gender translation, which still shows potential for improvement when the training is stopped (Savoldi et al., 2022a).

The contributions of this thesis are the result of several activities and collaborations carried out during the three years of the PhD. A list of such activities and collaborations, as well as of all the above-referenced publications[1] included in this thesis, is provided in the Appendix A.

[1] All works are the result of collaborative and multidisciplinary efforts. As such, the development of the translation systems – used in the publications and the resulting thesis – were contributed by my co-authors.

1.3 Thesis Structure

The thesis is structured into nine chapters. The current **Chapter 1** contains the Introduction, which lays out the motivation, innovative aspects, and contributions of this work. What follows is the foundational **Chapter 2**, which guides the reader through the thesis by providing information on the three main pillars upon which this interdisciplinary work stands: *i)* ethical and social aspects of language technologies; *ii)* the relation between gender and language; *iii)* the workings of current neural approaches to automatic translation and its evaluation. Moving to the most innovative aspects of the thesis, **Chapter 3** represents a theoretical contribution, where state-of-the-art research on gender bias in MT is discussed, systematized, and critically reviewed in light of a multidisciplinary perspective and by identifying blind spots and challenges. The identified under-investigated aspects inform the empirical contributions of the following chapters, which foreground gender bias in the neglected task of audio-to-text ST. Accordingly, **Chapter 4** presents the multilingual, natural MuST-SHE benchmark and its proposed gender-sensitive evaluation method, which allows monitoring gender bias in both MT and ST. **Chapter 5** examines the implications of direct ST systems that leverage the speaker's voice as a gender cue to translate gender. Moreover, it explores different data-centric approaches and solutions to mitigate gender bias and override speech cues. Instead, **Chapter 6** investigates the algorithmic factors that can concur to the emergence of bias, by exploring the impact of different word segmentation techniques on gender translation in ST. **Chapter 7** foregrounds the impact of gender bias on specific linguistic phenomena and in relation to systems' design choices; as such, it presents a fine-grained multifaceted manual and automatic evaluation on the output of different ST systems. **Chapter 8** delves into the dynamic inspection of gender capabilities over the whole course of systems' training and how they progress in relation to generic translation performance. Such an approach allows to make ST systems' learning process more transparent and account for the integration of gender-related factors in their development. Finally, **Chapter 9** closes the thesis with remarks from the work carried out in these three years and presents future research directions.

A note on the style and terminology employed in this thesis is warranted. First, to welcome and guide readers with different areas of expertise, the foundational chapter 2 has favored a more informative and engaging tone compared to the following, more stylistically concise chapters. Second, since this work sits at the intersection of different disciplines, it has indeed been confronted with the issue of conflicting terminology across fields.[2] For consistency with the original publications included in the thesis, I largely adhere to the terminology more commonly

[2] For instance, a *phrase* is a conceptually and grammatically significant unit in linguistics, whereas in the context of statistical MT it can refer to any – grammatically relevant or not – sequence of words.

employed in the areas of computational linguistics and NLP, with explicit clarifications when needed. Finally, the thesis alternates the use of different pronouns: ‘we’ is typically employed in parts of the thesis that are the result of collaborations with other researchers, whereas ‘I’ refers to new, individual work (such as the writing of this thesis) or opinions that are not necessarily endorsed by my colleagues.

2

Foundations

> *Technology is not the design of physical things. It is the design of practices and possibilities.*
>
> — **Suchman (2007)**

During the 1978-79 UK electoral campaign, Tim Bell, the advertising executive of the Conservative Party's election campaign, realized that the main obstacle Margaret Thatcher had to overcome to win the national election was her voice. When speaking on the radio or TV, she sounded "strangulated" and "patronizing", completely unfit to gain the trust and vote of the lower-middle class. To solve the issue, he arranged a meeting with the Shakespearean actor Laurence Olivier, who taught Thatcher how to modulate her voice by lowering the pitch (Karpf, 2006). Training a performative voice was the identified countermeasure for what was actually an engineering problem. As Tallon (2019) reports, technology that transmits human voices such as microphones has historically been optimized for lower (typically male) voices, thus negatively impacting female speech and its radio intelligibility. By not identifying and tackling the design issue, the result was that women on air were perceived as "affected," "stiff," "forced," and "unnatural". Aside from audio quality, the long-term effect of such a technological limitation has been on women's public credibility, on the perceived "unsuitability" of their speech. Few women would thus make a radiophonic appearance. And when they did, the poor definition affecting high frequencies and consonants acted as a self-fulfilling prophecy: “A woman speaker is rarely a success, and, if I were a broadcast manager, [...] I would permit few women lecturers to appear” said a KDKA radio publicist (Tallon, 2019).

Time has passed, and the core technology is different. Nonetheless, although typically regarded as neutral, objective tools (Waseem et al., 2020), contemporary language technologies have been increasingly found to perpetuate analogous disparities. Indeed, many NLP tools were found to exhibit gender asymmetries (Costa-jussà, 2019): automatic speech recognition (ASR) that produces Youtube’s captions better capture male than female speech (Tatman, 2017), and machines that automatically translate across languages often fail to adequately render women’s text and mentions (Vanmassenhove et al., 2018; Hovy et al., 2020).

Language technologies are a ubiquitous component of our world and transform our interactions with it. Automatic translation allows us to access information in languages we don’t even speak. We ask the voice-controlled personal assistant in our cars to search for a number in our contact list and make a phone call. Also, we might ask such an assistant about the shortest way to reach a certain location and make our driving experience easier. However, when we think about technology, we typically approach it as artifacts, or systems that are abstracted away from their context of use and creation (Gebru, 2020). Also, we may be more inclined to observe developments merely through a technical lens. For instance, by inspecting components, functioning, and the tasks they accomplish, or by judging them against (often implicit) values such as productivity and efficiency (Birhane et al., 2020). Such a perspective however can only tell us so much about the longstanding impact that technology can exercise, if not hinder the analysis on the potential adverse outcomes. Instead, by echoing a long tradition rooted in Science and Technology Studies (SST) (Winner, 1980; Wajcman, 2007; Feenberg, 2010) that emphasizes how technologies are inherently situated within a specific cultural, political, and societal context, works in the field of NLP – and more generally autonomous Artificial Intelligence (AI) systems – have advocated the use of a sociotechnical lens (Selbst et al., 2019; Waseem et al., 2020; Benjamin, 2019). That is, technologies are regarded as *sociotechnical* systems, which do not exist in a vacuum and consist only of code, data, and architectural components. Rather, these systems interact with and affect human practices, just like human practices are implicated in the stages of creation and deployment of technology.

In this sense, the analysis of the history of voice electronics by Tallon (2019) illuminates often neglected aspects of technology, by emphasizing its impact on society and users. Tallon (2019)’s insightful inquiry waives together several considerations: it inspects development and testing procedures; it draws on phonetics and gendered communicative practices to grasp how language is acoustically transmitted and perceived; finally, it interprets such aspects in light of women's conditions in the Western 20th century. Namely, it requires an interdisciplinary framework. Such is the spirit behind this work.

This is an interdisciplinary thesis that investigates gender bias in the quintessential NLP task: Automatic Translation. Accordingly, it explores linguistic and ethical issues related to gender in the practical and applied field of NLP. In light of the above, systemic gender bias in automatic translation is not only regarded as a technical concern. By waving together the ethical, linguistic and technical aspects related to gender bias in translation technologies, this work builds on three main areas of research. I touch upon each of them before delving deeper into the experimental part of this thesis and its specificities related to gender bias in automatic translation. In this Chapter, I intend to clarify how this topic is part of a larger discussion that has invested language technologies. First, this thesis draws on and participates in the growing body of work concerned with Ethics in NLP. In Section 2.1, I thus introduce this growing research area, recently engendered by considerations on the negative impacts of automated systems. In doing so, I highlight work on bias in NLP, while emphasizing the specificity of language features for its understanding. Then, in Section 2.2 I frame the relevant background to discuss gender in language and its relationship to the extra-linguistic reality of gender. Towards this goal, I explore how gendered linguistic features interact with the perception and representation of individuals, which are deemed relevant for the recognition of gender groups and their linguistic visibility. I conclude in Section 2.3 by describing the underlying mechanisms and components of MT. This last section provides the technical background required to examine gender bias and translation in current translation technologies.

2.1 Ethical and Social Considerations for Language Technologies

May 23rd, 2016. Microsoft released Tay, an (English) conversational chatbot designed to interact with people on Twitter. More specifically, it targeted 18- to 24- year-olds in the U.S. for entertainment purposes. Its online presence, however, lasted less than 24 hours.[1] Within such a timeframe, Tay managed to spit racist comments, share sexually-charged messages, praise Hitler, and deny the Holocaust – among other types of inflammatory content. Tay was meant to be a clever experiment in AI. The bot would speak like millennials, learning from the people it interacted with on Twitter. Instead, it made ethical dilemmas surrounding AI grow thornier.

Actually, as a Microsoft spokesperson said, "It [was] as much a social and cultural experi-

[1] https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

ment, as it [was] technical" (Shah and Chokkattu, 2016). Allegedly, "a coordinated attack by a subset of people" (Lee, 2016) tweeting offensive content steered Tay's disastrous behaviors. Indeed, this case made the news and turned into one of the earliest and most notorious controversy surrounding technologies that produce language to interact with users, by revealing these tools' potential for spreading fake news, heating debates, and propagating malignant stereotypes (Weidinger et al., 2022). In the years, controversies have piled up.

Flash-forward to 2022, Amazon Alexa suggested a 10-year-old girl to touch a coin to the prongs of a half-inserted plug (Shead, 2022), a so-called "Penny challenge" that had become viral on TikTok (Prance, 2021). Meanwhile, Meta also released a chatbot – BlenderBot 3 – that would improve its communicative skills in interaction with users. As a feature, BlenderBot 3 can search the open web to talk about different topics and look for answers. Once again, it ended up spreading Antisemitic conspiracy theories, as well as conspiracies marring the 2020 US Presidential elections that allegedly scammed Trump out of the presidency (Hancock, 2022). On the one hand, these cases are raising concerns about the impact that language tools mediating our interactions with the world might have. To what extent can automatized language generation be reliably deployed? And how can such undesirable – if not dangerous – outputs affect people? On the other hand, however, the goal itself of achieving human-like language generation is also brought into question. As these technologies are modeled on human language and online activities, which aspects of human language behavior do we expect them to reproduce? Human textual productions and communications are turned into and represent the (language) data that underlies the development of current systems. But whose language? Encoding which characteristics, perspectives, and worldviews? While conversational agents have been particularly in the spotlight for the public,[2] similar questions and issues have invested several other NLP tasks and applications, if not the whole NLP field.

Natural Language Processing[3] is the research area at the crossroad between linguistics, computer science, and engineering. It is concerned with the creation of tools able to automatically process and generate language (Khurana et al., 2022) As a field, NLP has taken our world by storm. Over the last decades, it has risen from a relatively "niche academic area to an extremely influential topic of widespread industrial and political interest" (Hovy and Prabhumoye, 2021). Its boom – part of the larger so-called AI spring (Mitchell, 2021) – is attested in terms of actively developed and deployed applications around the world, public interest, and market size. How-

[2] I suspect the high level of coverage concerning these tools can be understood in light of the direct experience users have with voice assistants and chatbots. This makes their behaviors immediately tangible for many individuals.

[3] Occasionally, the term NLP and Computational Linguistics (CL) are used interchangeably. While CL and NLP often rely on the same methods and may be considered near-synonyms, I echo the terminological distinction by Tsujii (2021) to underscore different research goals. I intend CL as the development or use of computational methods for the scientific study of language (e.g., to provide an explanation for a linguistic phenomenon). NLP, instead, is motivated by technological and engineering concerns towards components and applications.

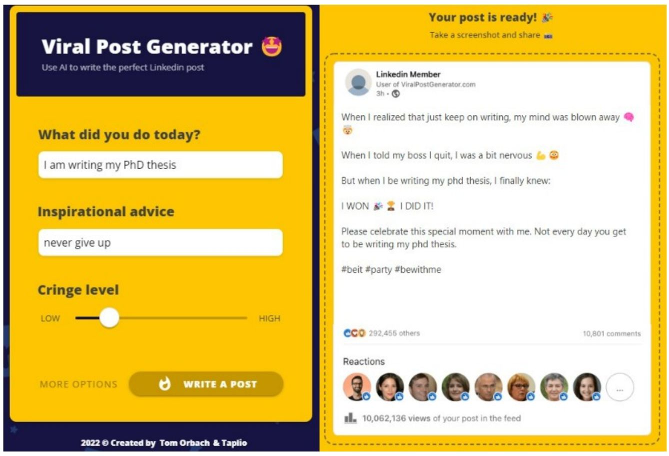
ever, as NLP has morphed into practice and into daily lives, unintended negative consequences have also emerged (Jin et al., 2021).

Earlier discussions on the societal impact of language tools were led by the landmark study by Hovy and Spruit (2016). As the authors remark, NLP as a discipline has overshadowed the involvement of human subjects in its studies and experiments, thus removing the need to consider how to minimize the risk of harm to users. That is because NLP used to mostly work with somewhat anonymous collections of texts – i.e., corpora –, which were not linked to some particular author (e.g., newswires), and with the goal of enriching linguistic analyses. However, with the widespread use of NLP also for commercial purposes, and the increasing reliance on contemporary user-generated data from social media, "the outcome of NLP experiments and applications can now have a direct effect on individual users' lives" (Hovy and Spruit, 2016). For one, reliance on user-generated data raises privacy concerns. Besides, the paper underscores the *dual nature* of technology, where e.g., a tool for textual detection of depression symptoms can help us investigate medical issues, but also be exploited for malicious targeting of fragile individuals. Also, the authors discuss how overlooking demographic and linguistic differences in NLP research might lead to issues of *exclusion*, where certain people do not fully benefit from NLP technologies.

Over the years, ethical and social considerations have expanded in the NLP as well as in the broader AI community. In 2017, the first Workshop on Ethics in NLP was introduced (Hovy et al., 2017). Inspired by the decision undertaken for the 2020 NeurIPS conference,[4] starting with ACL[5] 2021 NLP conferences now expect authors to include a broad impact statement for their work.[6] Also, individual papers can be subject to ethics reviews. To encourage transparency and appropriate scenarios of use for both data and models, sets of best practices and recommendations have been released (Gebru et al., 2021; Bender and Friedman, 2018; Mitchell et al., 2019). Also, Ethics Sheets are now dedicated to fleshing out the assumptions and ethical considerations underpinning the framing and design of AI tasks (Mohammad, 2022). At a higher level, the broad field of AI ethics has established itself as foundational. It has gathered under its umbrella crucial works on technology accountability, fairness, and – of course – bias. (Benjamin, 2019; Noble, 2018; D'Ignazio and Klein, 2020; O'Neil, 2016). Along this line, it is worth mentioning the seminal computer vision study by Buolamwini and Gebru (2018), who found huge disparities in image classification; in a range, faces from lighter-skinned men were recognized with the highest accuracy, whereas black women with the lowest one.

Within the broad horizon, I acknowledge the research being conducted in NLP-neighbouring

[4] AI conference on Neural Information Processing Systems.
[5] Association for Computational Linguistics.
[6] `https://2021.aclweb.org/ethics/Ethics-FAQ/`.



I am writing my PhD thesis

never give up

When I realized that just keep on writing, my mind was blown away 🧠🤯

When I told my boss I quit, I was a bit nervous 💪😳

But when I be writing my phd thesis, I finally knew:

I WON 🎉🏆 I DID IT!

Please celebrate this special moment with me. Not every day you get to be writing my phd thesis.

#beit #party #bewithme

Figure 2.1: LinkedIn Viral Post Generator. By prompting what you did today and including a piece of inspirational advice, the tools will generate a post that mimics the tone and style of LinkedIn posts that have gone viral.

fields and for several language applications as both inspirational and foundational for this work. While a comprehensive coverage of all ethical discussions currently undergoing NLP is beyond my scope, this portion intends to set forth the background and concepts necessary to scrutinize NLP technologies, and which will underlie the rest of the thesis. Accordingly, in Section 2.1.1, I present the social and technological context in which NLP has emerged. Section 2.1.2 highlights the often tacit assumptions underling the current paradigm for the development of NLP. Section 2.1.3 foregrounds current work on bias in the NLP community. I conclude in Section 2.1.4 by discussing the relation between linguistics and NLP, and how (socio)linguistically grounded knowledge can guide research on bias in NLP.

2.1.1 Social and Technical Context

NLP tools have entered widespread use. And yet, for a layperson, it might be hard to gauge how often we interact with these systems; examples abound. When you accept the suggestion of an autocorrector that flagged a typo in your writing, you are interacting with an application of NLP for assisted writing. When the answer to your query pops up in the upper part of a web page, that is the output of a question-answering system. Language technologies are integrated into our phones, cars, and online activities. Commercial NLP products like Amazon Echo and Google Home have entered millions of households (Perez, 2020). Day by day, a plethora of new language applications enter the public arena. Some have entertainment purposes (see Figure 2.1)[7] others, though, are put to use for sensitive contexts, such as healthcare assessment and treatment (Benton et al., 2017; Roy et al., 2021), and educational settings (Alhawiti, 2014).

[7] Available at: https://viralpostgenerator.com/?target=998942e43ce445e7a8ed43cc51862f42&params=%7B%7D.

NLP has become ubiquitous and permeates corners of our lives we might be even unaware of. As Weiser (1999) remarked, “The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.” Content moderation, spam filtering, intelligent web search integrated into our online queries, targeted advertising: albeit to different degrees, NLP tools are employed in all of these areas and more. In this way, NLP influences our information diet and social media presence (Mittelstadt et al., 2016). While the footprint of everyday applications might be frivolous if taken individually, collectively, they have a pervasive role and a powerful effect in shaping our culture (Selbst et al., 2019). Such a role transpired also with the COVID-19 health crisis. Indeed, when the pandemic broke out, NLP fostered the development of question-answering systems – CAiRE-COVID (Su et al., 2020) – and the detection of COVID-19 misinformation online (Hossain et al., 2020).

The space that language technologies have come to occupy as an integral part of our society is most likely the result of converging factors. Scientific and technological advances have met an increasing trend towards automatization.[8] With the transition to Industry 4.0, the digital processing of text and speech has become a vital component of the so-called ‘Information Age’. Within this landscape, the scope and range of NLP tools have wildly expanded. Progressively, scientific interest in computationally-aided theoretical inquiries into the nature of language has weakened. Instead, the (academic and industrial) field has shifted towards a stronger focus on applications and *predictive* models, by heavily benefiting from advances in machine learning (ML).

Debates between rationalist and empiricist approaches to the computational processing of language had long characterized the NLP field (Church, 2011; Gold, 2011). The 1990s and 2000s, though, attested to the replacement of knowledge-driven, rule-based approaches to language for the rise of data-driven, statistically-based machine learning methods (Jones, 1994; Manning and Schutze, 1999). Namely, “any methodology and set of techniques that can employ data to come up with novel patterns and knowledge, and generate models that can be used for effective predictions about the data” (Van Otterlo, 2013). Undoubtedly, the change towards an empirical paradigm was fostered by the increasing amount of available electronically stored data to which these methods can be applied; the most obvious example being the World Wide Web (Mitkov, 2022). In other words, there is no need to “explain” and “teach” a model how to perform a task. From the data themselves, the system will extract relevant patterns upon which its predictions can be based, without being explicitly taught how to do so. Data-driven, predictive methods prospered around 2010 when deep learning – which utilizes brain-inspired

[8] Consider automatic decisions making (ADM). Does this candidate show the right attitude for the job? (Naim et al., 2015). Given this set of charges, what should the prison term amount to? (Chen et al., 2019). These are just some of the (to say the least controversial) questions we have delegated to automated systems.

artificial Neural Networks (NN) – emerged as the dominant technique from the family of ML approaches (Mitchell, 2021). Deep neural networks had been around since the 1970s, but only recent increases in computational power and parallelization[9] (Raina et al., 2009) matched with the contemporary availability of large datasets provided the necessary ingredients to train NNs (Otter et al., 2020).[10] Conceptually, the bottom-up approach remains the same as other ML techniques. Feed a neural network with x-rated *vs.* family-friendly books, and it will: *i)* learn to discriminate between the two sets, *ii)* rely on the identified distinction to classify any new book we provide as input. However, deep learning proved comparatively superior at sorting and generalizing identified patterns, also from more complex, messy, and less structured data that humans couldn't possibly analyze and pin down into a set of rules.

Deep neural networks (Hinton et al., 2006; Bengio, 2009) are what powers all the major AI advances we have seen in the past decade, including image recognition, game playing, and protein folding. With astounding breakthroughs, deep learning approaches to NLP (Goldberg, 2017) have faced several arduous challenges, dramatically increasing the linguistic abilities of language technologies (Baroni, 2020). International, multilingual commerce is allowed by MT being integrated by platforms such as eBay. Improvements in speech recognition systems have been put to use to grant accessibility of multimedia content, such as automatic captions and subtitling. Benefits brought by language technologies are undeniable, and this work is by no means intended as a *tout court* criticism of current NLP approaches and garnered success. However, all that glitters is not gold.

While NLP technical advancements and NLP deployment have progressed relatively rapidly and optimistically, the understanding of their vulnerabilities and shortcomings in light of their societal impact has lagged behind. The promise of deep learning in the field has been achieving better performance by models that require more data but less (human) expertise to be trained and operate. But what are the underlying implications of such a promise? Has our notion of "better performance" been sufficiently long-sighted and validated?

2.1.2 Data-Driven Neural NLP: Assumptions and Shortcomings

The advances of data-driven neural models have consolidated the strong, bottom-up empiricist approach that now defines NLP. Such a turn has contributed to the mainstream characterization of NLP models as (fully) autonomous systems. Their hallmark is the alleged reduction of the

[9] Parallel computing allows several calculations or processes to be carried out simultaneously.

[10] NNs consist of several layers of computation, sometimes with billions of trainable parameters that require huge quantities of data and power to adjust themselves. More details on NNs and their functioning are provided in Section 2.3.1. Hereby, we are interested in providing the reader with a sense of their core assumptions and mechanisms.

human component, where neural networks "learn" by themselves how to successfully perform a task by sorting and spotting relevant features based on provided data. The absence of explicit human-defined rules and input has covered current approaches with a veneer of objectivity. Since the systems adjust themselves as a result of a direct observation of (factual) data, they are expected to rise above specific scientific frameworks or partial guiding principles on how language should be treated and how decisions in a model should be carried out. This is the argument that has often been used to justify the reliance on data-driven NLP methods for automatic decision-making, to prevent human subjectivity, preconceptions, or even prejudices from affecting the decision-making process (Barocas and Selbst, 2016).

From a broader perspective, the ML revolution powered by gazillions of data – i.e., big data – was hailed by some as the "end of theory" (Graham, 2012). Namely, as Crawford et al. (2014) reconstruct: "Some big data fundamentalists argue that at sufficient scale, data is enough; – algorithms find patterns where science cannot (Anderson, 2008)". However, such a stance disregards that even radically empiricist paradigms have epistemological commitments (Floridi, 2012), or in other words, 'big data is theory' (Crawford et al., 2014; Paullada, 2021). For one, this paradigm lies on the premise that data constitute an "exhaustive" and "neutral" representation of the reality of the phenomenon we aim to model. The pretence of objectivity is alluded to by the fact that the interpretation of the data is left to the model, which however exploits correlations without guarantee of causality. The employment of ML and neural networks for NLP stands on a set of implicit, often tacit assumptions. Their discussion is fundamental for inquiring the current issue of bias in NLP, whereas their disregard has proven scientifically and socially costly.

Some crucial issues with deep neural networks and ML are best exemplified by a hacker koan[11] surrounding Marvin Minsky — a pivotal figure in the landscape of AI research in the 70s – and the MIT student Sussman, who was training a neural net to play Tic-tac-toe. The net was randomly wired so that it wouldn't have any preconceptions of how to play: it had to come up with its own way, according to Sussman. In the koan, at that point Minsky shut his eyes: " – Why do you close your eyes? – Sussman asked his teacher. – So that the room will be empty." Although the metaphor is exquisite, the factuality of such a scene is uncertain. Allegedly, what Minsky really said was: "Well, it has [preconceptions], it's just that you don't know what they are." (Raymond, 1996).

The rationale behind neural self-assumed patterns is opaque to analysis. Neural networks are essentially black boxes that can be seen in terms of inputs and outputs. However, their inner functioning – which is based on non-linear functions and carried out in what are explicitly

[11] Hacker koans are funny short stories – often with a hidden exemplary meaning – about computer science culture. Several koans related to US computer science labs have been collected in the Jargon file (available at: http://www.catb.org/jargon/html/) and published by (Raymond, 1996).

called “hidden layers” – does not lend itself to transparent inspection. Namely, it’s hard to learn how systems learn.[12] In lack of the ability to provide human-understandable explanations for the predictions of a model, assessing whether the evidence it generates is inconclusive or not is a crucial endeavor. For instance, stylometric and text categorization results show that an author’s gender can be discerned in texts with relatively high accuracy, but due to which factors? (Koolen and van Cranenburgh, 2017). Are those tools actually distinguishing an author’s gender based on stylistic features, or have they just learned to discern authors based on spurious cues and correlations, e.g., women’s novels are printed more often by publisher x whereas man’s by publisher y?

Concurrently, tracing back such cues to the features of the training data is non-trivial, also because the size of current datasets hinders the thorough scrutiny of their content (Bender et al., 2021). However, as mounting evidence has shown, such data are far from an impartial representation of reality (Paullada et al., 2020), and rather entrench historical, social, and political inequalities. If the extent to which current NLP models are able to generalize beyond their training data is currently under debate (McCoy et al., 2021), it is nonetheless quite clear that current data-driven approaches are intrinsically bound to those same data. In this light, consider the quite straightforward – and yet real – case of an application trained to vet job applications for STEM positions on the basis of past records (i.e., CVs) over a 10-year period. Most resumes were from men, as a reflection of the current climate in the tech industry. As a result, when predicting the most likely successful candidates, the tool downgraded women’s applications. Rather than solely focusing on qualifications and skills, analyses revealed that the application would penalize explicit mentions to women (e.g., as in “women’s chess club captain”), or even more subtle cues, such as having attended all-women’s colleges (Dastin, 2018). Such predictions can seriously dismiss opportunities and feed into prejudicial commonplaces over women being unqualified for certain roles. Not only: the output of such a CV-screener would have generated the next set of data to consult for hiring decisions, as in a self-reinforcing loop.

Overall, rather than impartial, neutral tools, language technologies can replicate and feed into human-like biases and inequalities. Their investigation – to which we now turn – foregrounds human involvement and intervention.

2.1.3 The Trouble with Bias

The study of bias has come to occupy a central place in NLP, which has seen the topic growing apace into a rich area of research. Investigations on gender bias alone can attest to this exponential

[12] Crucially, it is opaque to the people developing the systems themselves, not only the lay public. As Matthias (2004) underscores, one of the implications of such a dynamic is the attribution of (moral and legal) responsibility. In principle, the manufacturer/operator of the machine is not capable of predicting its future behaviour.

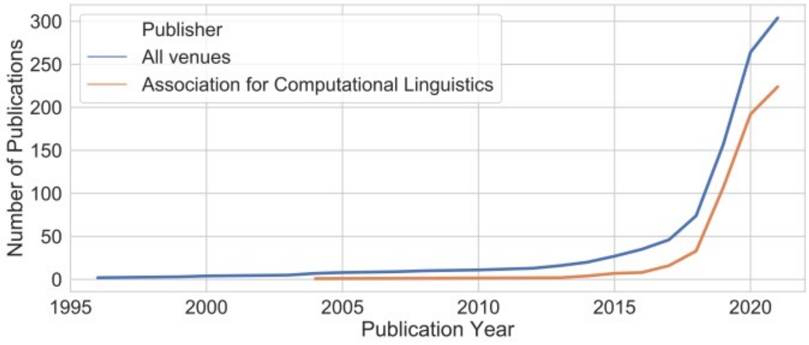


Figure 2.2: Cumulative number of papers published on *gender bias* prior to June 2021. Image from Stanczak and Augenstein (2021).

rise of interest: Figure 2.2 witnesses a spike in the number of papers published on the topic after 2019. However, it is worth clarifying that the relation between social biases and NLP is actually twofold. On the one hand, NLP methods have been used for computational social studies to detect how inequalities and stereotypes are expressed in language for several contexts, e.g., in movie plots (Madaan et al., 2017) or in the news (Garg et al., 2018). On the other hand, however, – and more central to this thesis – NLP applications can embed and perpetuate societal biases along the axis of race, gender, orientation, or class; their negative impact largely concerns the most marginalized and disadvantaged segments of the population.

While focused research in this area is still quite in its infancy, it has already produced a vast literature, exhibiting disparities in models' output (Sheng and Uthus, 2020; Rudinger et al., 2018) and internal representations (Bolukbasi et al., 2016; Nissim and van der Goot, 2020). Evidence includes sentiment analysis tools that process the word "Mexican" as being negative and associated with criminality (Speer, 2017); image captioning models that do not detect women sitting next to a computer (Hendricks et al., 2018); or tools favoring texts from wealthier, educated, and urban ZIP codes as examples of "appropriate language", even if unaligned with sensible parameters such as factuality or literary acclaim (Gururangan et al., 2022). This growing body of work has exposed structural issues: technology assumed to assist everyone actually offers unequal performance across demographic groups; also, it can reproduce[13] controversial and stereotypical associations.

To identify and tackle social biases in NLP, first efforts have been put into the creation of dedicated testing procedures (Webster et al., 2019, among others). Technical fixes and countermeasures have also promptly flourished. That is, mitigating and debiasing strategies aimed at adjusting models' predictions (Bolukbasi et al., 2016; Zhao et al., 2018c). To grasp

[13] Or amplify. Consider Zhao et al. (2017), showing how the activity *cooking* is over 33% more likely to involve women than men in their data used for the development of visual semantic role labeling models, which predict *who* is performing *which* activity in pictures. The models' prediction, though, increases such an association to 68%.

the extent of the issue, rather than directly introducing *new* models and debiasing methods, several works have retroactively put under scrutiny several *existing* resources, design choices and components, as well as standard procedures that underlie the creation pipeline of current technologies. Accordingly, data, the backbone of today's NLP, have been at the center of inspections, to expose how size does not guarantee diversity (Bender et al., 2021). Indeed, non-standard varieties are seldom included. Women are under- and misrepresented; even in Wikipedia texts, which are regarded as a high-quality resource in NLP (Sun and Peng, 2021). A contributing factor for such a lack of diversity can also be traced back to procedures for data collection and cleaning, which risk filtering out English texts written outside the US and from queer communities (Dodge et al., 2021).[14] Also, in the context of supervised learning,[15] for hate speech detection – where human data labeling (i.e., annotation) is required – it was found that annotators unfamiliar with African-American Vernacular English (AAVE) were themselves biased, and thus inclined to flag tweets with AAVE features as offensive. And this is just at the beginning of the pipeline. As I will discuss in detail in Chapter 3, such imbalances are to be examined within the definition of the task, in their interaction with models' architecture, the cascaded effects of such imbalances during training, and – crucially – at deployment time.

Overall, although the response of the NLP community has been consistent, several scholars pointed out that under many circumstances current attempts at addressing bias have been uncoordinated (Stanczak and Augenstein, 2021; Blodgett et al., 2020). For one, bias[16] has been used as an umbrella term for a range of all somewhat undesirable, but disparate models' behaviours and outcomes. Also, its conceptualization drastically varies across NLP tasks and papers. Coreference resolution[17] tools that hardly link the pronouns *she* with the occupational noun *doctor* (Zhao et al., 2018b); syntactic parsers degrading their performance on women's writings (Garimella et al., 2019); and the already mentioned (section 2.1.2) case of automatic résumé filtering that discards women's jobs applications regardless of their qualifications (Dastin, 2018). These are all commonly cited examples of gender bias. But in which way is gender used as a relevant variable? Does it concern texts written by women or (pronominal) references about women? Also, the impact of such models' behaviors is different: stereotypes raise long-term concerns regarding the representation of women, whereas the biased output of the CV screener can generate an immediate detraction of opportunities that amounts to discrimination (Crawford,

[14] It is crucial to underscore that such techniques are typically regarded as innocuous. Also, their reduction of datasets diversity is an unwanted consequence rather than a clear, willing design direction.

[15] Supervised learning (Caruana and Niculescu-Mizil, 2006) is an approach to ML that relies on human-labeled training data already sorted into categories to improve the predictions of a model. In this specific case, the model is fed with already distinctive examples of toxic *vs.* non-toxic language – judged as such by people – to learn from in order to detect hate speech.

[16] See chapter 3 for more on this point.

[17] Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

2017). But what is the negative impact for the syntactic parser?

On this basis, the review by Blodgett et al. (2020) takes stock of the situation by examining how NLP works on bias present "motivations [that] are often vague, inconsistent, and lacking in normative reasoning". Instead, the analysis of bias should forecast which people could be impacted, under which scenarios, and to what extent, also by recognizing the relationships between languages, their speakers, and social hierarchies. Along this line, consider language technologies that don't properly work on AAEV as an example of racial bias (Koenecke et al., 2020). A directly impacted stakeholder (Bender, 2019b) would be a speaker of such a variety who decides to rely on an ASR system, only to find out that it does not properly recognize their AAEV dialect. Depending on the context, such a malfunction could just imply a loss of time or a sense of frustration for the user.[18] Higher-risk scenarios, however, include dire errors in the context of medical voice-dictation software (Rodger and Pendharkar, 2004), or potential safety implications of voice recognition in cars. However, speakers of such an English variant can also be indirect stakeholders (Bender, 2019b), i.e., they can be indirectly affected by language technologies they do not personally choose to use. For instance, Twitter posts are being processed for health tracking (Jurgens et al., 2017), or to gather the community's input on public policy matters (Belkahla Driss et al., 2019). Clearly, social media are not (yet) the sole nor the central channel that enables health monitoring and community feedback. Still, if the employed NLP techniques fail to recognize the AAEV dialect, then this amounts to a contributing factor for potential delays of health interventions in the Afro-American community and its political underrepresentation.

Finally, as mentioned above, investigations of bias in NLP ought to include the recognition of the mutual relationship between language, individuals, and societal context. Accordingly, still focusing on AAEV racial bias, its analyses should be rooted in the long line of studies on (racio)linguistic ideologies in the US (Rickford, 2016). Research in sociolinguistics and linguistics anthropology highlights that, while variation is the natural state of language, people have metalinguistic beliefs and attitudes about certain varieties and dialects. However, what is qualified and perceived as "standard", "unmarked" or "prestigious" depends on social, hierarchical reasons rather than linguistic justifications (Campbell-Kibler, 2009; Preston, 2009). Such language ideologies, in turn, can play a key role in reinforcing negative attitudes not only towards the language variety itself but also towards its speakers (Loudermilk, 2015; Rosa and Flores, 2017; Craft et al., 2020). Along this line, AAVE has been historically stigmatized, considered offensive, grammatically sloppy and simple, and its speakers decried as ignorant

[18] Along this line, Mengesha et al. (2021) found that AAEV speakers would accommodate their communication with an ASR system: "I modify the way I talk to get a clear and concise response [...]. What usually works for me is when I talk real clear, and don't use slang words like my regular talk."

and lazy. The behaviors of NLP tools can thus feed into a larger phenomenon, where systemic dialect discrimination towards AAVE has been unveiled in the education system (Williams et al., 1971), courtrooms (Rickford and King, 2016), and workplace (Henderson, 2001; Grogger, 2011). Accordingly, bias in NLP tools raises issues that are not really new. However, in light of language technologies' pervasiveness and opaqueness, they can scale to a higher level and without possibility of recourse.

2.1.4 On the Interaction with Linguistics

Indeed, the NLP field is still adjusting to its recent success and to the new application scenarios, users, and responsibilities it brings. Admittedly, though, many of the challenges and concerns surrounding bias in current NLP can be traced back to a certain degree of "overlook" of the complexities of language as a phenomenon, treating it as homogeneous, static, and deprived of social factors. Now, a toolbox on how to address bias is just unimaginable. The topic is evolving, as well as our understanding of it. However, I echo Blodgett et al. (2020) and argue that a necessary starting point for studying bias in NLP ought to acknowledge and take advantage of the long-established lines of research grounded in the field that shares the same object of study as NLP: linguistics. Though its bond to NLP has weakened within the current ML approaches to language technologies, several sub-disciplines of linguistics have confronted aspects of language similar and informative to those surrounding bias in NLP.

For instance, linguistic typology systematically investigates how language systems and structures vary around the world (Greenberg, 1963; Comrie, 1989, *among others*). Today's NLP strives towards tools that are "language-independent". That is, the same techniques can be used to process different languages, e.g., regardless of their morphosyntactic features. However, – even though inadvertently – current NLP often *speaks English* (Bender, 2019a) and incorporates decisions primed from its structures (Bender, 2009; Gerz et al., 2018). In this regard, the integration of typological knowledge in NLP can prove useful for e.g., data selection and preprocessing (Ponti et al., 2019) as well as to help us unveil the vulnerability of NLP systems in multilingual scenarios (González et al., 2020), including bias. Additionally, typology informs the portability of mitigating approaches to languages other than English (Gonen et al., 2019).

Language use evolves constantly and dynamically, with any framing of language use being necessarily a moving target. As lexicographers and the field of pragmatics know well, meaning evolves and is negotiated within a social context (Tomaszczyk and Lewandowska-Tomaszczyk, 1990; Fellbaum, 2014). A case in point regards racial and sexist slurs, whose derogatory connotations can be positively reclaimed over time.[19] (Bianchi, 2014). Thus, recognizing the

[19] Consider *suffragette* as an example of a derogatory term now fully reclaimed (Steinmetz, 2015).

potential distance between the time of deployment and development of NLP systems becomes crucial. Otherwise, for instance, models that detect toxicity as learned from past data are deemed to maintain the fixed set of communicative intentions they were exposed to (Castelle, 2018), with the risk of censoring the communities leading the reclaiming process. By drawing on this knowledge, NLP technologies can cope with such mismatches by developing dynamic approaches to benchmarking (Kiela et al., 2021) and moving on to the incorporation of pragmatic factors in the development phase (Castelle, 2018).

As already introduced in 2.1.3, and hinted at throughout this section, even within assumed monolingual contexts, language variation is constant and inevitable. On the surface level, variation can be exhibited through pronunciation (e.g., UK ‘prɪvəsi/ vs. US ‘praɪvəsi/), lexical choices (e.g., US buddy *vs.* AU mate), or syntactic constructions (e.g., "I ain't got no time", double negatives as a typical feature of AAEV). More profoundly though, language variation carries indirect information about speakers' location, origin, education, or age (Eckert, 2017). Crucially, language use can be actively used to index group membership (Coupland, 2007; Bucholtz and Hall, 2004). Indeed, sociolinguists have historically departed from the idea of language homogeneity (Labov, 1972; Eckert and Wenger, 2005, *among others*). Rather, sociolinguistics and adjacent studies in linguistic anthropology foreground the relation among language, culture, and sociodemographic factors. This includes the understanding of how language shapes social identities (e.g., via personal pronouns) (McGlashan and Fitzpatrick, 2018), as well as the value attached to certain speech habits (e.g., high prestige and social status associated with the /r/ pronunciation in *car* (Labov, 1966)). Thus, sociolinguistic insights can provide relevant notions to frame human language as more than form and sound (Bender and Koller, 2020), and insights to articulate linguistic variability beyond the more general notion of *domain* (Plank, 2016).[20] When applied to NLP, such notions help us recognize the harms that can ensue from language technologies (Hovy and Spruit, 2016), and translate into the creation of demographically-informed tools (Nguyen et al., 2021; Hovy and Yang, 2021).

Finally, our utterances and communications do not only convey plain "information": *what* is said and *how* something is said can be strictly intertwined. In fact, utterances are situated and can reflect specific perspectives. The sentences "X is killed" vs "Y killed X" frame differently the visibility of the victim, killer, and the act of violence itself (Minnema et al., 2021). In this area, psycholinguistic studies explore the mechanisms that processing such phrasing can trigger (e.g., hindering the responsibility of the perpetrator (Bohner, 2001)). Analogously, the linguistic choices we make can reflect our (even implicit) worldviews and expectations (e.g., by distinguishing between "my friend" and "my *gay* friend"; "doctor" and "*woman* doctor"). A

[20] Notion used to interchangeably refer to variation in genre, context, or terminology.

longstanding line of psycholinguistic research has in fact unveiled how language is one of the most powerful means of stereotype transmission (Sczesny et al., 2018; Beukeboom and Burgers, 2019). While subtle, these mechanisms are pernicious. Recognizing the relation between expressions and percepetions of an utterance, sheds light on how NLP models can perpetuate negative generalizations towards social groups. Accordingly, knowledge of this relation can e.g., inform model reasoning (Sap et al., 2020), be used to spot asymmetries in language data (Minnema et al., 2021), and adjust them (Ma et al., 2020).

In light of the above, several disciplines within the field of linguistics can offer fundamental insights to guide and support research on bias and inequalities in NLP. While how to integrate and to what extent such far-reaching knowledge with technical aspects is up for examination, I believe that this is the most promising route forward. Indeed, as this thesis is concerned, it is for the study of gender bias within the multilingual scenario of automatic translation. Thus, in the next section, we move on to grounding the study of gender and language(s). As we will see, starting with an anecdote, the relation between gender as a social and (cross)linguistic phenomenon is not immediately self-evident, and not only for machines.

2.2 Gender and Language

Of all the accounts on second language learning, Mark Twain's may be the most amusingly unsuccessful. In the semi-autobiographical collection of essays 'A Tramp Abroad', the American writer recounts his frustration with the German language. Twain complains about compound words as long as alphabetical processions, and inhuman ways of cutting up verbs. The most memorable rants, however, are unleashed against the seemingly absurd German gender system:

> *Every noun has a gender, and there is no sense or system in the distribution; so the gender of each must be learned separately and by heart [...] To do this, one has to have a memory like a memorandum-book. In German, a young lady has no sex, while a turnip has. Think what overwrought reverence that shows for the turnip, and what callous disrespect for the girl.*
>
> **— (Twain, 1880)**

Besides being a hilarious read, Twain's take is insightful for anyone that wishes to venture into the study of gender and language. It should thus come as no surprise that this passage is frequently used to illustrate the seemingly opaque relationship, or perhaps sense of mismatch,

[21] DaLL-E is a system developed by Open AI that generates images from a text description in natural language. Available at:https://openai.com/dall-e-2/.

Figure 2.3: How would a 'sexed' turnip look like? Three images generated with DALL-E[21] given the prompt: *an abstract pencil and watercolor art of a female turnip*.

between our conventional understanding of 'gender' and grammar. As a matter of fact, the joke cleverly exploits inherent cross-lingual differences concerning the linguistic category of gender, so to twist them and play with the relationship between linguistic forms and the extra-linguistic reality of (human) gender. Consider German, where the word for 'young woman' (*das Mädchen*) is classified as neuter, whereas a 'turnip' (*die Rübe*) is feminine. For German speakers, the formal classification is decoupled from the conceptual nature of the referents (e.g., animate/inanimate, or in Twain's words, 'having or not sex'). Twain, as an English speaker, is however inclined to overlap the formal and conceptual, thus assimilating feminine German forms to *she* and neuter to *it*. Hence, the comical idiosyncrasy: why should a vegetable be female? And how could that be? We actually try to visualize it in Figure 2.3.

As Matasović (2004) puts it, gender is perhaps the only grammatical category that has ever stirred passion – and not just among linguists. Scholars have long considered gender to be an endlessly perplexing yet fascinating category (Corbett, 2013a), with highly debated origins, properties, and functionalities. Despite Twain's impression, grammatical gender is not completely arbitrary, and can rather correlate with feminine/masculine social categories, especially in the case of personal nouns (Dixon, 1982; Corbett, 1991, 2015). Indeed, it is precisely this intersection with interpersonal and social factors that makes it fascinating but also a challenging object of study, hardly to be restricted within purely formal analyses. Given its links to the real world, gender is a feature that speakers are partly aware of. Discussions about the "proper" or suitable usage of gendered language have in fact sparked the interest of linguists and non-linguists alike, in particular towards the appropriate recognition of gender groups and their status.[22] Appropriately and funny enough, even Twain mentions a sense of impoliteness or 'disrespect' that comes from attributing the wrong gender. Little did he know about the discussions that would later take place around the relation between language and

[22] For the interested reader, Conrod (2020) draws on McCready (2019) to adopt a socio-pragmatic approach and compare the appropriate use of *honorifics* to gendered language, as an analogous linguistic category that encode and conceptualizes social identities/relationships.

gender, and more recently, the concept of misgendering, i.e., the use of wrong pronouns or other gender-specific words when speaking or referring to someone (Conrod, 2019; McLemore, 2015).

As the reader probably already realized, the interplay between gender and language is a complex and multifaceted matter, which has been addressed from multiple standpoints. Indeed, to confront gender bias in MT, it is vital to first reach out to the body of research that has explored this interplay, by foregrounding how human gender interacts with language(s), translation, and implicit biases and inferences in our discursive practices. Admittedly, the examination of these relationships is first and foremost complicated by terminological and conceptual confusion. As Kaplan and Glover (2001) note, gender is "a busy term" with contentious meanings across disciplines. Accordingly, I start in Section 2.2.1 by laying out the necessary distinction between the notions of gender in linguistics and social sciences, as well as the notion of sex. With these distinctions in mind, in Section 2.2.2 I outline the different linguistic structures that are used to refer to the extra-linguistic reality of gender, and how they vary across languages. Then, in Section 2.2.3 I explore how gender is assigned and perceived in our verbal practices depending on contextual factors as well as assumptions about social roles, traits, and attributes. Finally, in Section 2.2.4, I introduce a line of research that has explored the tension between gender and language through speech practices, where language is viewed as a means for articulating and constructing personal identities.

2.2.1 Gender, Sex, and Grammar: Main Elements

The obstacle that we need to overcome to venture into the study of gender and language is the troubled relationship between overlapping terms and only partly interrelated concepts that this endeavor poses. Indeed, we are not the first to be confronted with this conundrum. As Di Sabato and Perri (2020) point out, "Gender is a polysemous term, often considered slippery and ambiguous, since it can be seen in both sociocultural terms and in terms of language as an abstract system." In the upcoming paragraph, we first need to detach ourselves from discussions about grammatical gender classes from a purely typological and abstract standpoint and rather explore how the social, extra-linguistic notion of gender is encoded and represented in language. Also, as we will see, the understanding of gender is often confounded with the "biological" classification of humans, based on the male-female categories. Although we reiterate on bimodal feminine/masculine linguistic forms and social categories, we emphasize that gender encompasses multiple biosocial elements not to be conflated with sex (Risman, 2018; Fausto-Sterling, 2019), and that some individuals do not experience gender, at all, or in binary terms (Glen and Hurrell, 2012). As Cao and Daumé III (2020) point out, linguistic, social, and biological categories might be interrelated, but they do not even remotely one-to-one map with each other.

2.2.1.1 Gender and Grammar

In linguistics, gender has a well-established technical sense, where it refers to a grammatically significant classification of nouns. The word ‘gender’ itself – which derives from Latin ‘genus’ via Old French ‘gendre’ – originally meant ‘sort’ or ‘kind’.[23] This etymology reconstruction guides us into the most basic understanding of ‘gender’ from a linguistic perspective. Namely, a noun-class system[24] (Dahl, 2004; Trudgill, 2011), where each noun is assigned to a value, i.e., gender. While most Indo-European languages commonly distinguish between *feminine*, *masculine* or — occasionally – *neuter* gender, in principle, many criteria could be used as the basis for noun classification (e.g., *common*, *animate*, *inanimate*). Indeed, grammatical gender distinctions are widespread across the languages of the world. According to a recent typological sample, they occur in 40% of the world’s languages (Corbett, 2013b). These systems, however, can highly vary in terms of number and type of classes, as well as assignment criteria. The hallmark of any grammatical gender system is however agreement (Corbett, 2006, 2013a), as best formulated in the longstanding definition by Hockett (1958): “Genders are classes of nouns reflected in the behavior of associated words.” In fact, although gender is a property of nouns, any other linguistic element dependent on the noun must take the same gender value, i.e., must agree in gender (Bloomfield, 1933; Hockett, 1958; Dixon, 1982; Matthews, 1997). To exemplify, consider the word ‘chair’ in Spanish (*silla* FEM), whose membership to the feminine class is reflected in the form of its related article and adjective (e.g., the old chair: la FEM vieja FEM silla FEM).[25] Accordingly, grammatical gender comprises formal and semantic features that are morphosyntactically defined (Ritter, 1993; Comrie, 1999; Schriefers and Jescheniak, 1999). These features, however, are properties of the morphemes themselves and may be independent of the properties of their real-world referents.[26] Hence, the ‘bed’ is formally FEM in Spanish (*la cama*), but MASC in Italian (*il letto*), and NEUT. in German (*das Bett*).

Compared to other grammatical categories such as *tense*, or *number*, grammatical gender

[23] See the entry for *gender* in the OED: `https://www.oed.com/viewdictionaryentry/Entry/77468`.

[24] In the wake of Corbett (1991), “grammatical gender” and “noun-class system” are here used synonymously (see also Dixon (1982)).

[25] Grammatical gender systems can be distinguished in: *i)* overt – where the form of the noun itself reveals its class membership, as for *silla* (chair) in Spanish; or *ii)* covert – where class membership mostly surfaces via adjectival or article agreement, as for German e.g., *der Stuhl* MASC. In German, the article (*der, die, das*) is generally taught along with a noun to memorize its gender. Hence, the additional difficulty in learning gender for each noun “by heart” lamented by Twain in Sec. 2.2

[26] Such a distinction has opened the door to the psycholinguistic investigation on the extent to which the perception of inanimate referents is semantically influenced by grammatical gender (Konishi, 1993; Sera et al., 2002). See Samuel et al. (2019) for a review.

does undoubtedly stand out. For one, unlike tense and number, which respectively convey time reference and notions of quantity, it is unclear what grammatical function is associated with the division of nouns into two, three, or more classes (Bittner, 2000). Also, tense and number are productive and functional inasmuch they offer a paradigm to choose from for each word (e.g., go vs. went; book vs. books). Differently, as Leiss (2000) reconstrues, gender behaves like a "category without a paradigm." For instance, in German we cannot choose among *der Blume (MASC), die Blume (FEM), and *das Blume (NEUT); each word is just assigned to a single gender value.

These issues are so puzzling that, according to some scholars, gender is little more than "an accident of linguistic history" (Ibrahim, 1973). Others have attempted to trace back old forms and historical developments, to a time when gender seemed to be linked to size (Lehmann, 1958; Leiss, 2000) and displayed paradigmatic options. Defenders of functionality have stressed the fact that gender agreement can help to keep track of referents across a stretch of discourse (Heath, 1975; Lyons, 1977; Contini-Morava and Kilarski, 2013). On a critical note, this effect is often overrated in languages that only have two or three genders as the disambiguating power of gender can only be convincing in languages with a larger number of gender values (Audring, 2016). To the best of my knowledge, the debate is still raging. Some scholars view gender as a "totally non-functional" (Trudgill, 1999) or "marginal" grammatical category (Trudgill, 2011), not marking "any real-world entity or category" nor serving "any communicative need" (McWhorter, 2001). As more of a middle ground position, others have seen gender as possessing (albeit partial) relevance in terms of meaning-making (Romaine, 1997; McConnell-Ginet, 2013). Indeed, this debate has converged into the wider discussion around the assignment of class membership, exploring whether masculine-feminine contrast at the grammatical level intertwines with the male-female contrast, exploited for gender specification.

Since the onset of linguistic research, scholars have pondered on the systems that regulate the assignment of gender. Indeed, the grammatical genders of inanimate nouns have been considered more idiosyncratic and less meaningful than the grammatical genders of animate nouns (Bloomfield, 1927; Aikhenvald, 2016), leading to the formulation of quite creative theories (Wheeler, 1899). Already in the fifth century, according to Romaine (1997), Protagoras was so convinced that 'sex' was inherent to the classification of nouns to argue that the Greek word for 'helmet' was not supposed to be assigned a feminine gender, but rather should be changed to masculine. Similarly, Grimm (1890) conceived grammatical gender as a metaphorical extension of sex. In his view, masculine nouns were stronger, bigger, whereas feminine ones were smaller, softer, and more flexible, hence explaining the masculine class membership of *Fuss* (foot), which is larger than the more 'delicate', hence feminine, *Hand* (hand) in German. Similarly, faulty claims were made by other grammarians according to Romaine (1997), which used the cross-lingual consistency of the feminine gender for family – e.g., die Familie (de), la familia (es), la famiglia (it) – as evidence of the relationship between genders and sex. Arguably, and as we will discuss below, these theories conflate the notion of sex with the sociocultural expectancies placed on this human categorization, and were largely disproved. While modern linguists have been quick to counter these post hoc rationalizations, scholars have often taken it to the other extreme by considering gender assignment to be completely idiosyncratic (Bloomfield, 1927) or arbitrary (Maratsos, 2013). To date, the extensive typological work carried out by Corbett (1991) represents the main source on grammatical gender systems around the world, which has systematized two main – although often interrelated – principles governing assignment: semantic and formal criteria.

In languages that strictly rely on semantic assignment, the meaning of a word suffices to determine its gender. That is the case of Ojibwe, from the Algonquian language family, which differentiates between animate and inanimate nouns.[27] Among those systems that include the feminine-masculine gender contrast, this strict assignment type is found in Dravidian languages from southern India like Kannada (Sridhar, 1990), where nouns denoting male humans are masculine, those denoting female humans are feminine, and these genders are only extended to personified entities, such as deities, demons, and heavenly bodies in these genders. Based on a formal principle, instead, nouns are assigned to gender according to their form, such as morphological or phonological factors. In Spanish, for instance, most nouns ending in '-o' are masculine, while most nouns ending in '-a' are feminine, although there are some exceptions, e.g., drama MASC and radio FEM (Comrie, 1999). Even French, often claimed to have no system for its gender assignment (Bloomfield, 1933), has a phonological assignment system: for example, of the nouns ending in '/ɛ/, 99% are masculine, like le pain '/pɛ/ 'the bread'.[28]

While a distinction between formal and semantic assignment serves systematization purposes, the two of them mostly interplay.[29] Indeed, this is the case for Indo-European languages, where the semantic core of grammatical gender seems to emerge in the context of living entities. In those cases, there is a tendency to correlate grammatical gender with the gender of the referent, particularly for humans and occasionally for animals (Hellinger and Motschenbacher, 2015; Comrie, 2005). Looking at this portion of the lexicon, Italian examples are e.g., *toro* (bull - MASC class, and conceptually male) vs. *mucca* (cow - FEM class, conceptually female). The same thing happens for *uomo* in Italian, *homme* in French, *hombre* in Spanish (man) and *donna*,

[27] To the best of my knowledge, languages with animate-inanimate classes do subdivide human referents into bimodal classes. This represents 25% of the languages having grammatical gender distinctions (Corbett, 2013b).

[28] For details see Tucker et al. (1977).

[29] Corbett (1991) argues plausibly that there are no purely formal gender systems, while there are few systems that seem to be purely semantic.

femme, *mujer* (woman). However, the interplay between formal and semantic principles give rise to counterintuitive instances of gender assignment, especially for nouns that refer to human referents. A case in point regards the same word Twain struggled with, namely *Mädchen*. Mädchen (young woman), indeed represents a case of competing principles, where a personal noun actually is attributed to the neuter class based on formal criteria; that is, the suffix '-chen' (young, little) is a diminutive attributed to the neuter value. Another well-known case regards the French word *sentinelle* (sentry), which is FEM, although it could be used indistinctively for any referent and, at least traditionally, the role of the sentry was occupied by men. It should be recognized, though, that in the context of personal nouns this sort of mismatch is relatively rare (Corbett, 2015).

In light of the above, we can see how the formal properties and gender-based noun classes do not reflect *tout-court* the features of their referents. Having made this distinction, I inform the reader that from now on, I will restrict the use of grammatical gender in relation to the syntactic and morphological processes that govern FEM and MASC[30] differentiation for human referents.

2.2.1.2 Extra-linguistic reality: Gender and sex

Until the second half of the 20th century, it was uncommon to use the term *gender* to refer to anything but the grammatical category outlined above. Our modern use of the word stems from pivotal work in the social sciences (Beauvoir, 1949), although it is the sexologist Money (1955) to be typically regarded as the first to introduce a distinction between biological sex and gender, describing the latter as a socially and culturally determined role or identity. Simply put, sex refers to a categorization of bodies based on biological, anatomical properties, e.g., based on chromosomes, reproductive organs, which influence the development of secondary characteristics (Johnson et al., 2012; Fausto-Sterling, 2019).[31] While still complex and contested, (West and Zimmerman, 1987; McConnell-Ginet, 2013; Kiesling, 2019) instead, the concept of gender covers at least a set of conventionalized norms and roles, such as attitudes and behaviors which are usually associated with masculinity/femininity (Rubin and Greene, 1991; Garnham et al., 2002; Gabriel et al., 2008; Eckert, 2014; Brutt-Griffler and Kim, 2018), as well as a person's own internal feeling towards those roles and how/if they decide to express it (Ansara and Hegarty, 2013; Zimman, 2017). Along this line, clothing, appearances but also gendered language are included as forms of gender expression.

Gender and sex are often widely confounded terms in linguistics and psychology (Cheshire,

[30] Henceforth, to avoid confusion throughout the thesis, we only use FEM and MASC to refer to the *linguistic* categories.

[31] Note that, although sex is still often used as a shorthand for distinguishing the bimodal nature of the male-female spectrum, physical traits can vary along multiple dimensions. Consider intersex individuals, which may present XY chromosomes but a predominately female phenotype (Hughes et al., 2012).

2004; Ansara and Hegarty, 2013). As Kiesling (2019) reconstrues, gender and sex are neither dichotomous nor necessarily completely independent of each other,[32] but overlapping them can create some serious confusion. This conflation between social and biological features was noted by Unger and Crawford (1993), because "occasionally, one can find a study in which rats are described as having a 'gender', although it is hard to imagine that they have pink or blue bows on their tails as they run the maze". Arguably, this same conflation can be found in the representation of our *female turnip* as well in Figure 2.3. Proceeding from left-to-right, we see a turnip wearing a pink collar, what seems to be a pair of earrings, and makeup. Indeed, such a characterization has nothing to do with sex-correlated traits. Rather, based on what Kramarae and Treichler (1985) call *social gender*, i.e., "the socially imposed dichotomy of masculine and roles and character traits", we recognize these as objects and appearances associated with femininity within given cultures. Along this line, the quite creative extension of metaphorical properties (e.g., delicate, strong) to inanimate objects (e.g., hand, foot) that linguists have employed to explain the function of grammatical gender (Grimm, 1890) are just cultural rather than biological projections.

Towards the understanding of bias, as well as towards a view of gender more compatible with trans-inclusive frameworks and performative theories (Butler, 1990), this works decouples biological sex and gender. Such a distinction is necessary to properly frame how gender is employed as a variable for the study of gender in NLP, and more broadly in language (Larson, 2017). Also, as we will see (more details in Sec. 2.2.3), embodied sex characteristics are only one among several factors that control gendered linguistic expression or morphology in our discursive practices. At this point, having separated the purely technical sense of gender in descriptive linguistics and its socio-cultural dimension, we are ready to look into the actual points of contact between gender and language. In particular, those that I deem more relevant for addressing gender bias in automatic translation. Towards this goal, since automatic translation operates within multilingual settings, below we first and foremost start by outlining how languages can linguistically express gender in different ways.

2.2.2 Linguistic Encoding of Gender

In this section, we thus intended to describe all the linguistic forms (*lexical*, *pronominal*, *grammatical*) that do bear a relation with the extra-linguistic reality of gender. To do so, we largely draw on (Corbett, 1991; Comrie, 1999; Hellinger and Bußman, 2001, 2002, 2003; Corbett, 2013a; Gygax et al., 2019). Also, since it is functional to the study of gender translation, we look at the qualitative and quantitative distribution of such forms across languages. Accordingly, we

[32] Ackerman (2019), for instance, introduces the notion of biosocial gender.

follow the classification by Stahlberg et al. (2007), which identifies three main language groups:

Genderless languages. Exemplary for this first group are languages such as Finnish, Turkish, or Hungarian, where the gender-specific repertoire is at its minimum. In fact, gender distinctions are only expressed at the level of basic lexical pairs, usually kinship or address terms (e.g., in Finnish *sisko*/sister vs. *veli*/brother). For such lexical items, the semantic property of femaleness or maleness is inherent to the word itself.

Notional gender languages. The most prominent representative of this group is the English language, although the group includes Scandinavian languages such as Danish and Norwegian.[33] On top of lexical gender (*mom*/*dad*), such languages display a system of pronominal gender (*she/he*, *her/him*). Note that these languages have long been described as having 'natural gender', a term used to express the overlap of gender and animacy in the pronominal forms used to address human referents. Following McConnell-Ginet (2013), we prefer notional to avoid terminological overlapping with "natural", i.e., biological/anatomical sexual categories.[34] English also displays some residual gendered morphology. In fact, it hosts some marked derivative nouns (*actor*/*actress*), and enables gender-specific compounds (*chairman*/*chairwoman*) via the combination of lexically gendered terms.

Grammatical gender languages. As anticipated in the previous section, in these languages (e.g., Arabic, Spanish, Italian), each noun pertains to a class such as masculine, feminine, and neuter (if present). Leaving aside inanimate objects where gender assignment is only formal, I discussed how in the case of personal nouns FEM and MASC genders are assigned on a semantic basis, and displayed via morphological markings. Note that, in the case of human referents, this distinction can be productive through derivation patters and allow for a paradigmatic FEM and MASC option on each word, e.g., it: amico/a (friend). Since grammatical gender is defined by a system of morphosyntactic agreement, several parts of speech beside the noun (e.g., verbs, determiners, adjectives) carry gender inflections.

In light of the above, the English sentence "*He/She* is a good friend" has no overt expression of gender in a genderless language like Turkish ("*O* iyi bir arkadaş"), whereas Spanish spreads several masculine or feminine markings ("*El/la* es *un/a buen/a amigo/a*"). Although general, such macro-categories allow us to highlight typological differences across languages. These are crucial to frame gender issues in both human and machine translation. Also, they exhibit to

[33] For a different classification, see the work by Gygax et al. (2019), which recognizes a combination of grammatical gender (common vs. neuter for nouns) and natural gender (in the pronominal system) for Scandinavian languages. As described in 2.2.1.2, we hereby only deem as relevant FEM and MASC gender marking. Note that English used to have a now lost grammatical gender system (Baron, 1971).

[34] For a wider discussion on the topic, see Nevalainen and Raumolin-Brunberg (1993); Curzan (2003).

what extent speakers of each group are led to think and communicate via binary distinctions,[35] as well as underline the relative complexity in carving out a space for lexical innovations which encode non-binary gender (Hord, 2016; Conrod, 2020). In this sense, while English is bringing the singular *they* into common use and developing neo-pronouns (Bradley et al., 2019), for grammatical gender languages like Spanish neutrality requires the development of neo-morphemes (“*Elle* es *une buene amigue”).*

2.2.3 Social Gender Connotations

To understand gender bias, we have to grasp not only the structure of different languages, but also how linguistic expressions are connoted, deployed, and perceived (Hellinger and Motschenbacher, 2015).

To start, consider the tendency of feminine words to acquire a derogatory connotation. This so-called phenomenon of ‘semantic derogation’ (Schulz, 1975) is of particular interest for grammatical gender languages, e.g., in French, *couturier* (fashion designer) vs. *couturière* (seamstress); in German *Sekretär* (secretary of an administration) vs. *Sekretärin* (secretary in an office). When it comes to derivational patterns, English forms are no exception (e.g., *governor/governess*). Such a dynamic has been linked to the relatively low-productivity and difficult acceptance of feminine forms, making them seem unsuitable for new social titles (Frank, 1985; Gheno, 2019).[36] Besides the lower perceived prestige of feminine marking, it has been pointed out that also at the lexical level supposedly “semantically symmetrical" pairings are often not symmetrical (Curzan, 2003), e.g., *spinster* carries negative connotations that are foreign to *bachelor*. Indeed, such cases have been targeted by the feminist language critique, which identified them as exemplary of unequal gender representation in language (Stahlberg et al., 2007).

Moreover, asymmetrical associations can lurk underneath seemingly neutral forms. Such is the case of epicene (i.e., gender invariant) nouns where gender is not grammatically marked. Here, gender assignment is often linked to (typically binary) social gender (Sec. 2.2.1.2). As an illustration, Danish speakers tend to pronominalize *dommer* (judge) with *han* (he) when referring to the whole occupational category (Gomard, 1995; Nissen, 2002). The tendency to attribute certain occupational nouns with masculine or feminine properties – although they do not exhibit any gender-specific formal features – underpins the confusion of the “riddle” cited

[35] Outside the Western paradigm, there are cultures whose languages traditionally encode gender outside the binary (Epple, 1998; Murray, 2003; Hall and O’Donovan, 2014).

[36] For two contemporary circumstances from the news that are arguably a by-product of semantic derogation, consider an orchestra director rejecting the feminine form of the profession (Repubblica, 2021), as well as the newly elected Italian prime minister Giorgia Meloni asking to be referred to as “il presidente” MASC (Post, 2022).

in Reynolds et al. (2006), originally from (Sanford, 1985).

> *A man and his son were away for a trip. They were driving along the highway when they had a terrible accident. The man was killed outright but the son was alive, although badly injured. The son was rushed to the hospital and was to have an emergency operation. On entering the operating theater, the surgeon looked at the boy, and said, "I can't do this operation. This boy is my son." How can this be?*

The enduring nature of this riddle is reflected in the difficulty of interpreting the surgeon as the son's mother.[37] As Ackerman (2019) underscores, this riddle demonstrates a strong male bias associated with surgeons. Despite there being no real nor linguistic requirement for that to be the case, there is a conceptual expectation that is hard to override. While certain associations seem to hold consistently and for large swathes of the population, in principle social gender assignment varies across time and space (Lyons, 1977; Romaine, 1999; Cameron, 2003). One example is the occupational title of secretary, which used to have male connotations in the 19th century. As women entered the UK labor market, the gender-specific *lady typist* term began to circulate (Lyons, 1977). In light of the foregoing, we can see how culturally-bound social gender assumptions impact our perceptions, to the point of requiring the introduction of new terms to override them (Hamilton, 1988; Gygax et al., 2008; Kreiner et al., 2008). Additionally, they can influence our behavior – e.g., leading individuals to identify with and fulfill stereotypical expectations (Wolter and Hannover, 2016; Sczesny et al., 2018) – and verbal communication, e.g., women are disproportionately misquoted in academic papers that attribute authors' gender (Krawczyk, 2017).

In translation studies, the problem of gender translation is well-documented (Jakobson, 1959; Chamberlain, 1988; Comrie, 1999; Di Sabato and Perri, 2020). Primarily, the problem arises from typological differences across languages and their gender systems. Nonetheless, socio-cultural factors deeply influence how translators deal with such differences. Consider the – admittedly secondary – character of the cook in Daphne du Maurier's *Rebecca*. In the whole English book, the chef's gender is never explicitly stated. In the lack of any available information, translators into five different grammatical gender languages independently represented the character as either a man or a woman (Wandruszka, 1969; Nissen, 2002). The decision to opt for either masculine or feminine inflections of the occupational noun can be ascribed to the expectations of translators as well as their respective communities of readers. Although extreme, this case can illustrate the situation of uncertainty faced by MT: the mapping of one-

[37] In principle, this is a heteronormative interpretation of the riddle. The riddle can also be solved by identifying two fathers as parental figures.

to-many forms when predicting gender across languages. However, as I will discuss in Chapter 3, mistranslations also occur when contextual gender information is available.

2.2.4 Gender and Language Use

From a slightly different perspective, here we discuss studies on gendered differences in language use. As previously mentioned in Section 2.1.4, language use varies across demographic groups and reflects their backgrounds, personalities, and social identities (Labov, 1972; Trudgill, 2000; Pennebaker and Stone, 2003). In this light, the study of gender and language variation has received much attention in socio- and corpus linguistics (Holmes and Meyerhoff, 2003; Eckert and McConnell-Ginet, 2013). It was the investigation by Lakoff (1972) to first explore the different communicative practices and attitudes that men and women seemed to exhibit, as if they spoke a different 'genderlect'.

Over the years, research conducted in speech[38] and text analysis highlighted several gender differences, which are exhibited at the phonetic and lexico-syntactic level. For example, women rely more on hedging strategies ("it seems that"), purpose clauses ("in order to"), first-person pronouns, and prosodic exclamations (Mulac et al., 2001; Mondorf, 2002; Brownlow et al., 2003). Although some correspondences between gender and linguistic features hold across cultures and languages (Smith, 2003; Johannsen et al., 2015), it should be kept in mind that they are far from universal and should not be intended in a stereotyped and oversimplified manner (Bergvall et al., 1996; Nguyen et al., 2016; Koolen and van Cranenburgh, 2017). It has been largely debated whether gender-related differences are inherently biological or cultural and social products (Mulac et al., 2001). Currently, the idea that they depend on biological reasons is rejected (Hyde, 2005) in favor of a socio-cultural or performative perspective (Butler, 1990).

Such purported differences in language use across gender groups have motivated NLP studies to incorporate them as a gender-based variable. Indeed, prior work has drawn on them to build demographically informed NLP tools (Garimella et al., 2019) and "gender personalized" MT models (Mirkin et al., 2015; Bawden et al., 2016; Rabinovich et al., 2017) achieving good results. Also, as we will see in Chapter 3, a few studies have discussed gender bias in MT under this lens. However, using personal gender – rather than linguistic expressions – as a variable requires a prior understanding of which categories may be salient, and a critical reflection on how gender is intended and ascribed (Larson, 2017). Otherwise, if we assume that the only relevant categories are "male" and "female", our analyses will be incapable of overcoming this assumption (Bamman et al., 2014), and opaque models will inevitably fulfill such a reductionist expectation. With this in mind, having now covered the relevant aspects of the relation between

[38] More on this point in Chapter 5.

languages and gender, I can close this foundational chapter by focusing on Neural Machine Translation.

2.3 Neural Machine Translation

> *Of all of the hopes and dreams that have been bestowed upon technology, one of the deepest and most persistent is the goal of automated translation.*
>
> **— (Sproat, 2010)**

If – as in the myth of the Tower of Babel – multilingualism is a curse, then translation might be the antidote. There are currently more than 7000 living languages spoken around the globe.[39] Crucially, the possibility to communicate depends on the ability to move and adapt information across these languages. Indeed, fantasies of technological wonders that could grant such a possibility have long inhabited – if not the collective imagery – at least the sci-fi world. *Star Trek* fans will remember the "universal translator", a small, hand-held device needed to communicate with aliens. The later *The Hitchhiker's Guide to the Galaxy*, instead, featured the yellow, leech-like Babel fish, an evolutionary marvel that could mediate speech into the first language of the characters: they just had to place it into their ear. Today, automated translation is literally a click away from reality.

By March 2021, the Google Translate app alone had been installed a billion times (Pitman, 2018). It is estimated to translate around 100 billion words per day, roughly the equivalent of 128,000 Bibles (Brooks, 2016). A survey over small and medium enterprises published by the European Commission in 2020 revealed that 70% of respondents relied on MT.[40] Due to the need to overcome language barriers, the interest and demand for translation keeps growing apace in many sectors, and MT applications are now deployed for a range of use-cases by millions of people on a daily basis, e.g., for leisure, traveling, to gist foreign content online, or for work-related activities (Vieira et al., 2022). Besides more traditional text-to-text MT, recent years have seen the surge of multimodal solutions, like image-guided models that process both image and text to e.g., to translate image captions. Also, applications for speech-to-text MT – or Speech Translation (ST) – are diverse and on the rise, including travel assistants, subtitling or simultaneous translation (Sperber and Paulik, 2020; Karakanta et al., 2021).

As I have anticipated in Sec. 2.1.1, the current popularity and mainstream use of MT is also to be accounted for by the advent of a recently novel paradigm: Neural Machine Translation (NMT).

[39] Data from the Ethnologue: `https://www.ethnologue.com/guides/how-many-languages`.

[40] Source: `https://ec.europa.eu/education/knowledge-centre-interpretation/news/human-or-machine-translation-survey-reveals-eu-sme-preferences-use-case_en`.

In the last decade, NMT made its way in the field and – while undergoing some refinements (Sutskever et al., 2014; Bahdanau et al., 2015; Sennrich et al., 2016; Vaswani et al., 2017) – quickly established itself as the dominant approach. Achieving and surpassing the results of previous data-driven statistical techniques (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017), neural models became the unchallenged state-of-the-art in MT.

While its current capabilities can be remarkable, automated translation is no wonder or marvel, and neither, as we will see, it is flawless (Arthur et al., 2016; Koehn and Knowles, 2017; Raunak et al., 2021). If we break it down to its core, machine translation is the automatic task of rendering content from one source language into a target language. Ideally, the result should be a good quality translation. Indeed, from an analogy with the intellectual activity performed by human translators, translation is a complex process. As a bare minimum, it requires high level cognitive and linguistic capabilities. The translator must be at ease with the two languages involved in the process, and transfer the source one to new target wordings, structure, and context. How do we make a computer emulate such a process and skills? And how do we verify its quality?

Leaving out of the scope of this section prior rule-based and statistical approaches to MT,[41] in the upcoming parts I describe the current NMT paradigm. To do so, in Section 2.3.1, I first introduce the neural nets underlying current deep learning approaches to MT, and the type of linguistic representations they handle. Then, I transition to the description of the MT and its architecture. Section 2.3.2 zooms on specific training and pre-processing aspects, explaining their relevance to this thesis. Before moving to the evaluation in Section 2.3.4, I describe ST models in Section 2.3.3, on which the main experimental focus of this work lies.

2.3.1 NMT Models: Looking under the Hood

As already mentioned, with the advent of deep learning based on ***artificial Neural Networks*** (NN), NMT has become the dominant paradigm in machine translation. In contrast to previous statistical approaches (Koehn, 2009), which relied on the concatenation of independently learned components, NMT functions in a so-called ***end-to-end*** fashion. That is, by training a single composite neural network that models the entire translation process. NN-based models are not provided with any explicit linguistic information on how to translate. Rather, they are designed to autonomously map patterns and features between the two languages involved in the translation process directly from the data used to train them. In the case of text-to-text MT, that would typically be up to millions of ***parallel sentences***: written sentences in the source language aligned with their corresponding translation in the target language, e.g., src-en: *I am*

[41] For the interested reader, see Hutchins (1995); Stein (2018).

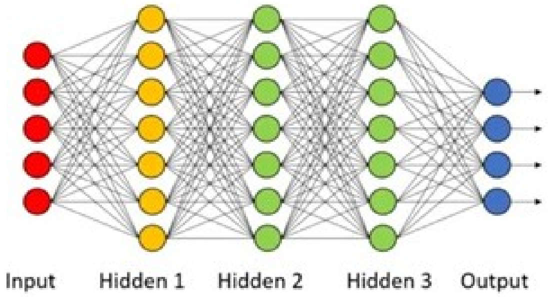


Figure 2.4: Simplified representation of an artificial neural network with three hidden layers. Each connection between neurons is represented by the lines across layers.[42]

a person; tgt-it: *Sono una persona*. These parallel data are the most basic element for NMT. Note, however, that NNs do not treat strings of language as is. Rather, as I will more thoroughly explain in the upcoming sections, such sentences are assigned a numerical ***representation***, or ***embeddings*** in the form of vectors.

The NMT task is typically framed as a ***sequence-to-sequence*** (seq2seq) learning problem (Sutskever et al., 2014). Namely, a seq2seq model takes in as input a sequence, based on which it generates another sequence as output. As an analogy, imagine having to translate into Italian the sentence “I am going home” as a two-step process. This source English sentence is the input of NMT. Ideally, the Italian “Sto andando a casa” would be the sequence output. Crucially, seq2seq models allow the input and output sequences to be of different lengths. This is a key property for translation, where the source and target sequences might not necessarily match in length. Now, let us follow the two-step analogy to outline the typical NMT ***encoder-decoder*** architecture. First, the encoder receives the source sentence as input and turns it into a representation. Then, the decoder, decodes such a representation into the target translation. This is done by unraveling the representation step by step to ***predict*** the target words. The encoder and the decoder are two NNs, which are however interconnected: hence, end-to-end NMT.

Hereby, I presented a panoramic view of NMT by also introducing its main features and terminology. In what follows, we give a more detailed look into how NMT works, its functioning, and components. Leaving aside the mathematical complexity, I move from the most essential basic NMT elements to finally describe its state-of-the-art ***attention***-based ***Transformer*** architecture.

2.3.1.1 NMT uses neural networks

To familiarize ourselves with and make sense of NMT, we first need to consider the neural networks (Goodfellow et al., 2016) that perform it. What is their composition? How do they work, and how are they trained? The main building blocks of NN are the *neurons*. Neurons

[42] Image retrieved from `https://omniscien.com/faq/what-is-neural-machine-translation/`.

work as a fully interconnected series of nodes structured into layers: an input layer, an output layer, and several hidden layers in-between. Put together, such an organized connection of nodes constitutes a neural network (see Figure 2.4), a structure that is loosely modelled on the human brain. I underscore that, despite their name and biologically-inspired structure, such artificial units only have a vague resemblance to actual neurons in the brain as well as to how the brain actually works (Baria and Cross, 2021). The analogy, however, also has to do with the resemblance of how these artificial neurons are activated (i.e., the degree to which they are excited or inhibited) depending on the stimuli they receive from their interconnected neurons, and the strength of the connection along which these stimuli are carried (Forcada, 2017).

Think about how our tongue interacts with the brain and – based on a stimulus, an input (e.g., chili) – the brain receives a signal (e.g., too spicy) and outputs a response (e.g., the muscles of the mouth move to spit the chili). For a NN, the stimulus are numeric values (e.g., text representations) fed to the nodes in the input layer. The core idea is that, to generate a proper response (output), 'the right' connections, or path forward, need to be established by activating neurons and pass the signal in the NN until the output layer. This involves a series of mathematical operations within each node (n), where: *i)* the input values received from the nodes interconnected with n are first multiplied by their corresponding weights (values that represent the strength of the connection), *ii)* these input-weighted products are then summed up, *iii)* and to the sum we finally add bias,[43] (value that represents the tendency of the neuron n to be excited). At this point, we have a new value, but are yet to determine the activation of the neuron. This is when the *activation function* comes into play. Essentially, such last operation serves to determine whether a connection with the next NN layer is weak, or rather sufficiently strong to be established and pass that value to the next neurons. This is the main idea, once that repeats multiple times in the NN: it is an incremental process, where the output of one node becomes the input of the next one. With this process in mind, however, it still remains unspecified how the values for weights and bias are assigned (see Sec. 2.3.1.2 for the input representations). That is because they are initially assigned randomly,[44] and understanding how they are subsequently updated brings us to discussing the training process of NNs.

When building a NN to solve a specific task, the first step is that of determining its architecture: how many layers and neurons it comprises and how they are connected. Once the architecture is fixed, the NN is to be trained. Training a neural net is basically the process of adjusting the NN connections given a set of training data as examples of how to perform a task. As mentioned above, NNs are initialized with random values. During the training process, however, these are progressively calibrated over and over for the whole network through a series of iterations over

[43] Here, *bias* is statistically intended as a constant value added to the NN layers to regulate the learning process.
[44] That is typically the case. Knowledge transfer techniques discussed in 2.3.3 imply an exception.

the given data. It is a process of trials and errors, where – for the specific NMT scenario – through a series of attempts the difference between the translation output of a model and the human translation (i.e., reference translation) available in the training data is minimized. To do so, the output *error* must be computed as the distance between the automatically generated and the reference translation. On this basis, the *gradient descent* – a training optimization algorithm – computes how much the weights should be changed to reduce the error and produce the correct output. In this process, the gradient of the output layer is computed, then the gradient is propagated backward through the network from the output layer to the hidden layers (i.e., *backpropagation*), assigning blame for the error and updating weights as they go. Once the error is minimized as much as possible, the neural model is considered ready, and the training completed (see also Sec. 2.3.2). Note that all of this is carried out via training algorithms – whose more detailed formalization exceeds the scope of this section – that adjust the net based on the provided training data. Hence the data-driven nature of NMT, and the depiction of neural models as systems that "learn by themselves". A main bottleneck of NMT – and NN in general – however, is precisely the availability of large amounts of data (Sennrich and Zhang, 2019; Ranathunga et al., 2021).

2.3.1.2 NMT relies on distributed representations

NMT – as with all neural computation – (Pérez-Ortiz et al., 2022) owes its power also to the processing of information based on distributed representations, i.e., the conversion of an input into numerical features that have comparative value (Ferrone and Zanzotto, 2020). In NLP applications such as MT, the information is made up of words, which are mapped to real-valued vectors intended to capture words' 'meaning'. Usually referred to as (word) embeddings, such representations are multidimensional and encode word properties in a relational way: namely, words which are closer in the vector space are expected to have similar meanings (Mikolov et al., 2013).[45] For instance, in Figure 2.5 we see how the word vectors – *v(London)* and *v(Paris)* – are close in space, both being capitals. Embeddings can also capture more subtle regularities, e.g., the 'number' property, where the distance between *v(King)* and *v(Kings)* is expected to be roughly equal to *v(Queen)* and *v(Queens)*. Also, the space between *v(King)*, *v(Queen)* and *v(Uncle)*, *v(Aunt)* is similar, thus capturing the gender relation.[46] Indeed, the same principle also applies to mapping vector similarities across languages, e.g., *v(Queen)*, *v(Regina)* in Italian. Note that the embeddings are fixed-length, regardless of the number of characters in the string of text they represent. Fixed-length embeddings are in fact computed for variable-length sequences.

[45] This approach owes much to distributional semantics theories (Boleda, 2020), which lie on the assumption that the meaning of a word can be inferred from its usage (i.e., its distribution in text), and that words occurring in the same contexts tend to purport similar meanings (Firth, 1957).

[46] In Chapter 3, I discuss how these properties underlie some mitigating strategies for gender bias.

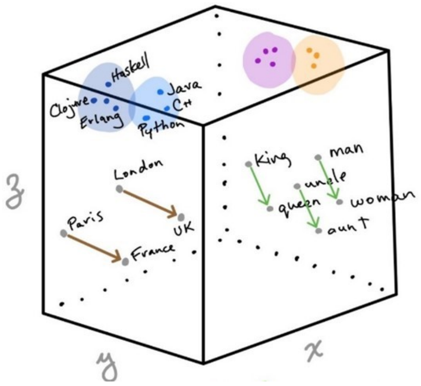


Figure 2.5: Words represented as vectors positioned in a multidimensional space, which capture similarities and relations across words.[47]

Zooming back to NMT, embeddings and weights are both learned at training time through repeated guesses based on the given data. However, embeddings are not restricted to the word level (or sub-word level, see Sec. 2.3.2). Rather, these vectors are built up together to generate sentence embeddings (Kalchbrenner and Blunsom, 2013). This is a key advantage for NMT, which can generate a translation beyond word-to-word mappings, but rather by bearing in mind the whole sentence context. It is however quite clear that, by dealing with numerical (sub-symbolic) representations instead of explicit words, the NMT translation process becomes inherently opaque to linguistic interpretation (Lipton, 2018; Costa-jussà, 2018).

2.3.1.3 Seq2seq learning: Attention mechanism and Transformer architecture

Knowing how NNs work, how they are trained, and the type of representation they deal with, we can now look at how NMT performs a translation task and introduce its encoder-decoder architecture. Then, I outline the largely employed Transformer architecture.

As introduced in Sec. 2.3.1, the typical NMT model has a seq2seq design that consists of two main interconnected components (encoder, decoder) (Sutskever et al., 2014; Cho et al., 2014) normally composed by several stacks of NN layers, hence *deep architectures*. The encoder receives the input textual sequence and transforms it into a fixed-length sentence representation formed from the vector embedding of individual words. Such a representation – in the form of an embedding matrix – can be seen as a lookup table for the *vocabulary* that the system is supposed to handle. The embedding is passed to the decoder, which turns the representation into the target language, by producing one output word at a time until the end of the sentence. Actually, the generation of each word entails a *prediction*, where the decoder seeks – or predicts – which word among all words from the target vocabulary has the highest likelihood of being the

[47] Source Image: Jon Krohn @untapt – Safari Live Lessons. Available at: `https://blog.trifork.com/2018/01/03/deep-learning-for-natural-language-processing-part-i/`.

correct one.[48] Imagine NMT as a text completion device (Forcada, 2017), where the likelihood of a word – and its consequential generation – depends both on its source representation, and on the target words that have already been predicted (e.g., how likely is word Y a continuation for word X).

NMT can rely on different types of NNs (Goodfellow et al., 2016) – e.g., feed-forwards, convolutional, recurrent – and researchers have experimented with several of such types to build the encoder-decoder architecture (Kalchbrenner and Blunsom, 2013; Jean et al., 2015b; Luong et al., 2015a; Gehring et al., 2017; Kaiser et al., 2017). In the earliest implementation by Sutskever et al. (2014), NMT relied on recurrent neural networks (RNN). Now mostly surpassed, RNNs were largely used for the translation task because their structure allows the output of a step (each hidden state) to be conditioned on the previous state too, in a recurrent way. In practical terms, the encoder takes in both the *i*) vector of an input word, *ii*) a context vector including information on the previous words, and then updates the context vector with the information collected during this step. The context vector is passed through and updated at each step. Accordingly, each step takes information from the new input word and the full sentence history (originally processed from left-to-right).[49] However, one of the main disadvantages of recurrent encoder-decoder architecture is that the importance of previous words diminishes as the context becomes longer: as the whole sentence information is compressed in one context vector, relevant – but older – pieces of information 'drown' under more recent one. Indeed, RNNs particularly struggled with so-called long-term dependencies, i.e., translation of grammatically related words that are however distant from one another within the sentence (e.g., *Since my grandma was feeling tired yesterday, I think I'll give her a call*).

Attention. Along this line, a substantial improvement came with the introduction of the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015a), which almost immediately extended the decoder-encoder architecture. Intuitively, the attention device is a method that supports NMT by providing focus on the parts of the source sentence that are more relevant to translate a particular word. First introduced in the field of computer vision (Larochelle and Hinton, 2010), this mechanism attempts to mimic cognitive attention in humans. When we are looking for an insect in a field of grass, our focal point adjusts on the insect region, while perceiving less its surrounding. Analogously, if we take as an example the above-mentioned sentence about the *grandma*, a translation is not merely performed by remembering the list of all preceding words. Rather, the translation of *her*, or *tired* will require giving more attention to

[48] For instance, if the vocabulary contains 30,000 words, the prediction is a 30,000-dimensional vector with the probability of each word, which is computed with the *softmax function*.

[49] There are other approaches to enhance this representation – from left-to-right and right-to-left – such as bidirectional RNNs (Schuster and Paliwal, 1997). We stick here with the simplest case.

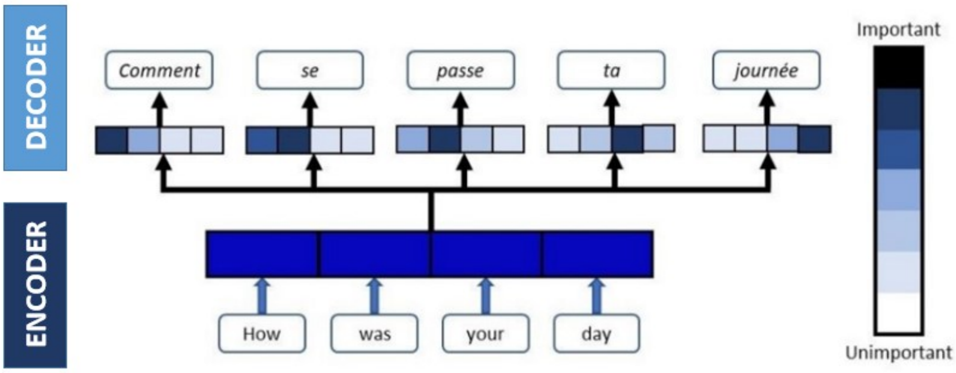


Figure 2.6: Simplified representation of the general attention mechanism between input and output. For each target word, a vector – exemplified as a colored series of squares – maps the degree to which each source word contributes to translate the target one. Note the relevance of this cross-lingual mapping beyond word-to-word translation, e.g. *How was* → *Comment se passe*.[50]

the noun they refer to. How does the attention mechanism achieve that? This device, placed in the decoder as a jointly trained feed-forward NN, generates a (context) attention-vector which weights how much each source encoded word should contribute to the generation of a target one (see Figure 2.6).

Transformer. Building upon the attention mechanism, the Transformer architecture (Vaswani et al., 2017) represents the current state-of-the-art in NMT, which thus also underlies the automatic translation systems employed in this thesis. The core aspect of the Transformer design is that it completely eliminates recurrent networks. In fact, Transformers consist of an encoder and decoder, each consisting of a stack of layers based on feed-forward NNs and *self-attention*. What are the implications of such a design? First, self-attention - as a specific type of attention mechanism - does not only work as a bridge between the encoder and decoder to inform which parts of the input sequence are more relevant to generate the target word. Self-attention, additionally, informs the relation between source representations themselves as to get more refined encoder embeddings. The same occurs in the decoder with the target sequence, resulting in an overall improvement of the output translation quality. The fact that the Transformer architecture is not recursive, instead, might seem counterintuitive given that feed-forward NNs, unlike RNNs, do not capture information about the order of words in a sequence. If we think about the difference between, e.g., *the dog bites the cat* vs. *the cat bites the dog*, this appears problematic, since linear word-order conveys meaningful information. The issue is solved by

[50] Image adapted from: https://pradeep-dhote9.medium.com/attention-mechanism-in-deep-learning-1cd0dafd0c21.

[50] Namely, we have attention among source words, among target words being generated, and between source and target words. As an additional detail, consider that Transformers use multiple *attention heads*. In practical terms, multiple heads allow generating even more than one attention embedding for each single word, so to capture as many nuanced representations as possible.

enriching word embeddings with positional encodings, serving as unique ID about the position of a word within the sentence. In this way, such NN do not need to remember how sequences are fed into a model. This, precisely, is a crucial advantage of Transformers. Since we do not need any sequential order to compute representations, we can run self-attention in parallel for all the tokens in the input sequence, and vectors can flow through the encoder and decoder stacks simultaneously. In this way, the Transformer architecture also speeds up the time needed for training and translate new data, i.e., *inference* time.

In the last few years, Transformer-based architectures have pushed the field forward, achieving unmatched results, and not only for NMT. Transformers now ground several of the latest NLP achievements, such as large language models like BERT (Devlin et al., 2019). All in all, we can see how the gold rush that followed the advent of the neural paradigm has been largely focused on architectural decisions and advances. Indeed, technical constraints have initially shaped the field more than any other concern.

2.3.2 Additional Settings

2.3.2.1 Subword units

So far, I have referred to *words* as the most basic unit of representation in NMT. Concretely, however, operating with embeddings using words as the smallest representational units within a sentence poses several challenges for NMT.

NMT typically allows to efficiently process a restricted vocabulary – i.e., number of words (embeddings) – of about 30,000-50,000 items (Sennrich et al., 2016). Translation tasks, however, entail a so-called "open-vocabulary" problem. Clearly, 50,000 items do not represent every possible word in a language. For highly inflected or agglutinative languages, in particular, the number of word forms can easily reach hundreds of thousands. Just consider the following forms for *child* in Italian: *bambino* (MASC) as well as *bambina* (FEM), but also *bimbo/a*, or *bimbetto/a*. Even by bounding the size of the vocabulary to the unique word forms (i.e., types)[51] available in the training data only, *i)* the number of such forms can nonetheless be more than 30/50,000 given the dimension of current corpora, and *ii*) a model will perform clumsily if confronted with the translation of new sentences involving words that are not contained in the training set, i.e., out-of-vocabulary words (OOV).

Among other proposals to alleviate these issues (Jean et al., 2015a; Luong et al., 2015b), a remedy was found in the task of word segmentation. Namely, to split up words into *subword*

[51] With *tokens*, we refer to all words appearing in a corpus, even repeatedly, whereas *types* refer to the unique word entry. For instance, in the sentence "We looked in all the rooms and all the corners" we have 10 tokens, but 8 types (*We, looked, in, all, the, rooms, and, corners*).

units that can reduce the vocabulary size and provide meaningful representations for the model. But how? Among a wide range of alternatives (Mielke et al., 2021), the most commonly employed segmentation strategy to date is the *byte pair encoding* (BPE) compression algorithm, introduced by Sennrich et al. (2016). Created as a statistical, language-independent strategy, BPE automatically finds how to segment words in a given training set and language: first, by splitting them into characters, and then by combining the most frequent sequences of characters into a single sub-word unit. Such a merging process proceeds until we have attained the desired vocabulary size, which can be defined by setting the number of so-called “merge rules" (e.g., 50K merging rules roughly correspond to a vocabulary of 50K subword units). The result entails subwords that are statistically significant, e.g., *looked* and *baking* split as *look* + *ed*, *bak* + *ing*. By relying on such units, models can thus *generalize* beyond the sole words encountered in the training data, e.g., to translate the new input ***looking***, by combining the subword units obtained from ‘***look*** + *ed*’ and ‘*bak* + ***ing***’. Besides improvements on OOV words (also called *unseen*), BPE was found to allow for a better treatment of *rare* words with low frequency in the training data, a common shortcoming for NMT.

Word segmentation is an essential step applied to the text fed to NMT models, including the target side of speech-to-text systems. Often regarded as a routine pre-processing technique, in Chapter 6 we approach the translation of feminine forms as a case of rare words, and inquiry the potential side effects of BPE in terms of gender bias.

2.3.2.2 Stopping criteria and development set

In Sec. 2.3.1.1, I outlined how training a NN is a trial and error process that tries to minimize an error function. Or, from a different perspective, at this point we can say that NMT training attempts to “maximize the probability of the target sentences in the training corpus” (Pérez-Ortiz et al., 2022). Deciding when the NMT training process is complete – hence, should stop – implies a delicate balance, though. On the one hand, the model should be put in the position to “learn” as much as possible from its training data. On the other, however, the model must not just end up memorizing the training data, i.e., *overfit* them. Overfitting, indeed, hurts NMT ability to generalize from the training set to translate new unseen input sentences. The recipe is training just enough: but how to identify when enough is enough? This is when the development set (also called validation set) comes into play. Compared to the training set, the development one is smaller, and held-out just for the purpose of monitoring the learning trend of the model and validate its performance. On top of computing error against the training corpus, the independent development data are used to compute the *validation loss*, which is used to progressively check the performance of the model on the validation set and inform about generalization. When the loss starts to increase, the performance on the development set degrades, hence the training

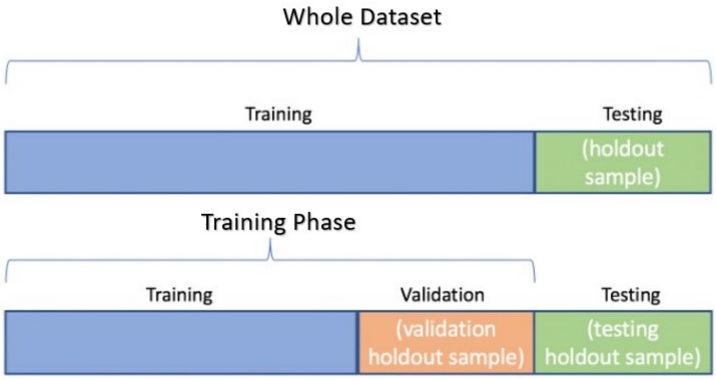


Figure 2.7: Dataset splits for training and evaluation. Training data are typically around 80% of the whole dataset, from which a 20% can be set aside for the development set. The remaining 20% is the test set (of at least 20/30k tokens) dedicated to the evaluation phase.[52]

should be stopped.

As we can see in Figure 2.7, the development set is obtained as a subset of the data employed at training time. Accordingly, the two sets usually roughly reflect the same language distribution. This means that, if the training data under-represent feminine gender, the development set will check the model's learning trend on the same type of under-representation, and indulge its replication.[53]

2.3.3 Speech Translation

Unlike the longstanding task of text-to-text MT (henceforth simply MT), ST translates spoken language from acoustic signals. The research history of ST dates back to 40 years ago. Since the first demo given in 1983 (Release, 2004), its evolution has been closely related to the adjacent fields of speech recognition and machine translation. Traditionally, ST has been in fact approached with so-called *cascade* architectures, which couple ASR and MT to address the ST task (Stentiford and Steer, 1988; Waibel et al., 1991; Eck and Hori, 2005). In the past few years, however, *direct* ST architectures (Bérard et al., 2016; Weiss et al., 2017) have gained traction, by translating the speech signal in an end-to-end fashion. A representation of both cascade and direct approaches is provided in Figure 2.8.

Note that in ST, whether direct or cascade, the speech input is pre-processed. In fact, the audio signal – stored as a waveform – is *i)* split into short-time frames (~ 25 ms) with overlapping

[52] Image adapted from: https://datascience.stackexchange.com/questions/61467/clarification-on-train-test-and-val-and-how-to-use-implement-it.

[53] In light of this, in Section 5.4.1, we create and release a new development set, intended to avoid rewarding the unbalanced gender distribution a model might be inferring from the training set portion of the dataset. As far as stopping criteria are concerned, the analysis in Chapter 8, instead, orthogonally investigates whether commonly used stopping metrics are insensitive to gender bias.

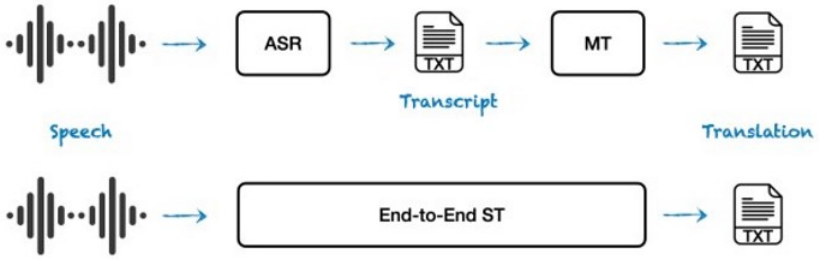


Figure 2.8: ST cascade (top) and direct (bottom) architectures.[54]

windows of 10ms between consecutive frames (∼ 10 ms), from which *ii*) a frequency spectrum is computed, and *iii*) log Mel filterbanks (i.e., frequency bands) are extracted as vector features given as input to the system (Mohamed et al., 2012). Besides this common pre-processing step, the specificity of each ST approach is described below in more detail.

Cascade Speech Translation. The traditional approach to ST entails a pipelined architecture consisting of two components. The ASR, which transcribes the source audio input, and MT, which translates the automatic transcript into text in the target language. Cascade systems are a reasonable solution, which grants the advantage of plugging state-of-the-art technology for each subcomponent, and exploit the wealth of training data available for both ASR (monolingual *audio-transcript*) and MT (*source text - target translation*). Besides the additional cost of separate maintaining procedures for each component, however, the pipelined architecture entails some intrinsic drawbacks. One is error propagation: sub-optimal ASR transcripts (e.g., *bay*→*pay*) have significant impact on the final output produced by the MT component (wrongly translated as *pagare* in Italian). To cope with this issue, recent works focused on making MT models more robust to noisy input transcripts (Sperber et al., 2017; Di Gangi et al., 2019d; Martucci et al., 2021). A further downsize, more relevant to this work, is the loss of the richer audio information that is not retained in (written) transcription, and thus cannot be grasped by the MT component (e.g., prosody, pauses, speaker's vocal traits). Their preservation has been attempted through acoustic feature vectors (Do et al., 2016; Anumanchipalli et al., 2012), or by injecting external knowledge to support the MT step (Elaraby et al., 2018).

Direct Speech Translation. By directly translating source speech into target text without intermediate representations, Direct end-to-end approaches can potentially cope with the limitations of cascade systems. Direct ST is defined as a seq2seq task, where a sequence of audio features in the source language are mapped to a sequence of written words in the target language with a composite NN. As it historically happened also for MT (Sec. 2.3.1), first ST solutions

[54] Image adapted from (Cortès Sebastià, 2020).

were built with RNNs encoder-decoder architectures (Bérard et al., 2016; Weiss et al., 2017), whereas current better-performing systems rely on ST-adaptions of the Transformer architecture (Di Gangi et al., 2019d; Bahar et al., 2020; Gaido et al., 2020a). Such adaptions keep the decoder as in the original Transformer architecture (Vaswani et al., 2017), but integrate the encoder with convolutional layers that reduce the audio input length to make it computationally manageable (Dong et al., 2018). Also, a distance penalty is added to the encoder self-attention, which explicitly brings it to attend local context and model short-range dependencies (Sperber et al., 2018; Di Gangi et al., 2019c). Over the years, direct models have achieved increasingly promising results, as shown across IWSLT[55] evaluation campaigns (Niehues et al., 2018, 2019; Potapczyk and Przybysz, 2020). However, the critical bottleneck for direct solutions is data paucity. Indeed, *audio-translation* parallel paired data are scarce compared to ASR and MT data available. To compensate, ***(i)*** data augmentation, and ***(ii)*** knowledge transfer techniques are employed.

(i) Data augmentation consists in the creation of additional *silver* data, i.e., artificial. Data can be augmented on the textual side, by automatically translating the transcript of ASR corpora with MT. Differently, the source side of textual translation corpora can be synthesized to create a corresponding automatic audio counterpart (Jia et al., 2019). Slightly different approaches to data augmentation, instead, corrupt the audio input to increase the variability of the training data. Among these, SpecAugment (Park et al., 2019; Bahar et al., 2019b) is directly applied to the audio features and corrupt spectrograms in both frequency and time dimensions. Analogously, the time stretch method (Nguyen et al., 2020) manipulates the time series of the frequency vectors (in the audio features) to achieve the effect of speed perturbation, i.e., adding noise by introducing temporal patterns with varying duration to make models more robust to temporal variations.

Finally, ***(ii)*** knowledge transfer (Gutstein et al., 2008) consists in ‘passing’ the information – or knowledge – already learned by a trained model with good performance onto another model. For the ST scenario, once again, knowledge transfer relies on the closely related ASR and MT tasks. One strategy is encoder-pretraining, where the ST encoder – rather than starting training with randomly assigned weight values – is initialized with the already optimized weights of an ASR system (Bérard et al., 2016; Bansal et al., 2019; Bahar et al., 2019a). On the other side, existing MT models are used for (word) knowledge distillation (KD) (Liu et al., 2019; Gaido et al., 2020b, 2022), where a “student” ST model is taught to produce the same output distribution as the MT “teacher”. This can be done by computing the divergence between the output distribution of an ST model against the pre-stored MT output distribution during training.

As the latest emerging paradigm, most works on direct ST compare their results against the

[55] International Conference on Spoken Language Translation: `https://iwslt.org/`.

cascade counterpart, as we also do in Chapter 5. At the time of our experiments, direct solutions still lagged behind traditional ones. As the gap seemed to be progressively reducing (Pino et al., 2019; Indurthi et al., 2020; Inaguma et al., 2020), the comprehensive study by Bentivogli et al. (2021) took stock of the situation, concluding that the overall performance of the two approaches is now *on par*. At a more fine-grained level, however, the two approaches result in different types of errors, where direct models emerge with poorer syntactic capabilities, but as more robust to audio understanding and able to take advantage of speech features to support translation. While keeping in mind overall translation quality, it is on this specific advantage of direct solutions that the focus on ST for this work lies. As we mentioned above, direct ST receives audio as an input, whose spectrogram contains information on *i)* duration, i.e., the temporal dimension of speech, and *ii*) frequency, such as prosody elements, but also *pitch*. As a feature of speakers themselves, pitch can be perceived as a strong proxy for gender. [56]

2.3.4 System Evaluation

An essential component of any translation model is its evaluation. Arguably, the development of translation technology is inseparable from its evaluation: "After all, how can progress be made if it cannot be identified as such?" (Doherty, 2019). As critical as it might be, evaluating translations is a thorny enterprise (Castilho et al., 2018). How do we define its quality? And quality for whom? For instance, mistranslating a negation (e.g., "Do not take this pill before lunch" rendered as "Do take this pill before lunch") implies a local omission that could be quickly post-edited by a translator. In terms of fidelity, however, the error is dire, especially if encountered by an end-user that cannot recognize it (Martindale and Carpuat, 2018; Martindale et al., 2021). Between the polar opposites of very good and very bad translations, there is a world of nuances that can be hard to pin down (Liu and Gildea, 2005; Giménez and Màrquez, 2010). Faulty evaluation practices compromise the soundness of experimental research on automatic translation (Freitag et al., 2020; Marie et al., 2021) and can generate unmatched expectations about its actual quality and limits (Toral et al., 2018; Läubli et al., 2018). Given the richness of the literature and the complexity of the issue, some may argue it is impossible to write a comprehensive overview of MT evaluation (Hovy et al., 2002; Lopez, 2008). Nonetheless, certain trends should be mentioned. In what follows, I touch upon the key aspects of MT evaluation relevant to this work. Accordingly, manual and automatic methods to MT evaluations and their respective metrics are hereby described.

[56] In Chapter 5, and from then onwards, I will explore the ability of direct ST systems to leverage such cue in translation, as well as its ethical implications.

Human evaluation. Human evaluation represents the "gold" approach to assessing MT quality. This is a manual method, where the output translation is judged by – ideally – bilingual evaluators, or linguists being proficient in both the source and target language. *Fluency* and *adequacy* are the two longstanding quality criteria employed to decide “how good” a translation is. Fluency, quite intuitively, questions how fluent, grammatically acceptable a translation is and can be potentially judged without considering its source. Adequacy, instead, reflects whether the translation correctly preserves the source meaning and information, an increasingly relevant aspect for NMT given its tendency to generate highly fluent, but semantically misleading, translation (Forcada, 2017). Based on these two criteria, evaluators rate the quality of the output on five or seven-point scales, comparing it to the input sentence (Callison-Burch et al., 2007).

An obvious advantage of human evaluation is that it is more accurate and informed of language nuances than automatic assessments, in spite of its higher costs and time to be performed. Manual analysis, however, can be influenced by evaluators’ own subjectivity and preferences (Lommel et al., 2014), and as research has shown, there is broad variability in human judgments of fluency and adequacy (Callison-Burch et al., 2008). Among some of the most canonical methods that have led to higher agreement, there is *relative ranking* (Bojar et al., 2018). Here, instead of rating a system on an absolute scale, evaluators pick a preferred translation between the output of two systems. Alternatively, *direct assessment* (DA) (Graham et al., 2013) entails judgments on a continuous scale (0-100). In reference-based DA, evaluators compare multiple outputs against a single human translation, so to scale the evaluation to several models by relying on monolingual raters only. Both these approaches, though, condense a judgment of quality into a single score or preference, without any specific observations over the nature of translation problems involved.

Going more fine-grained, MT output can be evaluated against error typologies (.e.g, morphology, lexicon etc.) typically hierarchically macro-organized into adequacy/fluency issues (Flanagan, 1994). One can opt from several of such available frameworks – e.g., Multidimensional Quality Metrics (Lommel et al., 2013) – or independently outline a specific taxonomy based on the specific focus of a given evaluation (see Chapter 7).

Finally, evaluations more grounded in a scenario of applicability are those based on the human task of post-editing (PE), i.e., the task of manually correcting an automatic translation according to the source. In PE-based evaluation, instead of judging the ‘goodness’ of a translation, focus is given to how much work is required to fix it. Thus, the automatic output can be compared to its post-edited version by means of metrics, such as Human translation edit rate (HTER) (Snover et al., 2006b). HTER calculates the edit-distance as the minimum number of edits (e.g., deletions, insertions, shifts, substitutions) needed for an output to match its post-edited version.

Automatic evaluation. Despite attempts at assessing text generation in a referenceless fashion (Lo, 2019; Zhang et al., 2019; Rei et al., 2020), MT automatic evaluation typically compares the model output (*hypothesis*) against collected human translations (*reference*). To compute the overlap between references and systems' output, a wide variety of metrics that correlate with human judgments have been proposed, thus allowing the evaluation process to be considerably quick and easily reproducible. Among those, TER (Translation Edit Rate) (Snover et al., 2006b), similar to HTER, calculates the edit-distance as the proportion of deletions, insertions, shifts, substitutions that would be required for an output to match the reference translation. Some metrics are designed to better deal with aspects like semantics – e.g., METEOR (Banerjee and Lavie, 2005) – or morphology, such as CharacTER (Wang et al., 2016) or chrF (Popović, 2015). The most longstanding and employed string-matching approach is BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), a precision metric based on *n*-grams, i.e., a contiguous sequence of n words. Namely, BLEU calculates precision as the proportion of the number of *n*-grams of a specific size (1-grams, 2-grams, 3-grams, 4-grams) appearing in the hypothesis, and the total number of *n*-grams of the same size generated in the reference. The higher the BLEU score, the better the translation output. One of the main shortcomings of BLEU, and reference-based metrics in general, is that it evaluates the output against a selected possible reference translation, although many possible translation alternatives could be acceptable. Even by potentially relying on multiple references as a point of comparison (Zbib et al., 2013), linguistic variability remains an intrinsic issue. Additionally, automatic scores are hard to interpret, e.g., 22% BLEU score is better than 20%, but it is not clear what it means in terms of translation quality *per se*, nor for specific linguistic aspects, since the score is aggregated and all *n*-grams matches are equally rewarded.

A viable solution for fine-grained, focused automatic evaluation relies on dedicated test sets that represent a specific linguistic phenomenon of interest, on which *ad-hoc* metrics can be computed. In the next chapter 3, I discuss how research on gender bias has confronted the evaluation challenges hereby outlined, which have informed the creation of the gender bias evaluation benchmark MuST-SHE (see Chapter 4).

2.4 Conclusion

In this chapter, I set out the contextual, theoretical, and technical foundations of this thesis. First, in Section 2.1, I presented the growing research area of ethics in NLP, which is concerned with the social impact of increasingly ubiquitous language technologies. Towards this goal, I discussed the underlying assumptions of current data-driven approaches to NLP and outlined a path forward for investigating bias in NLP by foregrounding and benefiting from several subfields of linguistics. Accordingly, in Section 2.2, I examined the relationship between gender and language. This has required clarifying the use of gender as a variable, by inspecting its relation to linguistic, social and biological categories. To understand gender translation, I outlined how gendered linguistic forms are differently distributed across languages, and how their deployment affects – and reflects – the perception and representation of gendered groups. Finally, in Section 2.3, I laid out the necessary technical background for following up with the upcoming chapters. I have described the current paradigm of NMT – its components, settings, and architecture – and the state of the art in ST. Last but not least, I explained the key methodologies and metrics for evaluating automatic translation.

With the key notions, concepts, and terminology in hand, the reader is now ready to delve into the main part of this thesis, which focuses on gender bias in automatic translation. In the next chapter, I begin by taking stock of related work on gender bias in MT.

3

Taking Stock of Gender Bias in Machine Translation

Machine Translation technology has facilitated our daily tasks by providing accessible shortcuts for gathering and communicating information. However, it can suffer from biases that can harm users and society at large. As a relatively new and rapidly growing field of inquiry, studies of gender bias in MT lack cohesion. This advocates for a unified framework to ease and guide research on the topic. Thus, in this chapter I critically: *i)* review current work on bias and its conceptualization in MT in light of insights from related disciplines, *ii)* discuss the state of the research by identifying blind spots and challenges. The offered comprehensive review represents the first contribution of the thesis towards understanding, assessing, and mitigating gender bias.

3.1 Introduction

Interest in understanding, assessing, and mitigating gender bias is steadily growing within the NLP community, with several studies showing how gender disparities affect language technologies (Rudinger et al., 2018; Hendricks et al., 2018). Despite a prior disregard for such phenomena within research agendas (Cislak et al., 2018), in Section 2.1 I underscored how it is now widely recognized that NLP tools encode and reflect controversial social asymmetries for many seemingly neutral tasks, MT included. Admittedly, the problem is not completely new (Frank et al., 2004). A few years ago, Schiebinger (2014) criticized the phenomenon of “masculine default" in MT after running one of her interviews through a commercial translation system. Despite several feminine mentions in the source text, she was repeatedly referred to by masculine pronouns. Gender-related concerns have also been voiced by online MT users, who noticed how commercial systems entrench social gender expectations, e.g., translating engineers as MASC and nurses as FEM (Olson, 2018).

With language technologies entering widespread use and being deployed at a massive scale, concerns towards their societal impact have triggered several prompt, yet disparate, responses. To take stock of the situation, Sun et al. (2019a) reviewed NLP studies on the topic. However, their survey is based on monolingual applications, whose underlying assumptions and solutions may not be directly applicable to languages other than English (Zhou et al., 2019; Zhao et al., 2020; Takeshita et al., 2020) and cross-lingual settings. Moreover, MT is a multifaceted task, which requires resolving multiple gender-related subtasks at the same time (e.g., coreference resolution, named entity recognition).[1] Hence, depending on the languages involved, and the factors accounted for, gender bias has been conceptualized differently across studies. Also, since their onset, studies in MT have tackled gender bias by means of a relatively narrow, problem-solving oriented approach to the detriment of deeper analyses on the nature of such a multifaceted issue. Indeed, while technical countermeasures are needed, failing to adopt a wider perspective and engage with related literature outside of NLP can be detrimental to the advancement of the field (Blodgett et al., 2020).

Having at hand relevant notions between gender and different languages (see Sec. 2.2), I hereby intend to put such literature to use for the study of gender bias in MT. I go beyond surveys restricted to monolingual NLP (Sun et al., 2019a) or more limited in scope (Costa-jussà, 2019; Monti, 2020), and present a comprehensive review of gender bias in MT. In particular, I first outline a unified framework that introduces the concepts, sources, and effects of bias in MT. Accordingly, in Section 3.2 I start by clarifying the concept of bias itself, and then inquiry the factors concurring to its emergence in MT systems in Section 3.3. Then, Sections 3.4 and 3.5, respectively, discuss previous analyses aimed at assessing gender bias in MT, and the mitigating strategies proposed so far. While drawing this mapping for gender bias in MT, I set out the line of inquiry for the following chapters of this work.

[1] Named entity recognition (NER) is the task of identifying entities in a text, e.g, person names.

3.2 Bias Statement

Bias is a fraught term with partially overlapping, or even competing, definitions. Indeed, "bias" is colloquially, or even legally, used in a normative sense to indicate a partial judgment, based on prejudices or preconceived notions. Such a connotation, however, is traditionally foreign to the statistical, neutral notion of "bias", e.g., an estimation error[2] (Campolo et al., 2017). The probably most longstanding line of research on bias stems from the cognitive sciences, where bias refers to the possible outcome of heuristics, i.e., mental shortcuts that can be critical for prompt reactions (Tversky and Kahneman, 1973, 1974). AI research borrowed from such a tradition (Rich and Gureckis, 2019; Rahwan et al., 2019) and conceived bias as the divergence from an ideal or expected value (Glymour and Herington, 2019; Shah et al., 2020), which can occur if models rely on spurious cues and unintended shortcut strategies to predict outputs (Schuster et al., 2019; McCoy et al., 2019; Geirhos et al., 2020). Since this can lead to systematic errors and/or adverse social effects, bias investigation is not only a scientific and technical endeavor but also an ethical one, not to be reduced to a statistical sense.

As Blodgett et al. (2020) recently called out, and has been endorsed within the NLP community (Hardmeier et al., 2021), analyzing bias is an inherently normative process that requires identifying *what* is deemed as harmful behavior, *how*, and to *whom*. Following these recommendations, we hereby stress a human-centered, sociolinguistically-motivated framing of bias. By drawing on the definition by Friedman and Nissenbaum (1996), we consider as biased an MT model that *systematically* and *unfairly* discriminates against certain individuals or groups in favor of others. We identify bias per specific model's behaviors, which are assessed by envisaging their potential risks when the model is deployed (Bender et al., 2021) and the harms that could ensue (Crawford, 2017), with people in focus (Bender, 2019b).

Since MT systems are daily employed by millions of individuals, they could impact a wide array of people in different ways. As a guide, we rely on Crawford (2017), who defines two main categories of harms produced by a biased system: *i)* **Representational** harms **(R)**, implying the detraction from the representation of social groups and their identity, which, in turn, affects attitudes and beliefs; *ii)* **Allocational** harms **(A)**, occurring when a system allocates or withholds opportunities or resources to certain groups. Considering the so far reported real-world instances of gender bias (Schiebinger, 2014; Olson, 2018) and those addressed in the MT literature that will be reviewed below, **(R)** can be further distinguished into *under-representation* and *stereotyping*.

Under-representation refers to the reduction of the visibility of certain social groups through language by *i)* producing a disproportionately low representation of women (e.g., most feminine

[2] For instance, the one we encountered in Section 2.3.1.

entities in a text are misrepresented as male in translation); or *ii)* not recognizing non-binary individuals (e.g., when a system does not account for gender-neutral forms). Under-representation can thus occur when translating text with mentions "about" such groups, but also impact the rendering of language "by".[3] speakers of such groups when *iii)* failing to reflect their identity and communicative repertoires (e.g., MT misgenders the speaker, or mishandles typical expressions and stylistic choices of the speaker's social groups).

Stereotyping regards the propagation of negative generalizations of a social group, e.g., belittling feminine representation to less prestigious occupations (teacher FEM vs. lecturer MASC), or in association with attractiveness judgments (pretty lecturer FEM).

Such behaviors are harmful as they can directly affect the self-esteem of members of the target group (Bourguignon et al., 2015). Additionally, they can propagate to indirect stakeholders. For instance, if a system fosters the visibility of the way of speaking of the dominant group, MT users can presume that such a language represents the most appropriate or prestigious variant – at the expense of the communicative repertoires of other groups. These harms can aggregate, and the ubiquitous embedding of MT in web applications provides us with paradigmatic examples of how the two types of **(R)** can interplay. For example, if women or non-binary[4] scientists are the subjects of a query, automatically translated pages run the risk of referring to them via masculine-inflected job qualifications. Such misrepresentations can lead to experience feelings of identity invalidation (Zimman et al., 2017). Also, users may not be aware of being exposed to MT mistakes due to the deceptively fluent output of a system (Martindale and Carpuat, 2018). In the long run, stereotypical assumptions and prejudices (e.g., only men are qualified for high-level positions) will be reinforced (Levesque, 2011; Régner et al., 2019).

Regarding **(A)**, MT services are consumed by the public and can thus be regarded as resources in their own right. Hence, **(R)** can directly imply **(A)** as a performance disparity across users in the *quality of service*, i.e., the overall efficiency of the service. Accordingly, a woman attempting to translate her biography by relying on an MT system requires additional energy and time to revise wrong masculine references. If such disparities are not accounted for, the MT field runs the risk of producing systems that prevent certain groups from fully benefiting from such technological resources.

In the following Sections 3.4 and 3.5, the above-described categories will be used to map studies on gender bias to their motivations and societal implications.

[3] See also the classifications by Dinan et al. (2020).

[4] We use non-binary as an umbrella term for referring to all gender identities between or outside the masculine/feminine binary categories.

3.3 Understanding Bias

In the following, we overview how several factors may contribute to gender bias in MT. We identify and clarify concurring problematic causes, accounting for the context in which systems are developed and used. To this aim, we rely on the three overarching categories of bias described by Friedman and Nissenbaum (1996), which foreground different sources that can lead to machine bias. These are: pre-existing bias – rooted in our institutions, practices, and attitudes (§3.3.1), technical bias – due to technical constraints and decisions (§3.3.2), and emergent bias – arising from the interaction between systems and users (§3.3.3). We consider such categories as placed along a continuum, rather than being discrete.

3.3.1 Pre-existing Bias

MT models are known to reflect gender disparities present in the data. However, reflections on such generally invoked disparities are often overlooked. Treating data as an abstract, monolithic entity (Gitelman, 2013) – or relying on "overly broad/overloaded terms like *training data bias*"[5] (Suresh and Guttag, 2019) – do not encourage reasoning on the many factors of which data are the product. First and foremost, the historical, socio-cultural context in which they are generated. A starting point to tackle these issues is the Europarl corpus (Koehn, 2005), a key resource in the MT community, which consists of the transcripts and aligned translations of the European Parliament sittings. In the corpus, however, only 30% of sentences are uttered by women (Vanmassenhove et al., 2018). Such an imbalance is a direct window into the glass ceiling that has hampered women's access to parliamentary positions.[6] This case exemplifies how data might be "tainted with historical bias", mirroring an "unequal ground truth"[7] (Hacker, 2018).

Other gender variables are harder to spot and quantify. Empirical linguistics research pointed out that subtle gender asymmetries are rooted in languages' use and structure. For instance, an important aspect regards how women are referred to. Femaleness is often explicitly invoked when there is no textual need to do so, even in languages that do not require overt gender marking. A case in point regards Turkish, which differentiates *cocuk* (child) and *kiz cocugu* (female child) (Braun, 2000). Similarly, in a corpus search, Romaine (2001) found 155 explicit female markings for *doctor* (*female*, *woman*, or *lady doctor*), compared to only 14 *male doctor*.

[5] See (Johnson, 2020a; Samar, 2020) for a discussion on how such a narrative can be counterproductive for tackling bias.

[6] To date, the highest peak of women's representation in the EU Parliament has been 40%. See Women in the European Parliament (infographics).

[7] In ML, the notion of ground truth refers to the factual information, or data, gathered from the real world. Supposedly, ground truth is the ideal expected result models should learn.

Feminist language critique provided extensive analysis of such a phenomenon by highlighting how referents in discourse are considered men by default unless explicitly stated (Silveira, 1980; Hamilton, 1991). A link to androcentrism in language has also been established with masculine generics, i.e., the prescriptive use of MASC forms for generic referents or mixed-gender groups (Moulton et al., 1978; Archer and Kam, 2022).[8] Indeed, prescriptive top-down guidelines still limit the linguistic visibility of gender diversity, as in the case of the Real Academia de la Lengua Española, which recently discarded the official use of non-binary innovations and claimed the functionality of masculine generics (Mundo, 2018; López et al., 2020).

By stressing such issues, we are not implying that pre-existing bias should be just reproduced in MT. Rather, the above-mentioned concerns are the starting point to account for when dealing with gender bias and to grasp its complex underlying dynamics.

3.3.2 Technical Bias

Technical bias comprises aspects related to data creation, model design, training and testing procedures. If present in training and testing samples, gender asymmetries are respectively learned by MT systems and rewarded in their evaluation. However, as just discussed, biased representations are not merely quantitative, but also qualitative. Accordingly, straightforward procedures – e.g., balancing the number of speakers in existing datasets – may not suffice to ensure a fairer representation of gender in MT outputs. Since datasets are a crucial source of bias, it is also crucial to advocate for a careful data curation (Paullada et al., 2020; Hanna et al., 2021; Rogers, 2021), guided by pragmatically- and socially-informed analyses (Hitti et al., 2019; Sap et al., 2020; Devinney et al., 2020) and annotation practices (Gaido et al., 2020c). We recognize that applying such practices on large training tests is non-trivial (Solaiman and Dennison, 2021). However, to date their incorporation for smaller, curated test tests paired with dedicated metrics is essential to unveil bias, and set which behaviour our models are expected to strive towards (Blodgett et al., 2021) (see Sec. 3.4).

Overall, while data can mirror gender inequalities and offer adverse shortcut learning opportunities, it is "quite clear that data alone rarely constrain a model sufficiently" (Geirhos et al., 2020) nor explain the fact that models *overamplify* (Shah et al., 2020) such inequalities in their outputs. Focusing on model components, Costa-jussà et al. (2022c) demonstrate that architectural choices in multilingual MT impact system behavior: shared encoder-decoders[9]

[8] As traced by Conrod (2020), the generic use of *he*, *she*, and *they* in English used to co-exist in apparently free variation. In the 18th century, grammarians forbade the use of *they*, some claiming its inappropriateness for formal writing (Curzan, 2003), others advocating for *he* due to a "hierarchy of the sexes" (Harvey, 1878).

[9] Unlike language-specific architectures, in multilingual MT a single shared encoder-decoder is trained with multiple input and output languages. This allows for generalizable translations between language pairs never seen during the training process.

retain less gender information in the source embeddings, thus disfavoring the generation of FEM forms. Vanmassenhove et al. (2019, 2021b), while investigating whether MT caused the loss and decay of certain words in translation, attest to the existence of an algorithmic bias that leads under-represented forms in the training data – as it may be the case for FEM references – to further decrease in the MT output. Finally, Roberts et al. (2020) prove that beam search – unlike sampling[10] – is skewed towards the generation of more frequent (MASC) pronouns.

To conclude, it emerges that efforts towards understanding and mitigating gender bias should not be reduced to a generically intended 'data problem', only. Rather, model design and its algorithmic implications should also be accounted for. To date, this remains quite unexplored.

3.3.3 Emergent Bias

Emergent bias may arise when a system is used in a different context than the one it was designed for, e.g., when it is applied to another demographic group. From car crash dummies to clinical trials, we have evidence of how not accounting for gender differences brings to the creation of male-grounded products with dire consequences (Liu and Dipietro Mager, 2016; Criado-Perez, 2019), such as higher death and injury risks in vehicle crashes and less effective medical treatments for women. Similarly, unbeknownst to their creators, MT systems that are not intentionally envisioned for a diverse range of users will not generalize for the feminine segment of the population. Hence, in the interaction with an MT system, a woman will likely be misgendered or not have her linguistic style preserved (Hovy et al., 2020). Other conditions of users/system mismatch may be the result of changing societal knowledge and values. A case in point regards Google Translate's historical decision to adjust its system for instances of gender ambiguity. Since its launch twenty years ago, Google had provided only one translation for single-word gender-ambiguous queries (e.g., *professor* translated in Italian with *professore* MASC).[11] In a community increasingly conscious of the power of language to hardwire stereotypical beliefs and women's invisibility (Lindqvist et al., 2019; Beukeboom and Burgers, 2019), the bias exhibited by the system was confronted with a new sensitivity. The decision of the service (Kuczmarski, 2018) to provide a double FEM/MASC output (*professor*→*professoressa*|*professore*) stems from current demands for gender-inclusive resolutions. As communicative practices evolve and for the recognition of non-binary groups (Richards et al., 2016), it is now an open question if and how such modeling could be integrated with non-binary inclusive strategies (Dev et al., 2021; Lauscher et al., 2022).

[10] Beam search and sampling are decoding algorithmic strategies that intervene when generating the output translation. That is, to pick the most likely sequence of target words.

[11] Note that the output gender is not pre-defined to be necessarily MASC.

3.4 Assessing Bias

First accounts on gender bias in MT date back to Frank et al. (2004). Their manual analysis pointed out how English-German MT suffered from a dearth of linguistic competence, as it showed severe difficulties in recovering syntactic and semantic information to correctly produce gender agreement. The inspected systems were rule-based models, which however presented gaps in feminine lexicon in their dictionaries and lacked morphological rules for feminine derivation. Similar inquiries were conducted on other target grammatical gender languages for several commercial MT systems, including statistical ones (Abu-Ayyash, 2017; Monti, 2017; Rescigno et al., 2020). While these studies focused on contrastive phenomena, Schiebinger (2014)[12] went beyond linguistic insights, calling for a deeper understanding of gender bias. Her article on Google Translate's "masculine default" behavior (see Sec. 3.1) emphasized how such a phenomenon is related to the larger issue of gender inequalities, also perpetuated by socio-technical artifacts (Selbst et al., 2019). All in all, these qualitative analyses demonstrated that gender problems encompass all three MT paradigms (neural, statistical, and rule-based), preparing the ground for quantitative work.

To attest the existence and scale of gender bias across several languages, dedicated benchmarks, evaluations, and experiments have been designed. We first discuss large scale analyses aimed at assessing gender bias in MT, grouped according to two main conceptualizations: *i)* works focusing on the weight of prejudices and stereotypes in MT (Sec. 3.4.1); *ii)* studies assessing whether gender is properly preserved in translation (Sec. 3.4.2). Finally, we review existing benchmarks for comparing MT performance across genders (Sec. 3.4.3).

3.4.1 MT and Gender Stereotypes

In MT, we record prior studies concerned with pronoun translation and coreference resolution across typologically different languages, accounting for both animate and inanimate referents (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010; Guillou, 2012). For the specific analysis of gender bias that follow here below, instead, such tasks are exclusively studied in relation to human entities.

Prates et al. (2018) and Cho et al. (2019) design a similar setting to assess gender bias. Prates et al. (2018) investigate pronoun translation from 12 genderless languages into English. They retrieve and then translate into the source 12 languages ~1,000 job positions from the U.S. Bureau of Labor Statistics, so to build simple constructions like the Hungarian "*ő* egy

[12] See also Schiebinger's project *Gendered Innovations*: `http://genderedinnovations.stanford.edu/case-studies/nlp.html`

mérnök" ("*he/she* is an *engineer*"). Following the same template, Cho et al. (2019) extend the analysis to Korean-English including both occupations and sentiment words (e.g., *kind*), an extension also later replicated for Hindi-English (Ramesh et al., 2021). As their samples are ambiguous by design, the observed predictions of he/she pronouns in translation should be random, yet they show a strong masculine skew.[13] To further analyze the under-representation of *she* pronouns, Prates et al. (2018) focus on 22 macro-categories of occupation areas and compare the proportion of pronoun predictions against the real-world proportion of men and women employed in such sectors. In this way, they find that MT not only yields a masculine preference, but it also underestimates feminine frequency at a greater rate than occupation data alone suggest. Such an analysis starts by acknowledging pre-existing bias – e.g., low rates of women in STEM – to attest the existence of machine bias, and defines it as the exacerbation of actual gender disparities.

Going beyond word lists and simple synthetic constructions, Gonen and Webster (2020) inspect the translation into Russian, Spanish, German, and French of natural yet ambiguous English sentences. Their analysis on the ratio and type of generated MASC/FEM job titles consistently exhibits social asymmetries for target grammatical gender languages (e.g., *lecturer* MASC vs. *teacher* FEM). Finally, Stanovsky et al. (2019) assess that MT is skewed to the point of actually ignoring explicit feminine gender information in source English sentences. For instance, MT systems yield a wrong MASC translation of the job title *baker*, although it is referred to by the pronoun *she*. Beside the overlook of overt gender mentions, the model's reliance on unintended (and irrelevant) cues for gender assignment is further confirmed by the fact that adding a socially connoted – but formally epicene – adjective (the *pretty* baker) pushes models towards FEM inflection in translation.

We observe that the propagation of stereotypes is a widely researched form of gender asymmetries in MT, one that so far has been largely narrowed down to occupational stereotyping. After all, occupational stereotyping has been studied by different disciplines (Greenwald et al., 1998) attested across cultures (Lewis and Lupyan, 2020), and it can be easily detected in MT across multiple language directions with consistent results. Current research should not neglect other stereotyping dynamics, as in the case of Stanovsky et al. (2019), and Cho et al. (2019), who include associations to physical characteristics or psychological traits. Also, the intrinsically contextual nature of societal expectations advocates for the study of culture and language-specific dimensions of bias. As a recent interesting perspective, consider the study

[13] Cho et al. (2019) highlight that a higher frequency of feminine references in the MT output does not necessarily imply a bias reduction. Rather, it may reflect gender stereotypes, as for *hairdresser* that is skewed towards feminine. This observation points to the tension between frequency count, suitable for testing under-representation, and qualitative-oriented analysis of bias conceptualized in terms of stereotyping.

by Ciora et al. (2021), which unlike the so far discussed inquiries, unveils gender bias into genderless languages too. In such a case, however, it emerges in a 'covert' manner, where only femininity is explicitly marked; e.g., "My *sister/brother* is a soccer player" into Turkish as *Kiz kardeşim* (female sibling) vs. *kardeşim* (sibling). As a final reflection, Blodgett (2021) underscores that not any statistical preference in models' output (e.g., *neighbor* skewed towards MASC, or *pregnant* as FEM), qualifies to be considered as normatively undesirable or harmful. That is, if they do not map to stereotypes or result in quality of service disparities, they might be just (occasionally undesirable) correlations.

3.4.2 MT and Gender Preservation

Vanmassenhove et al. (2018) and Hovy et al. (2020) investigate whether speakers' gender[14] is properly reflected in MT. This line of research is preceded by findings on gender personalization of statistical MT (Mirkin et al., 2015; Bawden et al., 2016; Rabinovich et al., 2017), which claim that gender "signals" are weakened in translation.

Hovy et al. (2020) conjecture the existence of age and gender stylistic bias due to models' under-exposure to the writings of women and younger segments of the population. To test this hypothesis, they automatically translate a corpus of online reviews with available metadata about users (Hovy et al., 2015). Then, they compare such demographic information with the prediction of age and gender classifiers run on the MT output. Results from the automatic classification indicate that different commercial MT models systematically make authors "sound" older and male. Their study thus concerns the under-representation of the language used "by" certain speakers and how it is perceived (Blodgett, 2021). However, the authors do not inspect which linguistic choices MT overproduces, nor which stylistic features may characterize – or be perceived as characterizing – different socio-demographic groups.

Still starting from the assumption that demographic factors influence language use, Vanmassenhove et al. (2018) probe MT's ability to preserve speaker's gender when translating from English into ten languages. To this aim, they develop gender-informed MT models that incorporate the gender of the speakers (more details in Sec. 3.5.1), and compare the outputs of such models with those obtained by their baseline counterparts. Evaluated on a test set for spoken language translation (Koehn, 2005), their enhanced models show consistent gains in terms of overall quality when translating into grammatical gender languages, where speaker's references are often marked. For instance, the French translation of "I'm *happy*" is either "Je suis *heureuse*" or "Je suis *hereux*" for a female/male speaker, respectively. But was such an

[14] Note that these studies establish *a priori* distinction of female/male speakers. As discussed in Section 2.2.4, we invite a reflection on the appropriateness and use of such categories.

Study	**Benchmark**	**Gender**	**Harms**
(Prates et al., 2018)	Synthetic, U.S. Bureau of Labor Statistics	b	R: under-rep, ster
(Cho et al., 2019)	Synthetic, Equity Evaluation Corpus (EEC)	b	R: under-rep, ster
(Ramesh et al., 2021)	Synthetic, Equity Evaluation corpus (EEC)	b	R: under-rep, ster
(Gonen and Webster, 2020)	BERT-based perturbations on natural sentences	b	R: under-rep, ster
(Stanovsky et al., 2019)	WinoMT	b	R: under-rep, ster
(Vanmassenhove et al., 2018)	Europarl (generic)	b	A: quality
(Hovy et al., 2020)	Trustpilot (reviews with gender and age)	b	R: under-rep

Table 3.1: For each **Study**, the Table shows on which **Benchmark** gender bias is assessed and how **Gender** is intended (here only in binary (b) terms). Finally, we indicate which (R)epresentational – *under-representation* and *stereotyping* – or (A)llocational **Harm** – as reduced *quality* of service – is addressed in the study.

approach equally beneficial and necessary for both genders? Through a more focused but still automatic cross-gender analysis – carried out by splitting their English-French test set into 1st person male vs. female data – they actually assess that *i)* the largest margin of improvement for their gender-informed approach concerns sentences uttered by women, and thus that *ii)* their baseline model exhibited a performance disparity in favor of male speakers. Besides morphological agreement, they also attribute such improvement to the fact that their enhanced model produces gendered preferences in other word choices (see Sec. 2.2.4). For instance, it opts for *think* rather than *believe*, in line with corpus studies claiming a tendency for women to use less assertive speech (Newman et al., 2008). Note, however, that the authors rely on manual analysis to ascribe performance differences to gender-related features (i.e., 1st person gender-marking or stylistic differences). In fact, global evaluations on generic test sets alone are inadequate to pointedly measure gender bias.

Toward this goal, dedicated benchmarks have been created, which we systematize in the following section. Also, in accordance with the human-centered approach embraced in this work, we contribute with Table 3.1, in which we map all works discussed so far to the harms (see Sec. 3.2) ensuing from the biased behaviors they assess.

3.4.3 Existing Benchmarks

MT outputs are typically evaluated against reference translations employing standard metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006a) (see Sec. 2.3.4). This procedure poses two challenges. First, these metrics provide coarse-grained scores for translation quality, as they treat all errors equally and are rather insensitive to specific linguistic phenomena (Sennrich, 2017). Second, generic test sets containing the same gender imbalance present in the training data can reward biased predictions. Differently, *Gender Bias Evaluation Testsets* (GBETs) (Sun et al., 2019a) are benchmarks designed to probe gender bias by isolating the impact of gender from other factors that may affect systems' performance. Note that different benchmarks and metrics respond to different conceptualizations of bias (Barocas et al., 2019). Common to them all in MT, however, is that biased behaviors are formalized by using some variants of averaged performance[15] disparities across gender groups, comparing the accuracy of gender predictions on an equal number of masculines, feminine, and neutral references.

Hereby, we describe the publicly available GBETs for MT. In doing so, we distinguish between: *i) challenge* test sets, which consist of *ad-hoc* created artificial sentences, and *ii) natural* test sets, which are built on naturally occurring instances of language.

3.4.3.1 Challenge test sets

Escudé Font and Costa-jussà (2019) developed the bilingual English-Spanish **Occupations test set.** It consists of 1,000 sentences, equally distributed across genders. The phrasal structure envisioned for their sentences is "I've known *{her|him|<proper noun>}* for a long time, my friend works as {a|an} <occupation>". The evaluation focuses on the translation of the noun *friend* into Spanish (*amigo/a*). Since gender information is present in the source context and sentences are the same for both masculine/feminine participants, an MT system exhibits gender bias if it disregards relevant context and cannot provide the correct translation of *friend* at the same rate across genders.

Stanovsky et al. (2019) created **WinoMT** by putting together two existing English GBETs for coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a). The corpus consists of 3,888 sentences with a Winograd-schema: that is, sentences presenting two human entities defined by their role and a subsequent pronoun that needs to be correctly resolved to one of the entities (e.g., "The *lawyer* yelled at the *hairdresser* because *he* did a bad job"). For each sentence, there are two variants – with either *he* or *she* pronouns – so to cast the referred entity (*hairdresser*) into a proto- or anti-stereotypical gender role. By translating WinoMT into grammatical gender languages, one can thus measure systems' ability to resolve the anaphoric relation and pick the correct feminine/masculine inflection for the occupational noun. On top of quantifying under-representation as the difference between the total amount of translated feminine and masculine references, the subdivision of the corpus into proto- and anti-stereotypical sets also allows verifying if MT predictions correlate with occupational stereotyping.

Along this line, Saunders et al. (2020) enriched the original version of WinoMT in two different ways. First, they included a third gender-neutral case based on the singular *they* pronoun, thus paving the way to account for non-binary referents. Second, they labeled the entity in the sentence which is not coreferent with the pronoun (*lawyer*). The latter annotation

[15] This is a value-laden option (Birhane et al., 2020), and not the only possible one (Mitchell et al., 2020). For a broader discussion on measurement and bias, we refer the reader also to (Jacobs, 2021; Jacobs et al., 2020).

is used to verify the shortcomings of some mitigating approaches, as discussed in Section 3.5.

SimpleGen (Renduchintala et al., 2021) is designed with the scope of testing MT systems on very 'simple' instances and identify egregious failures when translating gender within an English-German/Spanish setting. Unlike WinoMT – where examples are potentially open to different interpretations[16] – this dataset is built on sentence templates with a short, unambiguous syntactic structure, consisting of a lexically gendered item and an occupational noun (e.g., My *father* is a *mechanic*). It includes 1332 sentences, equally distributed across proto-/anti-stereotypical gender associations.

The above-mentioned corpora are known as *challenge sets*, consisting of sentences created *ad hoc* for diagnostic purposes. In this way, they can be used to quantify bias related to stereotyping and under-representation in a sound environment. However, since they consist of a limited variety of synthetic gender-related phenomena, they hardly address the variety of challenges posed by real-world language and are relatively easy to overfit. As recognized by Rudinger et al. (2018), "they may demonstrate the presence of gender bias in a system, but not prove its absence".

3.4.3.2 Natural test sets.

The evaluation subset of the **BUG** corpus (Levy et al., 2021) comprises 865 English sentences where a pronoun needs to be resolved to an entity in the form of an occupational noun (e.g., "With *his* dark hair and complexion, the ballet *dancer* was often cast in more exotic roles"). Such sentences, which are annotated as pro-/anti-stereotypical, do not however rely on a fixed set of artificial templates. Rather, BUG is created by extracting diversified, "in the wild" sentences with challenging cases of gender-role assignment that naturally occur in existing corpora, e.g., Wikipedia or medical data.

The Arabic Parallel Gender Corpus (APGC) (Habash et al., 2019) includes an English-Arabic test set[17] retrieved from OpenSubtitles natural language data (Lison and Tiedemann, 2016). Each of the 2,448 sentences in the set exhibits a first-person singular reference to the speaker (e.g., "I'm *rich*"). Among them, ~200 English sentences require gender agreement to be assigned in translation. These were translated into Arabic in both gender forms, obtaining a quantitatively and qualitatively equal amount of sentence pairs with annotated masculine/feminine references. This natural corpus thus allows for cross-gender evaluations on MT production of correct speaker's gender agreement.

[16] In the example above, "The lawyer yelled at the hairdresser because *he* did a bad job", where *he* could also refer to a third party not mentioned in the sentence.

[17] Overall, the corpus comprises over 12,000 annotated sentences and 200,000 synthetic sentences to be also used for training.

Unlike challenge sets, natural corpora enable the evaluation of bias under more authentic conditions, to assess if MT yields reduced feminine representation and if the quality of service varies across speakers of different genders. However, if they treat all gender-marked words equally, as in the case of the APCG corpus, it is not possible to identify if the model is propagating stereotypical representations.

All in all, we stress that each test set and metric is only a proxy for framing a phenomenon or an ability (e.g., anaphora resolution), and an approximation of what we truly intend to gauge. Thus, ideally, advances in MT should account for the observation of gender bias in real-world conditions to avoid that achieving high scores on a mathematically formalized esteem could lead to a false sense of security. Still, benchmarks remain key, valuable tools to continuously monitor models' behavior. As such, we believe that evaluation procedures should resemble authentic conditions as much as possible. Also, they ought to cover both models' general performance and gender-related issues. This is crucial to establish the capabilities and limits of mitigating strategies.

3.5 Mitigating Bias

To attenuate gender bias in MT, different strategies dealing with input data, learning algorithms, and model outputs have been proposed. As attested by Birhane et al. (2020), since advancements are oftentimes exclusively reported in terms of values internal to the machine learning field (e.g., efficiency, performance), it is not clear how such strategies are meeting societal needs by reducing MT-related harms. In order to conciliate technical perspectives with the intended social purpose, in Table 3.2 we map each mitigating approach to the harms they are meant to alleviate, as well as to the benchmark their effectiveness is evaluated against. Orthogonally, we hereby describe each approach by means of two technically-oriented categories: model debiasing (Sec. 3.5.1) and debiasing through external components (Sec. 3.5.2). As we discuss in the following, these two families of methods affect a model differently.

3.5.1 Model Debiasing

This line of work focuses on mitigating gender bias through architectural changes of general-purpose MT models or via dedicated training procedures. Thus, such approaches are applied to obtain a single mitigated model without needing further external components, but this comes with the cost of having to train such a model from scratch.

Approach	**Authors**	**Benchmark**	**Gender**	**Harms**
Gender tagging (sentence-level)	Vanmassenhove et al.	Europarl (generic)	b	R: under-rep, A: qual
	Elaraby et al.	Open subtitles (generic)	b	R: under-rep, A: qual
Gender tagging (word-level)	Saunders et al.	expanded WinoMT	nb	R: under-rep, ster
	Stafanovičs et al.	WinoMT	b	R: under-rep, ster
Adding context	Basta et al.	WinoMT	b	R: under-rep, ster
Word-embeddings	Escudé Font and Costa-jussà	Occupation test set	b	R: under-rep
Fine-tuning	Costa-jussà and de Jorge	WinoMT	b	R: under-rep, ster
Black-box injection	Moryossef et al.	Open subtitles (selected sample)	b	R: under-rep, A: qual
Lattice-rescoring	Saunders and Byrne	WinoMT	b	R: under-rep, ster
Re-inflection	Habash et al.; Alhafni et al.	Arabic Parallel Gender Corpus	b	R: under-rep, A: qual

Table 3.2: For each **Approach** and related **Authors**, the Table shows on which **Benchmark** it is tested, if **Gender** is intended in binary terms (b), or including non-binary (nb) identities. Finally, we indicate which (R)epresentational – *under-representation* and *stereotyping* – or (A)llocational **Harm** – as reduced *quality* of service – the approach attempts to mitigate.

Gender tagging. To improve the generation of speaker's referential markings, Vanmassenhove et al. (2018) prepend a gender tag (<M> or <F>) to each source sentence, both at training and inference time. As their model is able to leverage this additional information, the approach proves useful to handle morphological agreement when translating from English into French. However, this solution requires additional metadata regarding the speakers' gender that might not always be feasible to acquire. Automatic annotation of speakers' gender (e.g., based on first names) is not advisable, as it runs the risk of introducing additional bias by making unlicensed assumptions about one's identity.

Elaraby et al. (2018) bypass this risk by defining a comprehensive set of cross-lingual gender agreement rules based on Parts-of-Speech (POS) tagging. In this way, they identify speakers' and listeners' gender references in an English-Arabic parallel corpus, which is consequently labeled and used for training. The idea, originally developed for spoken language translation in a two-way conversational setting, can be adapted for other languages and scenarios by creating new dedicated rules. However, in realistic deployment conditions where reference translations are not available, gender information still has to be externally supplied as metadata at inference time.

Stafanovičs et al. (2020) and Saunders et al. (2020) explore the use of word-level gender tags. While Stafanovičs et al. (2020) just report a gender translation improvement, Saunders et al. (2020) rely on the expanded version of WinoMT to identify a problem concerning gender tagging: it seems to introduce noise if applied to sentences with references to multiple participants (e.g., The *lawyer* yelled at the *hairdresser* <M> because *he* did a bad job), as it pushes their translation towards the same gender (e.g., both *lawyer* and *hairdresser* rendered as MASC). Saunders et al. (2020) also include a first non-binary exploration of neutral translation by exploiting an artificial dataset, where neutral tags are added and gendered inflections are replaced by placeholders. The results are however inconclusive, most likely due to the small size and synthetic nature of their dataset.

Adding context. Without further information needed for training or inference, Basta et al. (2020) adopt a generic approach and concatenate each sentence with its preceding one. The underlying idea is that useful information to translate gender might be outside the single sentence. By providing more context, they attest a slight improvement in gender translations requiring anaphorical coreference to be solved in English-Spanish. This finding motivates exploration at the document level, i.e., with MT systems that perform translation for larger spans of text than isolated sentences. However, it should be validated with manual (Castilho et al., 2020) and interpretability analyses since the added context can be beneficial for gender-unrelated reasons, such as acting as a regularization factor (Kim et al., 2019). That is, positively affecting the training of the model, hence its performance, for causes that do not reflect improved linguistic ability, e.g., cohesion for retrieving gender information.

Debiased word embeddings. The two above-mentioned mitigations share the same intent: supply the model with additional gender knowledge. Instead, Escudé Font and Costa-jussà (2019) leverage pre-trained word embeddings,[18] whose representations are modified to be 'fairer' by using the hard-debiasing method (Bolukbasi et al., 2016) or the GN-GloVe algorithm (Zhao et al., 2018c). These methods roughly remove gender associations from the representations of English words that are formally neutral, yet highly socially connoted. To exemplify, these strategies alter the embedding space of words such as *nanny*, or *plumber*, to ensure that their semantic representations do not contain a gender component e.g., the embedding for *nanny* is not more closely related to *she* than *he* (see Section 2.3.1.2). Escudé Font and Costa-jussà (2019) employ such embeddings on the decoder side, the encoder side, and both sides of an English-Spanish model. The best results are obtained by leveraging GN-GloVe embeddings on both encoder and decoder sides, increasing BLEU scores and gender accuracy. Note that the authors generically apply debiasing methods developed for English also to their target language. However, being Spanish a grammatical gender language, other language-specific approaches should be considered to preserve the quality of the original embeddings (Zhou et al., 2019; Zhao et al., 2020). Indeed, as described in Sec. 2.2.1.1, in Spanish all nouns are marked for gender, whether on a formal basis (e.g., chair, *silla* FEM), or for semantic distinctions (e.g.,

[18] Namely, using embeddings that are already trained on a given dataset. Instead of initializing NMT weights randomly, pre-trained embeddings are supplied to the model and used as initialization weights.

friend *amigo* MASC vs *amiga* FEM) In both cases, preserving the gender information is relevant, whether to ensure grammatical agreement or actually convey gender distinctions. We also stress that it is debated whether depriving systems of some knowledge and "blind" their perceptions – rather than 'teaching' them how to perform – is the right path towards fairer language models (Dwork et al., 2012; Caliskan et al., 2017; Gonen and Goldberg, 2019; Nissim and van der Goot, 2020). Also, Goldfarb-Tarrant et al. (2021) question the underlying assumption that adjusting models' inner representations will reduce bias in their output, as no reliable correlation is found between intrinsic evaluations of bias in word-embeddings and cascaded effects on models' biased behavior.

Balanced fine-tuning. Costa-jussà and de Jorge (2020) rely on Gebiotoolkit (Costa-jussà et al., 2020) to build gender-balanced datasets (i.e., featuring an equal amount of references across genders) based on Wikipedia biographies. They fine-tune (i.e., adjusting by means of a final training step) their models on such natural and more even data, and evaluate them on the WinoMT set. By generically looking at the amount of FEM forms in the output, they find that their generation is overall improved. However, by specifically checking results on the anti-stereotypical portion of WinoMT – where FEM translations are expected for highly masculine connoted occupations, e.g., *doctor* – the approach does not appear as effective. Indeed, they employ a straightforward method that aims to increase the amount of feminine Wikipedia pages in their training data. However, such coverage increase does not mitigate stereotyping harms, as it does not account for the qualitatively different ways in which men and women are portrayed (Wagner et al., 2015).

3.5.2 Debiasing through External Components

Instead of directly debiasing the MT model, these mitigating strategies intervene in the inference phase with external dedicated components. Such approaches do not imply retraining, but introduce the additional cost of maintaining separate modules and handling their integration with the MT model.

Black-box injection. Moryossef et al. (2019) attempt to control the production of feminine references to the speaker and numeral inflections (plural or singular) for the listener(s) in an English-Hebrew spoken language setting. To this aim, they rely on a short construction, such as "*she* said to *them*", which is prepended to the source sentence and then removed from the MT output. Their approach is simple, it can handle two types of information (gender and number) for multiple entities (speaker and listener), and improves systems' ability to generate feminine target forms. However, as in the case of Vanmassenhove et al. (2018) and Elaraby et al. (2018), it requires metadata about speakers and listeners.

Lattice re-scoring. Saunders and Byrne (2020) propose to post-process the MT output with a lattice re-scoring module. This module exploits a transducer to create a lattice by mapping gender-marked words in the MT output to all their possible inflectional variants. Developed for German, Spanish, and Hebrew, all the sentences corresponding to the paths in the lattice are re-scored with another model, which has been gender-debiased but at the cost of lower generic translation quality. Then, the sentence with the highest probability is picked as the final output. When tested on WinoMT, such an approach leads to an increase in the accuracy of gendered forms selection. Note that the gender-debiased system is created by fine-tuning the model on an *ad hoc* built tiny set containing a balanced amount of MASC/FEM forms. Such an approach, also known as *counterfactual data augmentation* (Lu et al., 2019), requires creating identical pairs of sentences differing only in terms of gender references. In fact, Saunders and Byrne (2020) compile English sentences following this schema: “The <profession> finished <*his*|*her*> work”. Then, the sentences are automatically translated and manually checked. In this way, they obtain a gender-balanced parallel corpus. Thus, to implement their method for other language pairs, the generation of new data is necessary. For the fine-tuning set, the effort required is limited as the goal is to alleviate stereotypes by focusing on a pre-defined occupational lexicon. However, data augmentation is very demanding for complex sentences that represent a rich variety of gender agreement phenomena[19] such as those occurring in natural language scenarios.

Gender re-inflection. Habash et al. (2019) and Alhafni et al. (2020) confront the problem of speaker’s gender agreement in Arabic with a post-processing component that re-inflects 1st person references into masculine/feminine forms. In Alhafni et al. (2020), the preferred gender of the speaker and the translated Arabic sentence are fed to the component, which re-inflects the sentence in the desired form. In Habash et al. (2019) the component can be: *i*) a two-step system that first identifies the gender of 1st person references in an MT output, and then re-inflects them in the opposite form; *ii*) a single-step system that always produces both forms from an MT output. Their method does not necessarily require speakers’ gender information: if metadata are supplied, the MT output is re-inflected accordingly; differently, both MASC/FEM inflections are offered (leaving to the user the choice of the appropriate one). The implementation of the re-inflection component was made possible by the Arabic Parallel Gender Corpus (see Sec. 3.4.3), which demanded an expensive work of manual data creation. However, such corpus grants research on English-Arabic the benefits of a wealth of gender-informed natural language data that have been curated to avoid hetero-centrist interpretations and preconceptions (e.g.,

[19] Zmigrod et al. (2019) proposed an automatic approach for augmenting data into morphologically-rich languages, but it is only viable for simple constructions with one single entity.

speakers of sentences like “that’s my wife” are flagged as gender-ambiguous). Along the same line, Google Translate also delivers two outputs for short gender-ambiguous queries (Johnson, 2020b), though the service is not available for any language pair.

In light of the above, we can see that there is no conclusive state-of-the-art method for mitigating bias. The discussed interventions in MT tend to respond to specific aspects of the problem with modular solutions. If and how they should be integrated within the same MT system remains unclear. Indeed, as we have discussed throughout the chapter, the umbrella term “gender bias” refers to a wide array of undesirable phenomena. Thus, it is unlikely that a one-size-fits-all solution will be able to tackle problems that differ from one another, as they depend on e.g., how bias is conceptualized, the language combinations, the kinds of corpora used (Ciora et al., 2021). In light of the foregoing, generalization and scalability should not be the only criteria against which mitigating strategies are valued. Conversely, we should make room for language-specific and context-aware interventions and inquiries. Finally, gender bias in MT is a socio-technical problem. As addressed in Section 3.3, it is thus crucial to highlight that gender bias in MT does not start nor end within technological artifacts. As such, engineering interventions alone are not a panacea (Chang, 2019; D’Ignazio and Klein, 2020), and are rather to be viewed as specific countermeasures towards a more equal, and considerate deployment of language tools.

3.6 Conclusion

In this chapter, I have confronted the rapidly developing line of research on gender bias in MT. To this scope, I presented current studies on the topic within a unified framework to critically overview current conceptualizations and approaches to the problem. Since gender bias is a multifaceted and interdisciplinary issue, this chapter has intertwined knowledge from related disciplines, which is instrumental for a comprehensive investigation of bias. Concurrently, the offered systematization serves to contextualize how this thesis relates to the existing literature on gender bias in MT. I thus identify unexplored areas and gaps, to be addressed in the upcoming chapters.

From the review, we can see how gender bias in non-textual modalities to automatic translations (e.g., audiovisual) has been largely neglected. Dealing with different input data and features, however, can present gender cues that are foreign to text-to-text MT. In fact, image-guided translation has been claimed useful for gender translation since it relies on visual inputs for disambiguation (Frank et al., 2018; Ive et al., 2019), e.g., processing an image to pick the correct gender form in a caption of a “tennis player”. Seemingly helpful visual information, however, also pose new risks. That is, the risk of bending towards stereotypical assumptions concerning gendered appearances (e.g., in terms of clothing, attitude, and traits), thus adopting a reductionist view on (human) gender, whilst making systems potentially harmful for a diverse range of users (van Miltenburg, 2016; Hamidi et al., 2018; Keyes, 2018; May, 2019). Along this line, I focus on ST to investigate the relation between vocal characteristics and gender, starting from Chapter 5.

Henceforth, I continue exploring gender bias in ST accounting for different factors. Since its onset, work on bias has been largely data-centric. However, as mentioned in Section 3.3.2, algorithmic components do constrain model predictions. State-of-the-art architectural choices and algorithms in MT, however, have mostly been studied in terms of overall translation quality without specific analyses regarding gender translation. On this basis, in Chapter 6, I audit a crucial component of translation models: segmentation techniques. To foreground the interaction between modeling choices and language specificity in light of gender bias, in Chapter 7 I then carry out a multifaceted evaluation of different ST systems. Such an evaluation, controls both the paradigmatic (gender inflection) and syntagmatic axes (gender agreement) of gender. Finally, in Chapter 8, I adopt a shift in perspective, which does not only consider how gender bias is exhibited in the final output of the model. Along the line of explainability-oriented studies attempting to make opaque neural approaches to NLP more transparent (Belinkov et al., 2020), I explore the evolution of gender bias and gender capabilities over the whole course of ST systems' training.

All in all, several analyses and insights on gender bias await the reader in the forthcoming pages. As discussed in Section 3.4.3, none of those would be possible without a dedicated, gender-sensitive benchmark, able to disentangle generic performance from gender as a variable. Accordingly, it is precisely from the creation of such a benchmark that the next Chapter 4 begins.

4

The MuST-SHE Benchmark

An essential requirement towards assessing gender bias is the availability of a dedicated test set and evaluation procedure, which can target and isolate the impact of gender phenomena in translation from other unrelated factors. To this aim, in this chapter I introduce MuST-SHE, a multilingual, natural benchmark for three language pairs (English → French/Italian/Spanish). Built on TED talks data (Cattoni et al., 2021), for each language pair it comprises ~1,000 (*audio*, *transcript*, *translation*) triplets, annotated with qualitatively differentiated and balanced gender-related phenomena. In light of these features, MuST-SHE allows testing both MT and ST systems, by complementing BLEU with accuracy-based evaluations on gender bias for a diversified range of phenomena.

4.1 Introduction

To assess and monitor gender bias we need dedicated resources able to unveil its presence in current systems. As discusses in Section 3.4, several GBETs have been created towards this goal. On the one hand, the vast majority of such evaluation sets are challenge datasets (Prates et al., 2018; Cho et al., 2019; Escudé Font and Costa-jussà, 2019; Stanovsky et al., 2019; Renduchintala and Williams, 2021), which are simple, artificially-built constructions for controlled experiments. However, artificial data characterized by a qualitatively limited variety of phenomena generate constrained environments, and provide a limited perspective on the extent of the bias phenomenon beyond occupational stereotyping. On the other hand, the generic natural corpora built on conversational language that were used in few studies (Elaraby et al., 2018; Moryossef et al., 2019; Vanmassenhove et al., 2018) include only a restricted quantity of not isolated gender-expressing forms, thus not permitting either extensive or targeted evaluations. The natural GBET created by Habash et al. (2019), instead, does include a higher degree of variability for a balanced number of FEM/MASC phenomena. However, since this bilingual corpus is only available for English-Arabic, it does not allow for comparison across language pairs. Finally, all the above-mentioned test sets are exclusively built on textual data, thus restricting their scenario of use to text-to-text MT. Accordingly, they do not allow inspecting if and how ST systems are affected by gender bias, especially in relation to their potential tendency to pick up on audio cues to perform gender translation.

In light of the above, in this chapter I present MuST-SHE,[1] a natural, multilingual GBET to assess gender translation and gender bias for both MT and ST. MuST-SHE was built on spoken language data as a subset of MuST-C (Di Gangi et al., 2019a; Cattoni et al., 2021), the largest freely available multilingual corpus for ST, which comprises (*audio*, *transcript*, *translation*) triplets extracted from TED talks. To compile MuST-SHE, we exploited subportions of the MuST-C dataset and newly web-crawled TED talks, where we targeted linguistic phenomena that entail gender translation from English into three grammatical gender languages: Italian, French, and Spanish. Such language pairs were chosen to capture cross-lingual differences between the source English language, and three Romance languages that extensively express gender via feminine or masculine morphological markers on nouns, adjectives, verbs as well as functional words (see Sec. 2.2.2). As MuST-SHE has been enriched with several annotated features, it allows for pinpointed evaluations on whether current systems yield an unequal quality of service and the under/over-representation of FEM vs. MASC forms in translation. We implement a new evaluation method to make BLEU scores informative about gender by removing unrelated factors that may affect the overall performance of a system to soundly estimate gender bias.

Starting from Section 4.2, I present the cross-lingual empirical study on a selected sample of the parallel MuST-C corpus (Cattoni et al., 2021) that has informed MuST-SHE design: its curation rationale and the classification of the categories of gender phenomena we decided to address (Section 4.2.1). In Section 4.3, I describe MuST-SHE creation and annotation process, as well as its statistics. Then, in Section 4.4, I introduce MuST-SHE gender-sensitive evaluation procedure and metrics.

[1] MuST-SHE is released under a CC BY NC ND 4.0 International license, and is freely downloadable at `ict.fbk.eu/must-she`.

To ensure transparency and reproducibility, a thorough account of MuST-SHE characteristics can also be found in its data statement, which is made available at:
https://ict.fbk.eu/must-she-data-statement/

4.2 On First Looking into MuST-C Parallel Data

The MuST-C corpus (Di Gangi et al., 2019a; Cattoni et al., 2021) is a key resource for the ST community in light of its size and features. As a matter of fact, the dataset comprises several hundred hours of audio recordings from English TED Talks, which are aligned at the sentence level with their manual English transcripts and target translations. Besides being multilingual, MuST-C – and more generally TED talks – represent a variety of topics and highly proficient English speakers, whose information is publicly available to ensure transparent validation (see Sec. 4.3). In light of these aspects, the types of data represented in MuST-C made ideal candidates for the compilation of a natural, multilingual test set to be used for both MT and ST. Thus, among other existing corpora (Post et al., 2013; Kocabiyikoglu et al., 2018; Sanabria et al., 2018), we opted for the creation of a MuST-C-like benchmark, following its (*audio-transcript-translation*) structure and TED talks spoken language genre.

To gain a better insight into TED talks linguistic data used in MuST-C and capture the implications of gender translation, we initially conducted a qualitative cross-lingual analysis on 2,500 parallel sentences randomly sampled from the corpus. More specifically, we focused on its parallel textual (*transcript*, *translation*) portion with the twofold intent of *i)* identifying the circumstances in which referential gender assignment is implied in translation, and *ii)* inspecting how gender is realized and distributed across our three language pairs. Such insights have, respectively, informed a classification of our phenomena of interest (Sec. 4.2.1), and an automatic approach to maximize the extraction of segments with gendered phenomena from TED data (Sec. 4.3).

Towards the inspection of cases relevant to the understanding of gender bias, the parallel sentences captured in our analysis all entail the translation of at least one source English unmarked word into the corresponding MASC or FEM target word(s), where such a distinction is made on a semantic basis for human referents. Hence, we discarded all instances that allowed for a one-to-one mapping of gendered forms in translation (e.g., en: *she* FEM is kind → it: *lei* FEM è gentile), as well as non-human referents (e.g., *the cat* → it: *il gatto* MASC vs. *la gatta* FEM). On this basis, examining the parallel sentences that met our criteria led to the classification of four distinctive circumstances for gender translation, both human and automatic.

4.2.1 Categorization of gender phenomena

Unlike template sentences from challenge datasets that replicate the same type of syntactic construction and gender assignment scenario, naturally occurring data exhibit a high degree of variability and complexity. To frame different translation scenarios and their implications for cross-lingual gender transfer, our analysis thus differentiates between 4 different categories of gender phenomena encountered in MuST-C spoken language data. The categories are classified based on the type and source of information needed to disambiguate gender translation. In what follows, we present them via English-Italian example sentences from MuST-C:

Category 1. First-person singular references (to the person uttering the sentence), which reflect and are to be translated according to the speaker's linguistic expression of gender. Such a piece of information, however, is not available in the transcribed sentence.

A.

En: I was **born** and **brought up** in Mumbai.
It: Sono **nata** e **cresciuta** a Mumbai. (FEM)

B.

En: So, I decided I was going to **become a scuba diver** at the age of 15.
It: Quindi, ho deciso che sarei **diventato un sommozzatore** all'età di 15 anni. (MASC)

From a purely textual perspective, the displayed examples A and B are completely ambiguous, and can be translated into both masculine/feminine-inflected Italian versions, without affecting grammaticality. The distinction only depends on the referent's gender. Indeed, as the human reference translations show, translators have access to external information and their own world knowledge concerning TED speakers – which are oftentimes publicly renowned figures – and thus pick the suitable gender inflection in the target language. For textually-bound MT, however, such an information is out of reach, and can only be supplied through injected metadata. Differently, and more crucially, Category 1 represents those cases where direct ST systems (Sec. 2.3.3) might be bound to exploit audio information to correlate speakers' vocal characteristics and gendered linguistic forms.

Category 2. Gender phenomena in reference to any participant, which are to be translated according to contextual gender information available in the sentence, like lexically gendered words (*sister*, *Mr*), pronouns (*he, she*) and proper nouns (*Annie, Dean Kamen*).[2]

[2] Although proper names do not generally unambiguously map to one gender Ackerman (2019), for this binary task they are treated as culturally specific cues congruent with the gendered inflections associated with specific individuals in the human reference translations.

C.

En: **My** mom is also **a** great **friend of mine**.
It: **Mia** madre è anche **una mia** grande **amica**. (FEM)

D.

En: I was **a** popular young man, socially **well-adjusted**.
It: Ero **un** ragazzo popolare, socialmente **inserito**. (MASC)

For phenomena from Category 2, target words can be unambiguously translated, as a "gender cue" within the sentence context signals the particular gender marking that should be adopted in the target language to ensure agreement. In examples C and D, such cues are both lexically gendered nouns (the kinship term *mom* and *man*), where their disregard can lead to factual errors and agrammatical translations. As shown in example D, the referred participant can be the speaker, too. Regardless of whether such a sentence is provided as a textual or audio input, MT, ST as well a human translators are all put in the same conditions to retrieve the necessary information to perform gender translation.

Category 3. Gender phenomena in reference to both the speaker and a second participant, where gender translation requires accounting for both speakers' linguistic gender expression and for contextual gender information.

E.

En: Camilla and I have now **been** to other organizations.
It: Camilla e io siamo **state** in altre organizzazioni. (FEM)

In example E, the inflection for *state* (feminine and plural) can be only assigned by retrieving information about both the speaker and the second participant, via the proper noun cue "Camilla". Indeed, according to the still-standing normative rules governing the use of masculine generics, MASC forms are used in reference to gender-mixed groups. Instead, the use of FEM forms is restricted to all-feminine plural references. This is a complex category, in-between Category 1 and 2, where the human translator can access the speaker's information to pick a target feminine form. Instead, as in the case of Category 1, MT is not provided with such a piece of information, and it remains an open question whether ST attempts to infer it.

Category 4. Gender phenomena for under-specified referents, where no information to disambiguate gender is available.

F.

En: There was **one black professor**.

It: C'era **un professore nero**. MASC

G.

En: What do you think a batting average for **a cardiac surgeon** or **a nurse practitioner** or **an orthopedic surgeon**, **an OBGYN** is supposed to be?

It: Quale credete debba essere la media di battuta per **un cardiochirurgo** (MASC.) o **una infermiera professionista** (FEM) o **un chirurgo ortopedico** (MASC), **un'ostetrica** (FEM)?[3]

No information about any of the referents is available within the sentence. For a human translator, the F and G cases might be different. In the former, the gender of the black professor might have been explicitly stated within the larger context of the whole TED talk. In the latter, however, the professional nouns are used generically, not in relation to any specific individuals. Here, for G, we can see in the reference translation how gender assignment is completely arbitrary, and potentially influenced by social expectations and gendered roles. In the case of automatic systems that work at the sentence-level, however, such a fine-grained distinction between F and G does not hold. Both sentences, in fact, remain equally ambiguous, and translation particularly prone to reflect stereotypical associations. Concurrently, which gender-marked forms should be suitably picked in translation, however, remains to be defined.

Overall, this 4-tiered categorization represents a useful scaffolding to frame the implications and challenges of gender translation, upon which we relied for the creation of the MuST-SHE corpus. Indeed, as we will describe in the upcoming section, we attempted to represent different categories of gender phenomena within the test set, so to explore the behavior of current systems under different conditions. Due to the rarity of their occurrences, however, only few phenomena from the 3rd and 4th category were detected, and they did not suffice for proper evaluation.[4] Thus, henceforth, we will only discuss the representation of Category 1 and 2 within MuST-SHE.

[3] Although not largely in use, the feminine inflection for surgeon (*chirurga*) is attested and would have been acceptable. See https://www.treccani.it/vocabolario/chirurgo/

[4] Overall, across the three language pairs only 28 segments for Category 3 and 103 for Category 4 segments could be extracted from our full TED sample. Category 3 cases were overall extremely rare. Differently, MASC gender phenomena abounded for Category 4 due to the use of masculine generics, but they could not be balanced with FEM ones. All these segments are included within MuST-SHE release, but have not been employed for any of the experiments included in this thesis.

4.3 Dataset Creation and Annotation

A critical requirement for any test set is that of being *blind*; namely, consisting of sentences that the model has not previously seen and learned from over the course of training. Indeed, this is essential to ensure that models have not simply memorized their training sample, and can rather handle and generalize to new, unseen data. Thus, to create MuST-SHE, we could not make use of any language data from MuST-C training portion. As the remaining MuST-C data (~40 TED talks) proved however insufficient to compile the test set, we acquired new TED talks data. To do so by ensuring conformity, we followed the automatic procedure used to create MuST-C (Di Gangi et al., 2019a; Cattoni et al., 2021). Accordingly, after having downloaded TED videos and the files containing both transcripts and translations from the official website, *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. As reported in the original MuST-C paper (Cattoni et al., 2021), this automatic procedure was assessed to generate 90% of properly aligned triples on average.

On this basis, we gathered more than 250 additional TED talks. Together with MuST-C usable data, we thus obtained a sample of more than 300 TED talks, from which the segments included in MuST-SHE have been retrieved. Below, we describe in detail how MuST-SHE has been compiled, manually checked, and annotated.

4.3.1 Data Collection and Filtering

By design, and informed by the qualitative analysis insights in Sec 4.2, we compiled MuST-SHE with (*audio, transcript, translation*) segments that include least one gender-neutral expression in the source English utterance translated into the corresponding masculine/feminine target word(s). To target such segments in our TED Talks sample, we started from the textual (*transcript, translation*) portion of the corpus by means of an automatic approach, which aimed to quantitatively and qualitatively maximize the extraction of an assorted variety of gender-marked phenomena belonging to categories 1 and 2.

First, we employed regular expressions[5] to transform gender-agreement rules into search patterns to be applied to TED parallel data. Our queries were designed and adapted to the targeted language pairs, categories, and masculine/feminine forms. For instance, a (loose) regular expression intended to capture English-Italian FEM phenomena from Category 1 specified that: *i)* the English source sentence had to include the pronoun "*I*" and verb "*have*", whereas *ii)* the Italian translation had to include the verb "*sono*", followed by a word with the suffixes *-ata,*

[5] Regular expressions are used to construct search strings that search and locate text patterns.

-ita, -uta. This allowed us to identify target segments with gender-marked past participles (e.g., I have *felt* → it: Mi sono ***sentita***). To specifically match a differentiated range of gender-marked lexical items, we also compiled three series of 50 human-referring adjectives in Italian, French, and Spanish (e.g., "adventurous" in it: *avventuroso/a*; fr: *aventureux/euse*; es: *aventurero/a*) as well as a list with more than 1,000 English occupation nouns obtained from the US Department of Labour Statistics (e.g., *carpenter, assistant, secretary*)[6] (Prates et al., 2018). These list-based queries were also designed to maximize the overlap of segments with gender phenomena across the three language pairs. Accordingly, to ensure the collection of a common subset of segments across en-it, en-fr, and en-es, corresponding source English transcripts were automatically identified by means of a string-matching algorithm.[7]

Once the automatic step was concluded, the pool of retrieved sentence pairs underwent a thorough, manual inspection. Such a quality check was carried for each language direction in order to: *i)* remove any noise and keep only pairs containing at least one gender phenomenon and *ii)* examine the remaining pairs and include in MuST-SHE a balanced distribution of categories, FEM/MASC forms, and speakers; *iii*) verify and include overlapping pairs among en-it/fr/es to create a common multilingual subset.[8] Finally, we ensured that the final pairs were not affected by misalignments resulting from the automatic procedure used to create MuST-C and the new TED Talks data. Accordingly, to finalize the creation of MuST-SHE textual portion, we adjusted any identified <*transcript, translation*> misalignment. As a last step, all the corresponding audio segments were also manually checked in order to correct possible misalignments in the <*audio, transcript, translation*> triplets to be included in MuST-SHE.

4.3.2 Data Annotation

The resulting dataset was then manually enriched with different types of information that allow for fine-grained evaluations along different dimensions.

For each segment, the annotation includes: category (1 and 2), form (FEM and MASC), and speaker's gender information. Note that speaker's gender information has been labeled based on the personal pronouns the speakers used to describe themselves in their publicly available personal TED section.[9] Accordingly, such an assignment is meant to account for the gendered linguistic forms by which the speakers accept to be referred to in English, and would want their

[6] http://www.bls.gov/emp/tables/emp-by-detailed-occupation.htm

[7] We used PyLucene: https://lucene.apache.org/pylucene/.

[8] Segments were considered 'common' if they met the following conditions: *i*) there is a complete or partial overlap of the same English source sentece among – at least two of – the language directions represented in MuST-SHE, *ii)* across language directions, the gender phenomena represented in the overlapping segments are of the same category and gender form.

[9] https://www.ted.com/speakers

		Overall	Segments
He	speaker is referred to by masculine forms (he/his/him)	144	1696
She	speaker is referred to by feminine forms (she/her/hers)	143	1596
They	speaker is referred to by gender-neutral forms (they/their/them)	1	10
Multi-He	multiple speakers, all referred to by feminine gender forms (she/her)	2	10
Multi-She	multiple speakers, all referred to by masculine gender forms (he/his/him)	2	15
Multi-Mix	multiple speakers, all referred to by different gender forms	3	40

Table 4.1: MuST-SHE speakers' gender information, annotated based on *i)* the number of English speakers for each TED Talk, *ii)* the personal pronouns found in their TED bio. The last two columns show, respectively, speaker's gender distribution across the 295 unique speakers represented in MuST-SHE, and the distribution for each MuST-SHE segment.

translations to conform to.[10] Accordingly, such an information has also been employed to check the congruence between speakers' pronouns and the proper rendering of gender phenomena from Category 1.[11] We show speakers' gender annotations in Table 4.1.

To allow performing a sound evaluation able to discriminate gender-related issues from other non-related factors that may affect systems' performance, for each target gender-marked word in MuST-SHE (e.g., en: *he is nice* → it: *è* ***carino*** MASC), we created a corresponding gender-swapped counterpart in the opposite gender form (e.g., ***carina*** FEM). These word forms were paired and annotated in MuST-SHE reference translations (e.g., it: *è* ***<carino/carina>***), and served to obtain for each correct reference translation (C-REF) an almost identical – but "wrong" – alternative (W-REF), that differs from the original one only for the swapped gender-marked words. As we will describe in more detail in the upcoming Section 4.4, the double references and annotated target gender-marked words are key features of MuST-SHE, enabling both gender-sensitive evaluations at the sentence/corpus level and fine-grained analyses focusing solely on the correct generation of target gender-marked words. In Table 4.2, we show some selected annotated examples from the common multilingual subset of MuST-SHE segments.

I personally created and annotated the whole dataset, by also producing strict and comprehensive guidelines[12] based on the preliminary manual analysis of a sample of MuST-C data (2,500 segments). To ensure data quality, a second linguist independently re-annotated each MuST-SHE segment with the corresponding category and produced an additional "wrong" reference. Being the annotation per category a straightforward task, it resulted in no disagreement for both Category 1 and 2. Thus, the dataset contains only segments in complete agreement.

[10] The personal pronouns were retrieved in 2020. Being a static annotation, it might no longer reflect the preferences of TED speakers.

[11] Among all MuST-SHE speakers, we found one referred to by the gender-neutral *they* pronoun. As the reference translations for Category 1 phenomena referring to this speaker relied on MASC forms, we considered them as unmatching and thus excluded them from the corpus. We however included segments uttered by the speaker for Category 2 phenomena.

[12] Available at:
https://docs.google.com/document/d/1ZPAG-JxcqRzmZ8GP2W1FvDuHgTfRKlVIIMpRt5Jwk50/edit

Form		Category 1	Speaker
FEM	SRC	I was 24 years old when... downhill skiing **paralyzed** me.	She
	C-REF$_{It}$	Avevo 24 anni quando... un incidente sciistico mi ha lasciato **paralizzata**.	
	W-REF$_{It}$	Avevo 24 anni quando... un incidente sciistico mi ha lasciato **paralizzato**.	
	SRC	I was 24 years old when... downhill skiing **paralyzed** me.	
	C-REF$_{Fr}$	J'avais 24 ans quand... une descente à ski m'a **paralysée**.	
	W-REF$_{Fr}$	J'avais 24 ans quand... une descente à ski m'a **paralysé**.	
	SRC	I was 24 years old when... downhill skiing **paralyzed** me.	
	C-REF$_{Es}$	Tenía 24 años cuando... esquiando quedé **paralizada**.	
	W-REF$_{Es}$	Tenía 24 años cuando... esquiando quedé **paralizado**.	
MASC	SRC	I **myself** was **one** of them, and this is what I talk about at the events.	He
	C-REF$_{It}$	Io **stesso** ero **uno** di loro, e parlo di questo agli eventi.	
	W-REF$_{It}$	Io **stessa** ero **una** di loro, e parlo di questo agli eventi.	
	SRC	I myself was **one** of them, and this is what I talk about at the events.	
	C-REF$_{Fr}$	Moi-même, j'ai été l'**un** d'eux, et voilà de quoi je parle aux événements.	
	W-REF$_{Fr}$	Moi-même, j'ai été l'**une** d'eux, et voilà de quoi je parle aux événements.	
	SRC	I **myself** was **one** of them, and this is what I talk about at the events.	
	C-REF$_{Es}$	Yo **mismo** era **uno** de esos y esto fue de lo que hablé en los eventos.	
	W-REF$_{Es}$	Yo **misma** era **una** de esos y esto fue de lo que hablé en los eventos.	
		Category 2	
FEM	SRC	She'd get together with **her dearest friends these older** women...	He
	C-REF$_{It}$	Tornava per incontrare **le sue** più **care amiche**, **queste** signore **anziane**...	
	W-REF$_{It}$	Tornava per incontrare **i suoi** più **cari amici**, **questi** signore **anziani**...	
	SRC	She'd get together with her **dearest friends**, these **older** women...	
	C-REF$_{Fr}$	Elle se réunissait avec ses **amies** les plus **chères**, ces femmes plus **âgées**...	
	W-REF$_{Fr}$	Elle se réunissait avec ses **amis** les plus **chers**, ces femmes plus **âgés**...	
	SRC	She'd get together with her **dearest friends**, **these** older women...	
	C-REF$_{Es}$	Ella venía a reunirse con sus más **cercanas amigas**, **unas** señoras mayores...	
	W-REF$_{Es}$	Ella venía a reunirse con sus más **cercanos amigos**, **unos** señoras mayores...	
MASC	SRC	Dean Kamen, **one of the** great DIY **innovators**, his technology...	She
	C-REF$_{It}$	Dean Kamen, **uno dei** più grandi **innovatori** fai-da-te, la sua tecnologia...	
	W-REF$_{It}$	Dean Kamen, **una delle** più grandi **innovatrici** fai-da-te, la sua tecnologia...	
	SRC	Dean Kamen, **one** of the **great** DIY **innovators**, his technology...	
	C-REF$_{Fr}$	Dean Kamen, l'**un** des **grands innovateurs** autonomes, sa technologie...	
	W-REF$_{Fr}$	Dean Kamen, l'**une** des **grandes innovatrices** autonomes, sa technologie...	
	SRC	Dean Kamen, **one** of **the** great DIY **innovators**, his technology...	
	C-REF$_{Es}$	Dean Kamen, **uno** de **los** grandes **innovadores** del bricolaje, su tecnología...	
	W-REF$_{Es}$	Dean Kamen, **una** de **las** grandes **innovadoras** del bricolaje, su tecnología...	

Table 4.2: MuST-SHE annotated segments organized per category. For each example in en-it, en-fr, and en-es the Correct Reference Translation (C-REF) shows the realization of target gender-marked forms (MASC/FEM) corresponding to English gender-neutral words in the source (SRC). In the Wrong Reference Translation (W-REF), Italian, French, and Spanish gender-marked words are swapped to their opposite gender form. The last column of the table provides speaker's gender information.

Disagreements were more common for the gender-marked words' identification and swapping (i.e., in the "wrong" references), since the task requires producing subtle variations that can be hard to spot, also entailing the production of ungrammatical sequences for Category 2, e.g., *my mother* → C-Ref it: <*mia*> *madre*, W-Ref <*mio*> *madre*.

Inter-annotator agreement (IAA) was measured using the Dice coefficient (Dice, 1945), which is computed as:

$$Dice = 2C/(A + B) \tag{4.1}$$

where C is the number of items annotated by both annotators (i.e., swapped gender-marked words), whereas A and B are the number of items annotated by each annotator, respectively. Thus, we measure agreement as the overlap of gender-marked words swapped by both annotators, with respect to all the gender-marked words that have been swapped. Dice coefficient of 1 indicates that there is an overlap between the two annotations in all samples observed, while 0 indicates that there is no overlap at all (Dice, 1945). Accordingly, the agreement rate is of 0.89, which means that 89% of the annotated gender-marked words were recognized as such and swapped by both annotators. Disagreements, amounting to 11%, were all oversights and thus reconciled. Finally, the corpus was again manually checked by graduate students with a background in linguistics/translation studies. All students had an excellent English level, and native/excellent level in the MuST-SHE target language they were assigned. Overall, they found 17 issues (e.g., misalignment, overlooked gender-marked words), which we have fixed in MuST-SHE.

4.3.3 Dataset Statistics

	En-It			**En-Fr**			**En-Es**		
	Fem	Masc	*Tot.*	Fem	Masc	*Tot.*	Fem	Masc	*Tot.*
Cat. 1	278 (401)	282 (415)	*560*	315 (424)	292 (410)	*607*	284 (392)	287 (419)	*571*
Cat. 2	237 (497)	264 (629)	*501*	225 (474)	237 (515)	*462*	267 (558)	268 (608)	*535*
Tot.	*515 (898)*	*546 (1044)*		*540* (898)	*529* (925)		551 (950)	555 (1027)	
Total	**1,061** (1,942)			**1,074** (1,823)			**1,106** (1,977)		

Table 4.3: MuST-SHE statistics. En-it, en-fr, and en-es number of segments split into fem and masc phenomena and categories. In parentheses, the number of annotated gender-marked words per dimension.

Overall, MuST-SHE comprises 3,241 (*audio*, *transcript*, *translation*) triplets (1,061 for en-it, 1,040 for en-fr, and 1,106 for en-es) uttered by 275 different speakers. Instead, the multilingual subset consists of a total of 1,040 segments: of these, 591 segments overlap between two language directions, whereas the other 449 are common to all three MuST-SHE language directions.

As shown by the statistics in Table 4.3, the corpus presents a balanced distribution across *i)* masculine and feminine forms, and *ii)* gender phenomena per category. As we have seen in

Table 4.1, also speakers are substantially balanced across genders. The gender of the speaker (i.e., their linguistic expression of gender) and that of the referred entity in the utterance coincide in Category 1 (where the speakers always talk about themselves), while it differs in about 50% of the segments in Category 2 (where the speakers refer to other persons). Thus, MuST-SHE is precisely designed to: *i)* equally distribute gender references as well as speakers, and *ii)* allow for a sound and focused evaluation on the accuracy of gender translation. As such, it satisfies the parameters to be qualified as a GBET (Sun et al., 2019b), and represents the very first of its kind for both ST and MT created on natural data.[13]

Disclaimer. As a finale note, we inform the reader that the initial release of MuST-SHE was only available for two language pairs: English-Italian and English-French. To allow for more extensive cross-lingual comparison and due to time constraints, the English-Spanish portion of the dataset was only completed and released at a later time. For this reason, the experiments carried out on MuST-SHE in Chapters 5 and 6 are only presented for en-it and en-fr.

4.4 Gender-sensitive Evaluation Method

MT evaluation metrics are used to provide a global score about translation quality as a whole (see Sec. 2.3.4). Used as-is, their holistic nature hinders the precise evaluation of systems' behavior on an individual phenomenon, since variations of global scores are only a coarse and indicator of better/worse overall performance (Callison-Burch et al., 2006). Thus, this represents a limitation when examining specific aspects such as gender translation and bias.

Indeed, previous works on gender bias that relied on holistic measurements has stumbled upon such a limitation. As discussed in Section 3.4, in fact, the gains in BLEU scores (Papineni et al., 2002) obtained by prepending gender tags or other artificial antecedents to the input source, as in Vanmassenhove et al. (2018) and Moryossef et al. (2019) might be contingent on different words generated in the MT output, and thus cannot be assuredly ascribed to a better control of gender features (e.g., src: *But I was **glad** it was all over*; it-ref: *Ma ero **contenta** fosse tutto finito* → output1: *Ma ero **contento** d'averlo finito*, output2+tag: *Ma ero **contento** fosse tutto finito*). To overcome this problem, Moryossef et al. (2019) rely on morphological and syntactic analyzers to pinpoint if overall gains are due to gender-related improvements. Such a procedure, however, adds complexity to the evaluation protocol, and still does not unveil the imprint of gender translation on overall quality scores. Instead, our goal is to conciliate in a

[13] At the time of its creation, MuST-SHE was the very first multilingual GBET built on naturally occurring data. At the time of writing, the natural MT-GenEVAl test set (Currey et al., 2022) has also been released, and it draws on the double-reference evaluation features that I described in both the previous and upcoming paragraphs.

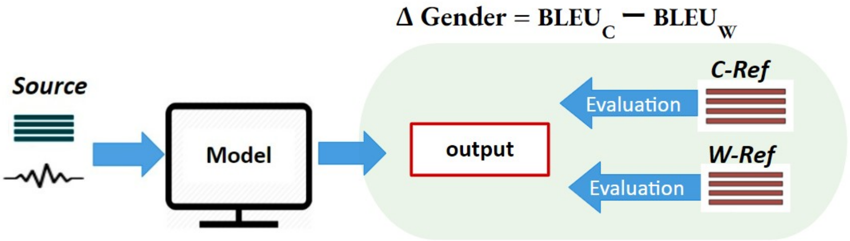


Figure 4.1: In the green background, the image shows the BLEU-based evaluation pipeline on MuST-SHE double references, given either an audio or textual input.

single reference-based evaluation protocol fine-grained analysis and overall quality assessment, but make global scores informative about systems' ability to produce the correct gender forms.

To this aim, as seen in the previous 4.3.2, for each "correct", MuST-C human reference c in the corpus we create a "wrong" one. Such a W-REF that is identical to c, except for the morphological gendered inflections. In particular, for each gender-neutral English word in the source utterance (e.g., *"one"*, *"great"* and *"innovators"* in the 4^{th} example of Table 4.2), the correct translation (containing the French words with MASC inflection *"un"*, *"grands"* and *"innovateurs"*) is swapped into its opposite gender form (containing FEM words *"une"*, *"grandes"* and *"innovatrices"*). The result is a new set of references that, compared to the correct ones, are "wrong" only with respect to the formal expression of gender. All (correct and swapped) target-gender marked words are annotated in the reference translations. On this basis, below we describe MuST-SHE evaluation protocol, which relies on *i)* gender-sensitive BLEU[14] scores (Papineni et al., 2002) and *ii)* gender accuracy computed on the single gender-marked words annotated in the references.

Gender-sensitive BLEU score. As the two reference sets in MuST-SHE differ only for the swapped gendered forms, the underlying idea is: results' differences for the same set of hypotheses produced by a given system can measure its capability to handle gender phenomena, hence isolate biased behaviors. In particular, we argue that higher values on the wrong set can signal a potentially gender-biased behaviour. In all the cases where the required gender realization is *feminine*, significantly higher BLEU results computed on the wrong set would signal a bias towards producing masculine forms, and vice versa. Thus, as shown in Figure 4.1, we compute a gender-sensitive BLEU score as the difference between the BLEU score obtained on the C-REF and the W-REF. Such a method draws on previous research relying on contrastive references for challenge datasets Sennrich (2017). Also, although this idea recalls the gender-swapping approach used in previous NLP studies on gender bias (Sun et al., 2019b; Lu et al., 2019; Kiritchenko and Mohammad, 2018; Zhao et al., 2018b; Cao and Daumé III, 2020), in such works

[14] We propose and rely on BLEU since it remains the *de facto* MT evaluation metrics. In principle, however, other metrics such as TER (Snover et al., 2006b) can be used, too.

it is only applied to pronouns; here we extend it to any gender-marked part of speech.

Gender accuracy. In addition to the quantitative BLEU-based evaluation, MuST-SHE also allows performing fine-grained qualitative analysis of systems' accuracy in producing the target gender-marked words. We compute accuracy as the proportion of gender-marked words in the references that are correctly translated by the system. An upper bound of one match for each gender-marked word is applied in order not to reward over-generated words by the system. Such is the notion of gender accuracy we have employed for the evaluation in Section 5.3.2.
However, as discussed in more detail in 5.4.3, this metric has some limitations. In fact, while its scores tell us the proportion of MuST-SHE annotated words that are correctly generated by the system, it remains unspecified if failures (i.e., unmatched MuST-SHE gender-marked words in the output) are due to _i)_ the generation of a different word (e.g., en: I have _been_, where the Italian output contains _andat*_ instead of <_stat*_>), or _ii)_ the generation of the wrong inflection, matched in the other reference (e.g., <_stato_> instead of <_stata_>). To overcome this limitation, from Section 5.4.3 onwards we rely on an improved definition of gender accuracy.[15]

It is worth remarking that the BLEU-based and the accuracy-based evaluations are complementary: the former aims to shed light on system's translation performance with respect to gender. Thus, it serves as a proxy for differential in the quality of systems offered across genders. The latter, which is more discriminative, is designed to exclusively focus on the actual gender-marked words, and thus unveil more precisely where feminine forms are underrepresented in the output. Compared to the standard BLEU-based evaluation with correct references only, we expect that the possible differences suggested by its extension with gender swapping will be reflected and amplified by sharper accuracy differences.

4.5 Conclusion

In this chapter, I have presented MuST-SHE, a multilingual, natural benchmark designed to evaluate and monitor gender bias in both MT and ST systems. First, I have described the cross-lingual, manual analysis that has informed its design and phenomena of interest. On this basis, I have outlined the creation and annotation process that underwent in the test set. Finally, I have explained its underlying, innovative evaluation method, which complements BLEU-based and focused accuracy-based evaluation protocols. MuST-SHE represents the first resource contribution of this thesis, and on which all the experiments in the upcoming experimental portion of this work rely.

[15] For adherence with the paper in which our first results on the MuST-SHE benchmark were published, in Chapter 5, Section 5.3.2, we used the first version of the gender accuracy metric. We have however verified that our findings and results are confirmed and consistent in both evaluation settings.

5

Gender Bias and Speech Translation: The role of speech cues

Gender bias in automatic translation emerges more prominently when translating across languages that express gender differently. Such an issue is also due to the asymmetries reflected in the training data on which models are built. On top of picking up on these imbalances, however, the generation of the correct gender form can also be hindered by the fact that the input sentence does not always provide cues about the gender of the referred human entities. Such is the case of ambiguous references to the speaker, as in "I am a glad". Being exclusively fed on *textual* data, MT systems are intrinsically constrained when confronted with such a type of input. But what happens with direct speech translation, where the input is an *audio* signal? Does audio provide additional information to support gender translation? If so, what are the implications of relying on audio cues leveraged from speakers' voice? We first explore these questions by means of a thorough investigation of gender bias in speech translation. Informed by our results and going beyond speech signals, we then explore different approaches to enrich direct ST models with external speakers' gender information and thus mitigate bias.

5.1 Introduction

Having introduced the MuST-SHE benchmark, we are now equipped to conduct thorough evaluations of gender bias for different translation models and experimental conditions. As a first step in this direction, we open a line of research focused on *speech* translation, the task of translating audio speech in one language into text in another language (2.3.3). In fact, as it emerged from our review in Chapter 3, inquiries on gender bias have been largely restricted to traditional text-to-text MT, and no systematic attempt has yet been made to investigate if and how ST technologies are impacted by this particular issue. More specifically, in this chapter, we foreground how gender bias interacts with the most distinctive feature of ST technology: namely, the role of audio input features as a variable for gender translation.

Following our MuST-SHE classification (see Section 4.2), we have identified different categories of phenomena, where the sentence-context either does (e.g., *She/He is a student*) or does not (e.g., *I am a student*) offer information for proper gender translation. For this latter ambiguous scenario, MT systems are bound to and constrained by their textual input, which does not offer any gender disambiguating information. The same applies to traditional *cascade* approaches to ST, which consist of a pipelined architecture where the source speech input is first transcribed by an ASR component, and then translated by an MT system, thus losing audio information (see Sec. 2.3.3). Conversely, *direct* ST systems that translate from audio to text without any intermediate representation have the potential to access audio cues such as speaker's voice, which might supply for the lack of contextual gender disambiguating information.

Accordingly, we start in Section 5.2 by outlining the relation between vocal cues and gender. In particular, we discuss fundamental frequency as a strong perceptual marker of gender, as well as its controversial implications for gender categorization. On this basis, in Section 5.3 we explore the following research questions (**RQs 1**): to what extent and how are ST technologies impacted by gender bias? Do direct ST systems leverage audio information to guide gender translation? We find that, while overall affected by a strong masculine skew, direct ST systems do leverage speech cues, thus showing an apparent advantage for speaker-related gender phenomena (e.g., *I am a student*). However, we claim that by relying on speakers' vocal features as a gender cue, these systems can be unsuitable or even harmful for a diverse range of users. For this reason, in Section 5.4, we investigate (**RQ2**): beyond speech signal, how can we control gender translation in direct ST by relying on external and validated gender information?

Overall, the main contributions of this chapter can be summarized as follows. **(1)** We present a systematic analysis aimed to assess ST performance on gender translation. To this aim, we compare the state-of-the-art cascaded approach with the emerging direct paradigm, investigating their ability to handle different categories of gender phenomena. **(2)** We created the MuST-Speakers resource.[1] It comprises the manual annotation of the TED talks data contained in MuST-C (Cattoni et al., 2021) with speakers' gender information based on speaker's personal pronouns. **(3)** We compare different approaches to integrate external knowledge about the speakers and mitigate gender bias in direct ST, also depending on the potential users, the available resources and the architectural implications of each choice.

5.2 The Gendered Voice

According to Karpf (2006), "the moment we open our mouth we leak information about our biological, psychological and social status." That is because the features of our voices, or *vocal cues*, are regarded as highly salient perceptual markers of multiple dimensions on which listeners rely to make inferences about us. Such dimensions can include age, origin, and of all the information gleaned from a person's voice, gender seems to be among the most prominent ones (Kreiman and Sidtis, 2011; Schweinberger et al., 2014; Leung et al., 2021). Indeed, the voice represents a highly salient cue in the perception of gender (Azul, 2015; Zimman, 2021), to the point that gendered differences in the voice represent a linchpin of much research in the area of phonetic and sociolinguistics (Zimman, 2018). The exploitation of these differences for engineering purposes, including automatic gender identification from speech features, has a long history dating back to the late 1990s (Parris and Carey, 1996; Vergin et al., 1996) and continues with current ML techniques (Levitan et al., 2016; Yusnita et al., 2017). But what are the characteristics that make our voices "gendered"? Or in other words, how does gender sound?

Despite a general tendency to treat the voice as a single dimension, speech pathologists and linguists examined which cues appear as gender-characterizing; or perhaps more interestingly, which are *perceived* as significant of masculine and feminine voices (Leung et al., 2021). For instance, differences driven by articulatory practices for sibilant consonants (i.e., the /s/ segment) and vowels – though less consciously salient to most speakers – are recognized as an area where women and men typically differ (Gelfer and Mikos, 2005; Pharao et al., 2014; Zimman, 2017; Li, 2017). But while such differences have been observed for a large variety of English dialects and other languages, there is evidence (Heffernan, 2004) that they have little to do with anatomical differences between (cisgender)men and (cisgender)women[2] (see Zimman (2021) for a review). The confrontation with physiological factors, however, is unavoidable when considering what is the most intuitive, widely recognized, and well-studied perceptual marker of gender in the

[1] The resource is released under a CC BY NC ND 4.0 International license, and is freely downloadable at `https://ict.fbk.eu/must-speakers/`.

[2] *Cisgender* refers to individuals whose gender identity corresponds with the sex the person was assigned at birth (Fuchs and Toda, 2010).

voice: pitch (Gelfer and Mikos, 2005). Here is where the physical and material reality in the production of sound really comes into play.

Pitch, whose primary acoustic measure is fundamental frequency (F0), roughly corresponds to the speed at which the vocal folds vibrate (Zimman, 2018, 2021). The vibration speed is dependent on several concrete factors, like the mass of the folds (Jacobs, 2017), where shorter, thicker folds will result in lower fundamental frequency, and vice versa (Larose, 2022). Hence, given that the F0 range a speaker can produce is constrained by the size of the vocal folds and that their size is sensitive to testosterone (Evans et al., 2008), pitch across cisgender men (avg. F0 100-120) and women (avg. F0 200-220) tend to differ substantially (Fitch and Holbrook, 1970; Simpson, 2009).[3] Since vocal traits do not significantly differ before puberty, these gendered differences in pitch tend to arise in adolescence.

While biology clearly plays a role in the gendered voice, it is also evident that sociocultural factors have a significant impact on the way in which men and women use pitch. In fact, while an individual's pitch range is constrained by their vocal anatomy, biology alone does not determine which part of that range a speaker will use. For example, the gap between the average pitch of men and women varies across languages, with Japanese-speaking women famously maintaining a higher pitch than English-speaking women, while Japanese-speaking men speak with a lower pitch than English-speaking men (Loveday, 1981; Yuasa, 2008). Additionally, studies on queer communities and transgender individuals provide great insights into the role of socialization factors for the articulation and perception of a voice as gendered, as well as on the wealth of control that individuals have over their own voice. For instance, studies of gay, lesbian, bisexual, and queer voices have made clear that not all men embody norms for 'male voices' and not all women exhibit the features attributed to 'female speakers' (Munson and Babel, 2007; Podesva et al., 2016), thus resulting in difficulties for speakers that attempt to simply place them along a straight male-female binary scale. Conversely, Strand (2000) provided evidence that stereotypical-sounding voices are processed faster than non-stereotypical-sounding voices.

Despite the fact that many people may not be aware of how they sound, the voice is a relevant aspect of gender presentation, in particular for those who are transitioning from one gender role or presentation to another (Zimman, 2018; Oates and Dacakis, 2015). Indeed, for trans and gender-diverse individuals, who can display voice and communication characteristics that are incongruent with their self-identified gender, such a mismatch can lead to difficulties. In the experience of Larose (2022): "when an individual is dressed in a fashion that is societally received as one of the binary genders, their speaking voice can immediately alter that perception. The voice has the ability to override any other gender referents such as clothing, stance, or

[3] Reported F0 values are for speakers of American English.

body [..]". This has created a demand among some trans people for speech therapy focused on feminizing (or, less frequently, masculinizing) the voice (Zimman, 2018). Concurrently, perceptual studies have identified the range of 150–165 Hz as a potential crossover point for trans women (Spencer, 1988; Gelfer and Schofield, 2000), though there are notable exceptions. Günzburger (1993) reports that certain trans women in her study were perceived as female despite having F0 means as low as 119–128 Hz. Inversely, Gelfer and Schofield (2000) report that speakers may be perceived as male even with an F0 as high as 181 Hz. Thus, while fundamental frequency is extremely salient as a gender cue, there are other factors that contribute to the perception of a voice as feminine or masculine.

Overall, existing inquiries on the relationship between voice and gender highlight that voice represents a crucial aspect of gender presentation. Physiological factors do influence our perception and categorization of feminine and masculine voices, though they should be filtered through socialization and cultural factors. Along the same line, it is worth underscoring how gender is not an objective, static feature inherent in a voice. Voices might change over time, and the interpretation of vocal cues can vary depending on the listener's cultural background. Accordingly, by asking "who might be speaking" we can also disclose information about the listeners, ourselves and our own gender's expectations (Larose, 2022). This can influence our use of vocal features as a variable, and their embedding in current technology. With this in mind, we can move onto the experimental portion of this chapter.

5.3 Gender and Speech Translation

In this section, we ask whether ST technology is affected by gender bias and to what extent, also by exploring the potential to exploit audio input signals to guide gender translation (**RQ1** in 5.1). Towards this goal, for two language pairs (en-it and en-fr) we carry out a systematic comparison of two current approaches to ST: *cascade* and *direct* (Sec. 2.3.3).

Cascade architectures represent the traditional approach to ST, which rely on the combination of an ASR component (trained on *audio-transcript* data) and MT component (trained on *transcript-translation* data) (Eck and Hori, 2005). The main advantage of this pipelined solution is that it can directly plug in state-of-the-art technology for both components and exploit the wealth of training data available for the two tasks. This approach, however, has some inherent drawbacks. One is error propagation: suboptimal transcriptions by the ASR component have significant impact on the final output produced by the MT component (see 2.3.3). Another drawback, of particular relevance for this work, comes from the loss of richer information – which cannot be grasped from the written modality – when passing from audio to text representations.

Differently, by avoiding intermediate text representations, direct translation from audio to text (Bérard et al., 2016) can access several speech dimensions, such as prodosic elements, including information about frequency. Due to the dearth of available training (*audio-translation*) corpora, the emerging direct still underperforms with respect to the cascaded counterpart. Here, however, rather than focusing on generic performance, we are interested in exploring whether this new architectural solution leads direct models to leverage speech signals from speaker's voice to translate gender.

Related work. To the best of my knowledge, no prior work has explored the issue that we hereby experiment with by means of focused comparison across ST systems. Rather, gender bias in ST constitutes a very small research niche that has been addressed in two other recent studies only. Namely, by Zanon Boito et al. (2022), who compared the overall performance of en-fr direct ST systems on male/female subsets of TED-based data. However, since they relied on holistic metrics of overall quality and generic test data, they did not detect any significant gender-related disparity, and did not actually inspect how these models handled gender translation phenomena in the output. Differently, the assessment of ST models carried out by Costa-jussà et al. (2022a) did inspect target gender-realization against the gender-sensitive ST-adapation of the WinoMT benchmark (see Sec. 3.4). By design, however, their study focuses on occupational terms and stereotyping in translation, but does not account for the role of speech input to examine gender bias.

5.3.1 Experimental Settings: *Cascade* vs. *Direct* Systems

Here below, we describe the architectures of the en-it/fr `Direct` and `Cascade` systems used in our experiments.

Our **`Direct`** system uses the S-transformer architecture, which has proved to work reasonably well for this task (Di Gangi et al., 2019d). As described in Section 2.3.3, the S-transformer is an encoder-decoder architecture that modifies the original Transformer architecture (Vaswani et al., 2017) in the encoder side by integrating a stack of convolutional neural networks (CNNs) (LeCun et al., 1998) to reduce the length of the audio input, which is pre-processed in the form of sequences of 40 Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980). In the implementation of Di Gangi et al. (2019c) that we hereby follow, the output of the CNNs in then processed by 2D self-attention networks (Dong et al., 2018), which model bidimensional dependencies along the time and frequency dimensions in the spectrogram. Also, a distance penalty is added to the non-normalized probabilities in the encoder self-attention networks in order to bias the computation towards the local context. As a countermeasure to data paucity and to improve translation quality, data augmentation techniques are applied. Accordingly, the

`Direct` systems are trained on the MuST-C and Librispeech (Kocabiyikoglu et al., 2018) corpora using SpecAugment (Park et al., 2019). Since Librispeech is a corpus for ASR consisting of only (*audio-transcript*) pairs, it was augmented by automatically translating the original English transcripts into both target languages with the MT systems integrated in the `Cascade` models. Translations are performed with character-level segmentation (see 2.3.2).

Our **`Cascade`** systems share the same core (ASR, MT) technology for both language pairs. The ASR component is based on the KALDI toolkit (Povey et al., 2011), featuring a time-delay neural network and lattice-free maximum mutual information discriminative sequence-training (Povey et al., 2016). The audio data for acoustic modeling include the clean portion of LibriSpeech (Panayotov et al., 2015) (~460h) and a variable subset of the MuST-C training set (~450h), from which 40 MFCCs per time frame were extracted; a MaxEnt language model (Alumäe and Kurimo, 2010) is estimated from the corresponding transcripts (~7M words). The MT component is based on the Transformer architecture, with parameters similar to those used in the original paper (Vaswani et al., 2017). The training data are collected from the OPUS repository,[4] resulting in 70M pairs for en-it and 120M for en-fr. For each language pair, the MT system is first trained on the OPUS data and then fine-tuned on MuST-C training data (~250K pairs). Byte pair encoding (BPE) (Sennrich et al., 2016) is applied to obtain 50K sub-word units. To mitigate error propagation and make the MT system more robust to ASR errors, similarly to Di Gangi et al. (2019b) we tune it on a dataset derived from MuST-C, which includes both human and automatic transcripts. First, the *audio-transcipt* training portion of MuST-C is is split in two equally-sized parts: the first one is used to adapt the ASR system to the TED talk language, while the second part is transcribed by the tuned ASR system. Then, the human transcripts of the first half and the automatic transcripts of the second half are concatenated and used together with their corresponding reference translations to fine-tune the MT system. This process makes the MT system aware of possible ASR errors and results in more than 2 BLEU points improvement on the MuST-C test set.

To check the overall quality of our cascade and direct systems, we compared them with existing results obtained on MuST-C test data that had been published at the time of these experiments. Our `Direct` systems (en-it: 21.5, en-fr: 31.0) outperform all the models proposed in Di Gangi et al. (2019d), which were trained only on MuST-C (en-it: direct 16.8, cascade 18.9; en-fr: direct 26.9, cascade 27.9). Our `Cascade` (en-it: 27.4 en-fr: 35.5) also outperforms the system described in Indurthi et al. (2019) (en-fr: 33.7). Our results are in line with the findings of IWSLT 2019 (Niehues et al., 2019), showing that the cascade approach at the time still outperformed the direct one, although with a gradually closing gap.

[4] `http://opus.nlpl.eu`

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	`Direct`	21.5	19.7	1.8	20.2	19.3	0.9	22.7	20.0	2.7
	`Cascade`	24.1	22.4	1.8	22.8	21.9	0.8	25.5	22.8	2.7
En-Fr	`Direct`	27.9	25.8	2.1	26.3	25.0	1.3	29.5	26.4	3.1
	`Cascade`	32.2	30.1	2.1	30.4	29.4	1.0	33.8	30.8	3.0

Table 5.1: BLEU scores for en-it and en-fr on MuST-SHE. Results are provided for the whole dataset (All) as well as split according to feminine and masculine word forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

5.3.2 Results and Discussion

To test our models with respect to gender translation, we rely on the English-Italian/French portion of the MuST-SHE benchmark (Chapter 4). Across the two language pairs, we follow the evaluation method described in Section 4.4, which computes gender-sensitive BLEU scores (Papineni et al., 2002) as the difference obtained on MuST-SHE double reference translations. Also, we compute the more fine-grained gender accuracy metric, which calculates the proportion of gender-marked target words in MuST-SHE that are correctly generated by the system. Besides global accuracy, we also calculate difference scores on both correct/wrong (i.e., gender-swapped) target words annotated MuST-SHE, across gender forms (FEM & MASC) and categories (1 & 2).

5.3.2.1 Gender-sensitive BLEU scores

In Table 5.1, we present translation results in terms of BLEU score[5] on the MuST-SHE dataset. Looking at overall translation quality (*All/Correct* column), the results on both language pairs show that the highest performance is achieved by `Cascade`, which are better than `Direct` by 2.6 points for en-it and 4.3 for en-fr. This is expected given the stronger performance of state-of-the-art cascade models, and it is in line with the results obtained on the standard MuST-C test set provided in the previous section. Besides highlighting a disparity in overall translation quality across cascade and direct ST approaches, however, these scores alone are not informative in terms of gender bias.

Differently, looking at the scores' gap between the *Correct* and the *Wrong* datasets (*All/Diff* column), it becomes evident that – despite the lower overall BLEU scores – for both language pairs `Direct` performs almost on par with `Cascade` as far as gender phenomena are concerned (1.8 on en-it and 2.1 on en-fr). The fact that the *All/Diff* values are always positive indicates that all the systems perform better on the *Correct* dataset (i.e., they generate the correct gender-marked words more often than the wrong ones). However, examining the results at the level of masculine/feminine word forms, we observe that *Diff* values are higher on the MASC subset (where the required gender realization is masculine) than in the FEM one (where the required gender

[5] We also computed TER, and the results are fully in line with the reported BLEU scores.

	Systems	**All**			**Feminine**			**Masculine**		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	`Direct`	43.3	16.4	26.9	34.2	24.0	10.2	51.3	9.6	41.7
	`Cascade`	41.1	17.5	23.6	33.7	24.5	9.2	47.6	11.2	36.4
En-Fr	`Direct`	46.0	19.0	27.0	35.8	25.0	13.8	55.3	13.8	41.5
	`Cascade`	49.6	20.5	29.1	39.6	26.2	13.4	58.7	15.2	43.5

Table 5.2: Accuracy scores for en-it and en-fr on MuST-SHE. Results are provided for the whole dataset (All) as well as split according to feminine and masculine word forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

realization is feminine). As discussed in Section 4.4, this signals a bias of the systems towards producing masculine forms to the detriment of feminine ones, which are under-represented.

Although our gender-swapping methodology allows us to measure differences across systems that cannot be observed with standard BLEU evaluations, the results obtained so far may still conceal further interesting differences. This can depend on the fact that BLEU works at the corpus level, whereas gender translation interests a small proportion of gender-marked words, even within a dedicated test set as MuST-SHE (~2,000 out of ~ 30,000 total words, avg. 1.8 per sentence). Accordingly, the evaluation of gender phenomena can have a limited influence on global measurements. To dig into these aspects, our final analysis relies on accuracy, which is exclusively focused on gender-marked words.

5.3.2.2 Gender Accuracy

In Table 5.2, we present the gender accuracy scores. As we can see, these results are not only consistent with the BLEU ones, but also highlight differences that were previously indistinguishable. While the *All/Diff* BLEU results for `Direct` and `Cascade` were identical on both languages, the *All/Diff* accuracy scores show cross-lingual differences. Namely, although `Direct` performs better than `Cascade` for en-it, it performs worse for en-fr. Still, by comparing the accuracy scores for FEM and MASC forms, the larger *Diff* values obtained on the MASC set indeed confirm systems' gender bias for both `Cascade` and `Direct` architectures.

With this in mind, we now move on to focus on systems' results for the two categories represented in MuST-SHE (see Section 4.2.1). Namely, Category 1, with contextually ambiguous first-person references where the audio input might offer a gender cue for disambiguation (TED speakers talking about themselves, e.g., *I am a friend*); and Category 2, where such information occurs in the utterance content (TED speakers talking about someone else, e.g., *she/he is a friend*). The results across categories are shown in Table 5.3.

As for **Category 1**, *Diff* values show that `Cascade` performance is the worst on both languages. We attribute worse performance on this category to the MT component integrated

	En-it						**En-Fr**					
	Feminine			**Masculine**			**Feminine**			**Masculine**		
	Direct						**Direct**					
	Corr.	Wro.	Diff.	Corr.	Wro.	Diff.	Corr.	Wro.	Diff.	Corr.	Wro.	Diff.
Cat. 1	26.7	27.2	-0.5	46.3	6.8	39.5	25.4	29.5	-4.1	48.0	7.7	40.3
Cat. 2	40.6	20.5	20.1	53.9	10.9	43.0	45.0	20.3	24.7	60.0	17.6	42.4
	Cascade						**Cascade**					
Cat. 1	15.9	34.5	-18.6	40.0	12.0	28.0	20.4	37.5	-17.1	49.1	13.0	36.1
Cat. 2	48.9	15.7	33.2	51.2	10.8	40.4	56.3	15.6	40.7	64.9	16.7	48.2

Table 5.3: Accuracy scores for en-it and en-fr split according to MuST-SHE categories (Cat 1: information in audio, Cat 2: information in utterance content). For each category, results are further split into masculine/feminine forms. Results are calculated for both the *Correct* and *Wrong* datasets, and their difference is provided (Diff).

within cascade models, whose pipelined architecture does not allow to access speaker's gender information from the audio input, which could guide models towards a correct translation. This becomes particularly evident in the feminine class, where `Cascade` systems achieve extremely higher values on the *Wrong* datasets, leading to negative values in columns *Feminine/Diff* (en-it: -18.6; en-fr: -17.1). This means that the models are producing the wrong MASC forms more often than the expected FEM ones, thus signalling a particularly strong bias.

Although the *Diff* values obtained by `Direct` models for the feminine class are still negative (en-it: -0.5; en-fr: -4.1), they are comparatively more promising than those obtained by their `Cascade` counterparts. In light of this, we can see how `Direct` systems prove to leverage and exploit gender cues from speakers' voice in the audio input to guide gender translation. Thus, while still being affected by a bias towards masculine forms and despite their overall lower translation quality – direct approaches seemingly have an advantage when it comes to speaker-related gender phenomena.

In **Category 2**, where having direct access to the audio is not relevant since gender cues are present in the textual transcript, results reveal a different scenario. While the scores obtained on the MASC class are not conclusive across languages, on the FEM class `Direct` always shows the worst performance. A possible explanation for this behavior is that, being trained on a small fraction of the data used by the cascade systems, `Direct` is intrinsically weaker and more prone to gender mistranslations, i.e., less capable of retrieving linguistic gender cues from the sentence context to produce the correct target gender realization. Also, this might be due to MuST-SHE design choice (see Section 4.3.3) to include ~50% of Category 2 segments where the speaker's linguistic expression of gender (e.g., *He*) does not overlap with the gender of the phenomenon to translate (e.g., FEM). This feature makes MuST-SHE particularly challenging for systems like `Direct` since, in these specific cases, gender cues extracted from the source audio are not only irrelevant, but can also introduce noise.

All in all, we can confirm that gender translation is problematic in ST, and that all current technologies are affected by gender bias, though to a variable extent. Through the analysis made possible by MuST-SHE, we have been able to pinpoint systems' specific "strengths" and "weaknesses" so as to inform future studies on gender bias.

5.3.3 Summary

ST technologies are increasingly deployed to translate speech in one language into text into another language. While the progress of their underlying cascade and direct architectures is steadily monitored in terms of generic performances, little has been done to assess the impact of gender bias on these technologies.

In this section, we have conducted a systematic analysis on gender translation by comparing the state-of-the-art ST cascaded approach with the emerging direct ST paradigm. To this aim, we have relied on the MuST-SHE benchmark, and for two language pairs (en-it and en-fr), we have inspected gender translation for different categories of gender phenomena. More precisely, we were guided by the hypothesis that, by translating from audio input without intermediate representation, direct systems had the potential to leverage the speaker's voice as a gender cue to generate the correct target form. For instance, as in the case of the contextually ambiguous utterance, (e.g., *I am a student*), where correct gender translation in Italian and French depends on the speaker's expression of gender.

Our evaluation shows that, while both ST approaches are affected by a strong masculine bias, direct ST systems show the ability to leverage audio information, thus resulting in a better handling of speaker-related phenomena. On the one hand, such an ability grants direct systems an apparent advantage with respect to their cascade counterparts and MT, where audio information is out of reach and the translation of ambiguous first-person references is prohibitive. On the other, however, we believe that relying on speaker's vocal features like fundamental frequency can be harmful and unsuitable for a diverse range of users (see Sec. 5.2). In light of these findings and considerations, we are motivated to further explore speaker-dependent gender phenomena in direct ST, and experiment on how to inform gender translation to mitigate bias.

5.4 Breeding Gender-aware direct ST

As seen in the previous section, by translating speech audio data without intermediate transcription, direct ST models are able to leverage input information (i.e., infer speakers' gender from their vocal characteristics), which is otherwise lost in the cascade framework. In spite of this potential for disambiguation, however, direct ST is nonetheless affected by gender bias just like its cascade counterpart and most NLP applications. As previously analyzed in 3.3, the language data that are collected to build the corpora on which NLP models are trained are often far from being homogeneous and rarely offer a fair representation of different demographic groups and their linguistic behaviors (Bender and Friedman, 2018). Consequently, as predictive models learn from the data distribution they have seen, ST models still tend to favor the demographic group most represented in their training data (Hovy and Spruit, 2016; Shah et al., 2020), hence better modeling men and masculine linguistic forms. Moreover, direct ST systems that exclusively rely on vocal biometric features as a gender cue can be unsuitable and potentially harmful for certain users. Indeed, as more thoroughly discussed in section 5.2, gendered differences in the voice tend to more strongly emerge after puberty, thus hardly capturing differences in speech among children. Still, not all speakers might conform to the canonical features attributed to feminine or masculine, whereas very prominent perceptual cues of gender such as pitch might lead to trans-exclusive gender categorizations (Zimman, 2021).

In light of the above, and considering that ST applications have entered widespread societal use, we believe that more effort should be put into further investigating and controlling gender translation in direct ST. In particular, going beyond speech signals for scenarios where the gender of the speaker is known in advance. To this aim, we manually annotated MuST-C with speakers' gender information and used them for experiments reflecting different possible real-world scenarios. We compare our gender-aware models against a "pure" model that solely relies on the speakers' vocal characteristics for gender disambiguation and test their ability to override speech signals as gender cues.

5.4.1 MuST-C Talks Annotation

Although current research on gender bias in ST can count on the MuST-SHE benchmark for fine-grained evaluations, large amounts of gender-annotated data are not yet available for development. This hinders the creation of gender-aware systems via mitigating techniques. Also, it limits the scope of research to application scenarios based on 'pure' models that leverage vocal features as cues to infer the speaker's gender. These scenarios can be intrinsically problematic, since someone's voice is not necessarily indicative of the gendered forms expected in translation. Rather, as discussed in Section 5.2, speakers might exhibit voice and communication patterns that are incongruent with their self-identified gender, or do not reflect the features that are attributed to and perceived as matching the canonically expected 'female' or 'male' voice (Azul, 2015; Zimman, 2018). Thus, such a mismatch can imply the wrong attribution of gender and lead to actual misgendering (McLemore, 2015) in translation.

In light of the above, building large training corpora explicitly annotated with validated

	Talks M	Talks F	Hours M	Hours F	Segments M	Segments F
en-it	1,546	715	313	134	176,972	70,755
en-fr	1,601	791	324	149	187,832	80,358

Table 5.4: Statistics for MuST-C training data with gender annotation. We show the number of talks, as well as the total number of segments and the corresponding hours of speech for each language pair. Differences in the number of annotated data for en-it and en-fr reflect their original distribution in the MuST-C corpus.

gender information becomes crucial. To this aim, rather than building a new resource from scratch, we opted for adding an annotation layer to MuST-C, which has been chosen over other existing corpora (Iranzo-Sánchez et al., 2020) for the following reasons: *i)* it is currently the largest freely available multilingual corpus for ST, *ii)* being based on TED talks it is the most compatible one with MuST-SHE, *iii)* TED speakers' personal information is publicly available and retrievable on the TED official website.[6]

To annotate the MuST-C corpus with speaker's gender information, we have employed the same annotation process and scheme used in MUST-SHE (see Section 4.3.2). Namely, following the MuST-C talk IDs, we have been able to *i)* automatically retrieve the speakers' name, *ii*) find their associated TED official page, and *iii*) manually label the personal pronouns used in their descriptions. Though time-consuming when applied to large amounts of (training) data, we opted for such a manual retrieval of information over automatic approaches to speaker gender identifications. The reasons for this choice partially overlap with those that move us towards systems that do not attempt to infer gender solely from vocal features. First, since automatic methods based on fundamental frequency are not equally accurate across demographic groups (e.g., women and children are hard to distinguish as their fundamental frequency is typically high (Levitan et al., 2016)), manual assignment prevents from incorporating gender misclassifications in our training data. Second, we avoid biological essentialist framework that assess gender based on vocal cues (Sec. 5.2), as they risk engraining stereotypical expectations about how feminine/masculine voices ought to sound, and can be particularly harmful for transgender and gender-diverse individuals (Oates and Dacakis, 2015; Zimman, 2020)

Differently, following the guidelines on gender as variable in Larson (2017), we do not want to run the risk of making assumptions about speakers' gender identity and introducing additional bias within an environment that has been specifically designed to inspect gender bias. By looking at the personal pronouns used by the speakers to describe themselves, our manual assignment instead is meant to account for the gender linguistic forms by which the speakers accept to be referred to in English (GLAAD, 2007), and would want their translations to conform to. Indeed, this is the only accessible and useful piece of information we need to inform gender translation.

[6] Available at https://www.ted.com/speakers/

MuST-Speakers. On this basis, we created MuST-Speakers, a new resource that comprises the annotation of MuST-C talks with speaker's gender information based on their personal pronouns. Such a resource covers all the English TED talks included in MuST-C *i)* training, *ii)* development, and *iii)* tst-common sets, and whose translations are available for Italian, French, and or Spanish. This amounts to a total of 2,545 English talks annotated with their corresponding speakers, who are all referred to by binary gender forms (*he*/*she*), except for one case (*they*).[7] Focusing on (meta)data needed to enhance ST systems for the two language pairs of our interest, in Table 5.4 we present the statistics for en-it/fr MuST-C annotated *training* data. These, are the data that will be used to train our gender-aware systems in following section 5.4.2. As we can see from the table, for both language pairs the distribution of female/male speakers[8] is highly unbalanced. Indeed, talks with female speakers represent only the ∼30%, thus unfortunately confirming the feminine under-representation that characterizes the language data used to develop translation models.

Gender-Balanced Validation Set. Being a sub-partition of the MuST-C data used for development, MuST-C standard *validation set*[9] reflects the same gender-imbalanced distribution found in the training set. Therefore, towards the development of gender-aware systems, we created a new, specifically designed validation set composed of 20 talks. Unlike the standard MuST-C validation set, it contains a balanced number of female/male speakers so to discourage rewarding models' potentially biased behaviour to be learned at training time.[10] Our models in Sec. 5.4.2 exploit this balanced dev set.

Overall, we enable research on gender bias in ST by means of two dedicated resources – MuST-Speakers and the gender-balanced validation set – paired with MuST-SHE.[11]

[7] Since our experiments focus on binary linguistic forms, and given the very limited representation of a single *they* speaker, such an instance was not used for training. By design, some sparse talks with multiple speakers of different genders (*Multi-mix*) were also excluded. Detailed information about all MuST-C speakers and corresponding talks can be found in the resource release and Data Statement (Bender and Friedman, 2018) at `ict.fbk.eu/must-speakers`.

[8] Some authors distinguish female/male for sex and woman/man for gender (among others (Larson, 2017)). For the sake of simplicity, in our study we use female/male to respectively indicate those speakers whose personal pronouns are *she/he*.

[9] As described in Section 2.3.2, the validation – or development – set serves to monitor the model fit on the training data as a sort of intermediate test during the development phase.

[10] This new resource is released under a CC BY NC ND 4.0 International license, and is freely downloadable at `https://ict.fbk.eu/must-c-gender-dev-set/`.

[11] To ease future research, all three resources are available and released for en-it, en-fr and en-es. The experiments presented in this chapter, however, have been carried out for en-it/fr only.

5.4.2 ST Systems

For our experiments, we built three types of direct systems. One is the `Base` system, a state-of-the-art 'pure' model that does not leverage any external information about speaker's gender (Sec. 5.4.2.1). The others are two gender-aware systems that exploit speakers' gender information in different ways: `Multi-gender` (Sec. 5.4.2.2) and `Specialized` (Sec. 5.4.2.3). All the models share the same architecture, a Transformer (Vaswani et al., 2017) adapted to ST. Their architecture slightly differs from the one used in the experiments in the previous Section 5.3.1, as it follows the latest, competitive solution proposed by Gaido et al. (2020b) at the IWSLT-2020 evaluation campaign.[12] Within this ST-adapation of the Transformer, the encoder processes the audiosignal – preprocessed and given as input as a Mel-filter-bank sequences (see Sec. 2.3.3) – with two 2D convolutional layers with stride 2, returning a sequence that is four times shorter than the original input. The vectors of this sequence are projected by a linear transformation into the dimensional space used in the following encoder Transformer layers and are summed with sinusoidal positional embeddings. The attentions in the encoder layers are biased toward elements close on the time dimension with a logarithmic distance penalty (Di Gangi et al., 2019c). The decoder architecture, instead, is not modified.

5.4.2.1 Base ST Model

We are interested in evaluating and improving gender translation on strong ST models that can be used in real-world contexts. As such, our `Base`, gender-unaware model is trained with the goal of achieving state-of-the-art performance on the ST task. To this aim, we rely on data augmentation and knowledge transfer techniques – previously described in Section 2.3.3 – that were shown to yield competitive models at the IWSLT-2020 evaluation campaign (Ansari et al., 2020; Potapczyk and Przybysz, 2020; Gaido et al., 2020b). In particular, we use three data augmentation methods: SpecAugment (Park et al., 2019), time stretch (Nguyen et al., 2020), and synthetic data generation (Jia et al., 2019). Also, as described below, we rely on transfer knowledge both from ASR and MT through component initialization and knowledge distillation (Hinton et al., 2015).

The ST model's encoder is initialized with the encoder of an English ASR model (Bansal et al., 2019) with a lower number of encoder layers (the missing layers are initialized randomly, as well as the decoder). This ASR model is trained on Librispeech (Panayotov et al., 2015), Mozilla Common Voice,[13] How2 (Sanabria et al., 2018), TEDLIUM-v3 (Hernandez et al., 2018), and the English utterance-transcript pairs of the ST corpora Europarl-ST (Iranzo-Sánchez et al., 2020) and MuST-C. These datasets are either gender unbalanced or do not provide speaker's gender

[12] The 2020 campaign is concurrent to the time of our hereby proposed experiments and models.

[13] https://voice.mozilla.org/

information; the only exception being Librispeech, which is balanced in terms of female/male speakers (Garnerin et al., 2020). However, since these speakers are just book narrators, first-person sentences do not really refer to the speakers themselves.

Knowledge distillation (KD) is performed from a *teacher* MT model by optimizing the KL divergence[14] between the distribution produced by the teacher and by the *student* ST model being trained (Liu et al., 2019). For both en-it and en-fr, the teacher MT model is trained on the OPUS datasets (Tiedemann, 2016).

The final `Base` ST model is trained in three consecutive steps. In the first step, we use the synthetic data obtained by pairing ASR audio samples with the automatic translations of the corresponding transcripts. In the second step, the model is trained on the ST corpora (Cattoni et al., 2021; Iranzo-Sánchez et al., 2020). In these first two steps, we use the KD loss function. Finally, in the third step, the model is fine-tuned on the same ST corpora using label-smoothed cross entropy (Szegedy et al., 2016). SpecAugment and time stretch are used in all steps.

5.4.2.2 Multi-gender Systems

Our first type of gender-aware systems are the `Multigender` models, which are created with the intent of informing models about speakers' gender. This gender information has to be artificially injected in the model, so to enhance it and control gender translation.

The idea of "multi-gender" models was inspired by one-to-many multilingual neural MT systems (Johnson et al., 2017), in which a single model is trained to translate from a source into many target languages by means of a *target-forcing* mechanism. This approach entails the introduction of an artificial token that is prepended to the source sentence to specify in which target language the translation is required, hence target-forcing. Here, we adapt this mechanism for "*gender-forcing*". Accordingly, ST `Multi-gender` systems are fed not only with the input audio, but also with a tag (*token*) representing the speaker's gender: <F> or <M>. This *token* is converted into a vector through learnable embeddings. The main idea behind this approach is that such a token, provided a both training and inference time, will lead models to associate <F> and <M> tags with their corresponding FEM and MASC gender forms in the language data and generate them accordingly.

Note that this tag-method behind our `Multi-gender` systems was already introduced in the context of gender bias by Vanmassenhove et al. (2018) and (Elaraby et al., 2018) (see "gender tagging" in Section 3.5), where models were informed about the speaker's gender with a tag prepended to the source sentence. To decide how to incorporate gender information about

[14] The KL divergence is a measure of the distance between two probability distributions.

the speakers in our ST models, however, we do not follow their approach, as it is specifically dedicated to MT. Rather, we consider the incorporating options that obtained the best results in multilingual direct ST (Di Gangi et al., 2019e; Inaguma et al., 2019), namely:

- **Decoder prepending.** The gender token replaces the <\s> (*EOS*, end-of-sentence)[15] that is added in front of the generated tokens in the decoder input.
- **Decoder merge.** The gender embedding is added to all the word embeddings representing the generated tokens in the decoder input.
- **Encoder merge.** The gender embedding is added to the Mel-filter-bank sequence representing the source speech given as input to the encoder.

In all cases, multi-gender models' weights are initialized with those of the `Base` models. The only randomly-initialized parameters are those of the gender embeddings.

Overall, the design of `Multi-gender` models has two main potential advantages: *i)* the creation of a single model that supports both male and female speakers (which makes it particularly appealing for real-world application scenarios). Also, by training a single composite model *ii*) each gender direction (feminine and a masculine) can benefit from the data available for the other at training time, potentially learning and leveraging words encountered in the masculine set that are not available in the feminine one, and vice versa (*transfer learning*). Differently, separate training based on two subsets (feminine vs. masculine) would entail that models are trained on a reduced amount of available data.

5.4.2.3 Gender-specialized Systems

The second type of gender-aware systems that we build for these experiments are the `Specialized` models. In this approach, for each language pair, two different gender-specific models are created. Each model is initialized with the `Base` model's weights and then fine-tuned only on samples of the corresponding speaker's gender. That is, it is fine-tuned on the data available for female or male speakers, only.

The idea behind this approach is that, by being exposed to a dedicated set of FEM and MASC data only for fine-tuning, each model will improve on the generation of each specific gender form. Hence, the model fine-tuned of feminine data will be used for female speakers only, and vice versa. Note that this fine-tuning approach somewhat resembles the mitigating method for MT based on gender-balanced training data used by Costa-jussà and de Jorge (2020) (see

[15] Logically, this should be the beginning of the sentence. We however use the EOS token to indicate both beginning and end of sentence.

“balanced fine-tuning" in Section 3.5). Their intent, however, is that of reducing stereotyping (for unambiguous cases of gender translation) by training a single model with more even gender representations. Their goal and scenario is thus different from ours, as we work with gender-ambiguous inputs (e.g., *I am a student*), thus requiring a dedicated model depending on the gender of the speaker.

Compared to the multi-gender architecture, the `Specialized` solution has the drawback of a higher maintenance burden as it requires the training and management of two separate models. Moreover, no transfer learning is possible: although each model is initialized with the base model trained on all the data and the low learning rate used in the fine-tuning prevents catastrophic forgetting (Mccloskey and Cohen, 1989),[16] data scarcity conditions for a specific gender are likely to lead to lower performance on that direction.

5.4.2.4 Experiments

As just described in Sec. 5.4.2.1, our ST models adopt knowledge transfer techniques that showed to significantly improve ST performance. In particular, knowledge distillation (KD) is especially relevant as it allows the ST model to learn and exploit the wealth of training data available for MT, which otherwise would not be accessible. According to the studies of Gaido et al. (2022); Vamvas and Sennrich (2021), however, such an architectural choice might come with drawbacks. Namely, distilling knowledge from (higher performing) MT systems appear to negatively impact gender translation by overgeneralizing towards masculine forms and stereotyping. In light of this, we are interested in monitoring the progressive effect of KD and our three-step ST development on overall quality and gender bias. Hence, in our experiments we include: *i)* the teacher MT models, *ii)* the intermediate ST models trained on KD, and *iii)* the final `Base` ST models obtained with fine-tuning without KD. All those, of course, are compared with our gender-aware ST solutions.

Both `Multi-gender` (Sec. 5.4.2.2) and `Specialized` models (Sec. 5.4.2.3) are initialized with the weights of the `Base` model are then fine-tuned on the MuST-C gender-labeled dataset. As seen in Table 5.4, this training set shows a skewed male/female speaker distribution, amounting to around 30%/70% in terms of samples. Thus, we tested both approaches in two different data conditions: *i)* a balanced condition (*-Bal), where we use all the female data available together with the same amount of a random subset of the male data, and *ii)* an unbalanced condition (*-All) where all the MuST-C data available are exploited. It must be noted that there are differences between the two approaches on the usage of data. In the `Specialized` approach,

[16] Catastrophic forgetting refers to the tendency of neural systems to “forget” the information previously learned (here, from `Base`) upon learning new information (here, from the fine-tuning step).

since we have two separate systems, the one which is fine-tuned with talks by female speakers remains the same in both data conditions. Differently, in the multi-gender approach, where a single model is trained on both genders together, the same number samples for <F> and <M> gender must be provided for training. Thus, when all MuST-C data are used, the female gender pairs – which are underrepresented – are over-sampled.[17]

5.4.3 Evaluation Method: *Coverage* and *Accuracy*

Also for these experiments, we rely on the en-it/fr portion of the gender-sensitive MuST-SHE benchmark. Indeed, this test set, which consists of of (*audio, transcript, translation)* aligned triplets, can be used to assess both MT and ST systems. However, since we are interested in the evaluation of our gender-aware models for speaker-dependent gender phenomena, we do not exploit the whole MuST-SHE corpus. We rather restrict our focus to its Category 1 subset, which – for each language pairs – represents ~600 segments where gender agreement only depends on the speaker's gender (see Section 4.2.1 and 4.3.3). For each of such segments, we can rely on the speakers' gender information, which is annotated as in MuST-Speakers.

As described in Section 4.4, a key feature of MuST-SHE is that, each target gender-marked word is annotated together with its swapped gender, so to obtain for each reference translation an alternative "wrong" reference (e.g., *I have been* uttered by a female speaker is translated into the correct Italian reference *Sono <stat**a**>* FEM, and into the wrong reference *Sono <stat**o**>* MASC). The idea behind gender-swapping is that the *Difference* between the scores computed against the "correct" and the "wrong" set of gender-marked words captures the system's ability to handle gender translation. Indeed, as proved with the results in the previous set of experiments in Section 5.3.2, such difference values are extremely revealing, but we have come to find that they overlook a crucial aspect. Namely, these scores do not allow distinguishing between those cases where the system "fails" by producing a word different from the one annotated in the references (e.g., *andat** in place of *<stat*>*) and failures specifically due to the wrong realization of gender (e.g., *stat**o*** in place of *<stat**a**>*).

Thus, while still following the gender-swapping principle, we now introduce and rely on a more informative evaluation, which adapts our previous *gender accuracy* measure, and pair it with the new *term coverage* metric.

- First, we calculate the **term coverage** as the proportion of gender-marked words annotated in MuST-SHE that are actually generated by the system, on which the accuracy of gender

[17] Oversampling is a technique employed to adjust the distribution of different classes in a dataset and to compensate for an imbalance. In this case, simply put, data from female speakers are replicated, and ST systems iterate over this same set more than once while, instead, progressively seeing new male data.

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 new evaluation method[18] 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 bias. The new Gender accuracy has the further advantage of informing about the margins for improvement of the systems.

5.4.4 Results and Discussion

Our automatic evaluation comprises the discussions of overall (5.4.4.1) and cross-gender (5.4.4.2) results. Also, we analyze the experimental scenario where speaker's gender and voice provide conflicting information (5.4.4.3).

5.4.4.1 Overall Results

Table 5.5 presents overall results in terms of BLEU scores on the MuST-SHE test set, which are computed against the C-REF. Despite the well-known differences in performance between en-it and en-fr, both language directions show the same trend.

	en-it	**en-fr**
	BLEU	BLEU
MT for KD	**33.59**	**39.61**
Base-KD-only	23.58	31.97
Base	27.51	34.25
Multi-DecPrep-Bal	26.36	33.54
Multi-DecPrep-All	26.17	34.13
Multi-EncMerge-Bal	26.47	33.29
Multi-EncMerge-All	26.39	33.07
Multi-DecMerge-Bal	21.99	27.06
Multi-DecMerge-All	22.12	27.74
Specialized-Bal	27.43	34.32
Specialized-All	27.79	34.61

Table 5.5: BLEU scores on MuST-SHE.

First, the MT systems used by the ST models for KD achieve by far the highest performance. This is expected since the ST task is more complex and MT models are trained on larger amounts

[18] The evaluation script is publicly available with the MuST-SHE release at https://ict.fbk.eu/must-she/.

	en-it		en-fr	
	Cover.	Acc.	Cover.	Acc.
MT for KD	**63.83**	51.45	**63.10**	52.08
Base-KD-only	56.05	51.76	59.17	53.12
Base	56.17	56.26	62.02	56.24
Multi-DecPrep-Bal	56.91	64.86	60.95	69.34
Multi-DecPrep-All	56.54	66.81	61.31	70.29
Multi-EncMerge-Bal	57.04	62.55	60.60	62.67
Multi-EncMerge-All	57.65	60.39	62.38	61.83
Multi-DecMerge-Bal	49.88	59.41	54.52	64.63
Multi-DecMerge-All	50.74	60.58	56.31	65.96
Specialized-Bal	57.90	**86.35**	61.79	**86.13**
Specialized-All	58.02	**87.02**	62.38	**86.45**

Table 5.6: Term coverage and gender accuracy scores on MuST-SHE.

of data. However, all our ST results are competitive compared to those published for the two target languages. In particular, on the standard MuST-C test set, the scores of our ST `Base` models are 27.7 (en-it) and 40.3 (en-fr), respectively 0.3 and 4.8 BLEU points above the best *cascade* results reported in Section 5.3.1.

Moving on to the ST systems, we attest that the models after the first two training steps based on KD (`Base-KD-only`, see 5.4.2.1) achieve a lower translation quality than the `Base` models, thus showing that the third training step is crucial to boost overall performance. In general, except for the `Multi-DecMerge` system (whose performance is significantly lower), we do not observe statistically significant differences between the `Base` models and their gender-aware extensions (`Multi-*` and `Specialized-*`), which also perform on par when fine-tuned with varying amounts of annotated data (balanced vs. all). Accordingly, due to the very small percentage of speaker-dependent gender-marked words in MuST-SHE (< 3%, 810-840 over ~30,000 words), systems' ability to translate gender is not reflected by BLUE scores.

Now, we delve deeper into our more informative evaluation (as per Sec. 5.4.3) and turn to the term coverage and gender accuracy values presented in Table 5.6. The overall results assessed with BLEU are confirmed by **term coverage** scores for both en-it and en-fr: the MT systems generate the highest number of annotated words present in MuST-SHE (63.83% on en-it and 63.10% on en-fr), while we do not observe large differences among the ST models (between 56.17% and 58.02% for en-it and 60.60% and 62.38% for en-fr).

Instead, looking at **gender accuracy**, we immediately unveil that overall performance is not an indicator of the systems' ability to translate gender. In fact, the best performing `MT` systems show the lowest gender accuracy (51.45% for en-it and 52.08% for en-fr): intrinsically constrained by the lack of access to audio information, they produce the wrong target gender in half of the cases. Such a trend is directly reflected in the `Base-KD-only` models, which exploit as much as possible MT knowledge and are strongly influenced by the MT behaviour. Thus, although effective for overall quality, KD appears detrimental to gender translation within this speaker-dependent scenario. By undergoing the third training step without KD (i.e., fine-tuning on ST data from MuST-C), the `Base` models are in fact able to improve on gender translation, although with limited gains. Differently – but as expected – the gender-aware models fed with the speaker's gender information display a noticeable gender translation improvement, with `Specialized-*` models outperforming the `Multi-*` ones by 16–20 points and the `Base` ones by 30 points.

Among the multi-gender architectures, our results show that `Multi-DecPrep` approach has an edge on the other two models, both in overall and gender translation performance: for the sake of simplicity, from now on we thus ony present that model. As a single-model architecture, the multi-gender design would make a more functional solution with respect to specialized models. However, by being trained on both female and male speakers' utterances, it is noticeably weaker than multiple specialized models (trained on gender-specific data) at generating the correct gender forms. With regard to the different amounts of gender-annotated data used to train our gender-aware models, we cannot see any appreciable variation in term coverage and gender accuracy between the two *-BAL and *-ALL settings. Further insights on this aspect are provided in the next section.

5.4.4.2 Cross-gender Analysis

Table 5.7 shows separate term coverage and gender accuracy scores for FEM and MASC target gender forms. This allows us to highlight the models' translation ability for each gender form and conduct cross-gender comparisons to detect potential bias.

	En-it				**en-fr**			
	Feminine		**Masculine**		**Feminine**		**Masculine**	
	Cover.	Acc.	Cover.	Acc.	Cover.	Acc.	Cover.	Acc.
MT for KD	**66.25**	16.23	**61.46**	88.49	**63.76**	16.24	**62.41**	89.58
Base-KD-only	58.75	20.85	53.41	84.93	58.59	26.91	59.76	79.44
Base	58.75	33.62	53.66	80.45	60.47	32.30	63.61	79.55
Multi-DecPrep-Bal	60.00	68.75	53.90	60.63	61.41	68.58	60.48	70.12
Multi-DecPrep-All	58.00	69.83	55.12	63.72	61.88	65.78	60.72	75.00
Specialized-Bal	62.00	**79.84**	53.90	**93.67**	62.59	**79.32**	60.96	**93.28**
Specialized-All	62.00	**79.84**	54.15	**95.05**	62.59	**79.32**	62.17	**93.80**

Table 5.7: Coverage and accuracy scores for feminine and masculine word forms.

First, it is worth remarking that, also at this cross-gender level of analysis, results are consistent across language pairs. We assess that both the `MT` model and its strongly connected

Base-KD-only present a very strong bias since they almost always produce masculine forms: accuracy is always much lower than 50% on the feminine set (up to 20.85% for en-it and 26.91% for en-fr) and very high on the masculine set (up to 88.49% for en-it and 89.58% for en-fr). After fine-tuning without KD, the Base ST models improve feminine forms realization, although they still remain far from 50%. The comparison with the direct model presented in the previous Section 5.3.2 shows that, despite the much higher overall translation quality, our Base models are affected by a stronger bias. This further confirms the detrimental effect of KD on gender translation and that higher overall quality does not directly imply a better treatment of speaker's gender translation.

All gender-aware models significantly reduce bias with respect to the Base systems. This is particularly evident in the feminine set, where accuracy scores far above 50% indicate their ability to correctly represent female speakers. In particular, the Specialized models achieve the best results on both feminine and masculine sets (over 79% and 93% respectively). The higher performance on the masculine set can be explained considering that the two gender-specialized models derive from the Base model, which is strongly biased towards masculine forms. Interestingly, Multi-DecPrep shows similar feminine/masculine accuracy scores. This is possibly due to the random initialization of the gender tokens' embeddings (described in 5.4.2.2); as a result, the initial model hidden representations and predictions are perturbed randomly as well. Such a starting condition combined with balanced data leads to a comparatively fairer, similar behaviour across genders for Multi-gender models, although their absolute accuracy scores are lower with respect to the Specialized ones.

Finally, we observe that results obtained by training our models with balanced (*-Bal) and unbalanced (*-All) datasets are similar. Indeed, the masculine gender accuracy slightly improves by adding more male data, while there is not a clear trend for feminine accuracy. We can conclude that oversampling the data is functional inasmuch it keeps the performance on the feminine set stable.

5.4.4.3 Conflicting Vocal Characteristics and Gender Tags

So far, we have worked under the assumption that speakers' vocal characteristics match with those typically associated with the gender category she/he identifies with. In this section, instead, we explicitly explore systems' capacity to correctly produce speakers' linguistic expressions of gender for scenarios in which this assumption does not hold. As previously discussed in Section 5.2, this can be the case for some transgender individuals, children, or people with vocal impairment. However, we are hindered by the almost absent representation of such users within MuST-C. As such, as represented in Figure 5.1we design a counterfactual experiment where we

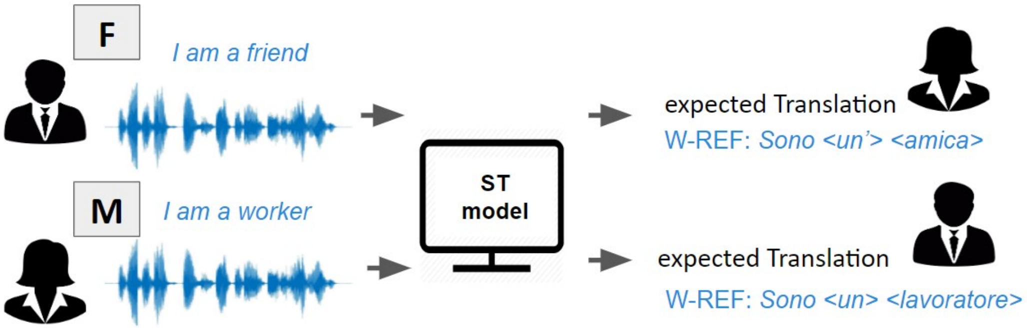


Figure 5.1: Design for counterfactual experiments on conflicting vocal/tag gender information. The depicted scenario includes the en-it language pair. English input sentences uttered by male/female speakers are, respectively, associated with <F> and <M> tags. We evaluate gender-aware systems ability of generating the FEM/MASC gender form required by the tag by relying on MuST-SHE W-REF translations.

associate the opposite gender tag to each female/male speaker and inspect models' behaviour when receiving conflicting information between the gender tag and the properties of the acoustic signal.

Table 5.8 presents the results for this experiment. In the *M-audio/F-tag* set, systems were fed with a male voice and a <F> tag and the expected translation is in the feminine form, while in the *F-audio/M-tag* set we have the opposite. As we can see, in both sets the `Multi-gender` model has a drastic drop in accuracy with respect to the results shown in Table 5.7, with scores below 50% for en-it. This behaviour indicates that `Multi-gender` systems rely on both the gender token and the audio features, which in this scenario are conflicting. Thus, this model is not usable in scenarios in which the vocal characteristics shall be ignored. On the contrary, the `Specialized` systems show higher accuracy scores on both sets. In particular, on *F-audio/M-tag* the performance is in line with the results of Table 5.7. This indicates that, independently of speakers' vocal characteristics, the model relies only on the provided gender information, being therefore suitable for situations in which one wants to control the gendered forms in the output and override the potentially misleading speech signals. For the *M-audio/F-tag* scenario the `Specialized` system works reasonably well (accuracy scores above 50%), although the generation of feminine forms still proves comparatively more difficult.

	En-it				**en-fr**			
	M-audio/F-tag		**F-audio/M-tag**		**M-audio/F-tag**		**F-audio/M-tag**	
	Cover.	Acc.	Cover.	Acc.	Cover.	Acc.	Cover.	Acc.
Multi-DecPrep-All	54.88	45.78	60.25	38.17	61.93	45.14	61.18	55.77
Specialized-All	54.39	64.57	60.75	94.24	62.17	59.69	61.41	94.25

Table 5.8: Coverage and accuracy scores when the correct translation is expected in a gender form opposite to the speaker's gender but in accordance with the gender tag fed to the system.

(a) Fem	SRC	I was **the classic Asian student**...	
	REF-it	Ero **la classica studentessa asiatica**...	
	`Base`	Ero **il classico studente asiatico**...	✗✗✗✗
	`Multi` & `Spec`	Ero **la classica studentessa asiatica**...	✔✔✔✔
(b) Fem	SRC	As a **researcher**, a **professor**...	
	REF-fr	En tant que **chercheuse**, **professeure**...	
	`Base`	En tant que **chercheur**, **professeur**...	✗✗
	`Multi` & `Spec`	En tant que **chercheuse**, **professeur**...	✔✗
(c) Masc	SRC	...the woman who wanted to know me as an **adult**.	
	REF-it	...la donna che voleva vedere come fossi da **adulto**.	
	`Base` & `Multi`	... una donna che voleva conoscermi da **adulta**.	✗
	`Spec`	... una donna che voleva conoscermi da **adulto**.	✔
(d) Fem	SRC	When I was a **kid**...	
	REF-fr	Quand j'étais **petite**...	
	`Base` & `Multi` & `Spec`	Quand j' ai été **tuée**...	?
(e) Fem	SRC	...while downhill skiing **paralyzed** me.	
	REF-it	...quando un incidente sciistico mi ha **paralizzata**.	
	`Base` & `Multi` & `Spec`	... quando mi **paralizzò**.	?
(f) Fem	SRC	I was **elected**...	
	REF-it	Sono stata **eletta**...	
	`Base`	Fui **eletto**...	✗
	`Multi` & `Spec`	Fui **selezionata**...	?

Table 5.9: Examples of feminine FEM and MASC gender-marked words translated by `Base`, `Multi-DecPrep-All` (`Multi`) and `Specialized-All` (`Spec`) on en-it and en-fr.

5.4.5 Manual Analysis

We complement our automatic evaluation with a manual inspection on the output of three models: `Base`, `Multi-DecPrep-All` (`Multi`), and Specialized-All (`Spec`). For each model, we analyzed the translation of 100 common segments across en-it/en-fr, which allow for cross-lingual comparisons.

We first take into account those instances where systems' accuracy in the production of gender-marked words was *measurable*, as in (a), (b), (c) in Table 5.9. A first observation, consistent across languages and models, is that a controlling noun (*student*) and its modifiers (*the, classic, Asian*) seem to always concord in gender in the systems' output. As per (a), this agreement is respected for both correct (`Multi`, `Spec`) and wrong gender realizations (`Base`). Differently, (b) shows that, whenever two words are not related by any morphosyntactic dependency, some words may be correctly translated (*chercheuse* – `Multi`, `Spec`), and some others not (*professeur*). Such a behaviour seems to attest that, although the systems are fed with sentence-level gender tags, gender predictions are still skewed at the level of the single word. Overall, (a), (b) and (c) clearly attest the progressively improved performance from `Base` to `Multi` and `Spec`. In particular, in (c), `Spec` is able to pick the required masculine form in spite of a contextual hint about a second female referent (*woman*), thus overcoming what is a difficult (masculine) prediction even for `Multi`.

We also inspect those cases where systems' accuracy on gender production was not *measurable* to cast some light on the reasons for a limited *term coverge*. We found that, while there are some generally wrong translations – (d) – such instances only amount to 1/3 of the cases. In the remaining 2/3, the output is fluent and reflects the source utterance meaning but it simply does not match the exact annotated word in the reference. We found that ST translations often offer alternative constructions that do not require an overt gender-inflection – (e) – or rely on appropriate gender-marked synonyms of the word in the reference – (f). We can hence conclude that many gender translations that do not contribute to *gender accuracy* confirm an improved ability of the enriched models in gender translation.

Overall, our manual analysis is in line with the automatic results, inasmuch it confirms the improved capability of our gender-aware models to handle speaker-dependent gender-translation, especially in the case of `Specialized` systems. Additionally, we have unveiled other qualitative insights concerning gender bias in ST. Namely, that gender-translation is skewed at the level of single – but syntactically independent – words in spite of gender tags operating at the sentence-level. While the hereby conducted qualitative assessment is limited in scope and extension, in Chapter 7, we will pick upon this finding to systematically explore *i)* which words are more impacted by gender bias, and *ii)* its effect on gender agreement among words.

5.4.6 Summary

Here, we rose to the challenge posed in the previous section 5.3 to further explore and mitigate gender bias in direct ST. Such an additional step has required exposing and recognizing how bias and harmful behaviours can first and foremost arise from our own (biased) conceptualization of gender, which we in turn embed in our systems (Keyes, 2018). Namely, from the conflation of perceptual markers of gender (i.e., voice), social identities and linguistic expressions, with the risk of perpetuating essentialist, binary distinctions based on biometric features (Zimman, 2018). Thus, going beyond audio signals, we developed gender-aware models suitable for operating conditions where speakers' gender is known. To this aim, we annotated the large MuST-C dataset with validated speakers' gender information, and used the new annotations to experiment with different architectural solutions: "multi-gender" and "specialized". Through results on two language pairs (en-it/fr), we demonstrated the improvements in gender realization brought by breeding gender-aware ST models. In particular, our specialized systems outperform the gender-unaware ST models by 30 points in gender accuracy without affecting overall translation quality. In addition, we show that specialized systems make a reasonable solution to control gender translation and override unreliable gender cues inferred from speakers' vocal characteristics.

5.5 Conclusion

In this chapter, we opened up a new line of research focused on gender bias in speech translation. Inquiries in this area were unprecedented, as previous work had largely focused on gender bias for text-to-text MT. In particular, we have started by foregrounding the role of audio input features as a variable for the study of gender bias in speech-guided translation technology. Indeed, by being fed on parallel *audio-translation* data, ST has the potential to access speakers' vocal features, which are largely regarded as perceptual markers of gender (Sec. 5.2).

Accordingly, in Section 5.3 we investigate for the first time whether ST systems leverage such audio information to translate gender. To this aim, we have analyzed the behaviour of cascade (pipelined ASR+MT solution) and direct ST technology, which generates target text from audio input without any intermediate representation. For two language pairs (en-it and en-fr), our evaluation shows that, altough to different extent, both approaches are affected by a masculine skew. Also, we find that – regardless of lower generic performance – the direct approach does actually exploit audio information to translate speaker-related gender phenomena. When tested on MuST-SHE, this emerges as a potential advantage of direct ST, as revealed by gender accuracy scores. But is that always the case?

Motivated by the fact that *i)* relying on vocal biometric features as a gender cue can be unsuitable and potentially harmful for certain users (e.g., transgender individuals), and *ii)* that direct ST does nonetheless exhibit a strong masculine bias due to the under-representation of feminine forms in their training data, in Section 5.4 we experimented with different solutions for bias mitigation going beyond audio signals. Rather, our proposed "multi-gender" and "specialized" ST approaches are informed about speakers' gender via externally validated information, and offer suitable solutions to override audio cues. To build them, we have relied on two newly created resources: a gender-balanced validation set and training data annotated with speakers' gender information (Sec. 5.4.1).

Overall, our findings pave the way for future studies in direct ST, an emerging technology that is gaining increasing traction (Bentivogli et al., 2021). For this reason, the upcoming portions of this work keep their focus on *direct ST* technology, whose quick advancements should be promptly informed by social concerns as well. Note that this chapter has adopted a rather data-centric perspective towards the causes and remedies for bias, i.e., modality diet (audio vs. textual data), balancing feminine and masculine training data distribution (multi-gender solution), enriching data with meta-information (specialized solution). In the next Chapter 6, instead, we delve into the exploration of algorithmic choices that concur to the emergence of bias.

6

How to Split: Word Segmentation and Gender Bias

Having recognized gender bias as a major issue affecting current translation technologies, most attempts towards mitigation have worked on the data side. However, less effort has been put into exploring whether algorithmic aspects concur to exacerbate the problem. In this chapter, we bring the analysis of gender bias in automatic translation onto a seemingly neutral yet critical component: word segmentation. Can segmenting methods influence the ability to translate gender? Do certain segmentation approaches penalize the representation of feminine linguistic markings? We address these questions by comparing 5 existing segmentation strategies on the target side of speech translation systems. Namely: the two statistically-driven methods of byte-pair encoding (BPE) and Dynamic Programming Encoding (DPE), the two morphologically-motivated Morfessor and the Linguistically Motivated Vocabulary Reduction (LMVR) methods, and finally segmentation at the level of single characters. Our results on two language pairs (English-Italian/French) show that state-of-the-art subword splitting (BPE) comes at the cost of higher gender bias. In light of this finding, we propose a combined approach that preserves BPE overall translation quality, while leveraging the higher ability of character-based segmentation to properly translate gender.

6.1 Introduction

When addressing gender bias, most works identified *data* as the primary source of gender asymmetries for current models. Accordingly, many pointed out the misrepresentation of gender groups in datasets (Garnerin et al., 2019; Vanmassenhove et al., 2018), which reflect the under-participation of women — e.g., in the media (Madaan et al., 2018; Devinney et al., 2020). Hence, preventive initiatives concerning data documentation have emerged (Bender and Friedman, 2018), and research has focused on the development of data-centered mitigating techniques (Zmigrod et al., 2019). For instance, as we also did in the previous Chapter 5, by training models on *ad-hoc* gender-balanced datasets (Saunders and Byrne, 2020; Costa-jussà and de Jorge, 2020), or by enriching data with additional gender information (Moryossef et al., 2019; Vanmassenhove et al., 2018; Elaraby and Zahran, 2019; Saunders et al., 2020; Stafanovičs et al., 2020).

As discussed in Section 3.3, however, although data are not the only factor contributing to bias, only few inquiries in MT devoted attention to other technical components that exacerbate the problem (Vanmassenhove et al., 2019) or to architectural changes that can contribute to its mitigation (Costa-jussà et al., 2022c; Saunders et al., 2022). Indeed, by focusing on the algorithmic choices considered for MT speed-accuracy optimization[1] (e.g., reducing the number of layers in the decoder to speed up inference time), Renduchintala and Williams (2021) find that such practices have an impact on gender translation by amplifying masculine bias. Still from an algorithmic perspective, Roberts et al. (2020) additionally expose how "taken-for-granted" approaches may come with high overall translation quality, but are actually detrimental when it comes to gender bias. Such is the case for the state-of-the-art beam search decoding technique that – with respect to sampling – is remarkably effective at maximizing BLEU scores (Papineni et al., 2002), while however yielding less linguistically diverse outputs and a higher, wrong generation of masculine pronouns.

Along this line, we focus on ST systems and inspect a core aspect of neural models: word segmentation. Byte-Pair Encoding (BPE) (Sennrich et al., 2016) represents the *de-facto* standard and has recently shown to yield better results compared to character-based segmentation in ST (Di Gangi et al., 2020). But does this hold true for gender translation as well? If not, why?

Languages like French and Italian often exhibit comparatively complex feminine forms, derived from the masculine ones by means of an additional suffix (e.g., en: *professor*, fr: *professeur* Masc vs. *professeure* Fem). Additionally, women and their referential linguistic

[1] Accuracy-speed trade-offs are largely considered in the production of MT models, with the goal of optimizing the process in terms of time required to translate and the overall quality of the model.

expressions of gender are typically under-represented in existing corpora (see also the distribution of TED speakers in Sec. 5.4.1). In light of the above, purely statistical segmentation methods could be unfavourable for gender translation, as they can break the morphological structure of words and thus lose relevant linguistic information (Ataman et al., 2017). Indeed, as BPE merges the character sequences that co-occur more frequently, rarer or more complex feminine-marked words may result in less compact sequences of tokens (e.g., en: *described*, it: *des@@critto* MASC. vs. *des@@crit@@ta* FEM.).[2] Due to such typological and distributive conditions, may certain splitting methods render feminine gender less probable and hinder its prediction?

We address such questions by implementing different families of segmentation approaches employed on the decoder side[3] of ST models built on the same training data. By comparing the resulting models both in terms of overall translation quality and gender accuracy, we explore whether a so far considered irrelevant aspect like word segmentation can actually affect gender translation. As such, our main contributions are as follows: **(1)** we perform a comprehensive analysis of the results obtained by 5 popular segmentation techniques for the two en-fr and en-it language directions in ST. **(2)** We find that the target segmentation method is indeed an important factor for models' gender bias. Our experiments consistently show that BPE leads to the highest BLEU scores, while character-based models are the best at translating gender. **(3)** Finally, we propose a multi-decoder architecture able to combine BPE overall translation quality and the higher ability to translate gender of character-based segmentation.

Accordingly, in Section 6.2, we discuss the interaction between different word segmentation methods and linguistic features, with a focus on gender. Then, Section 6.3 describes the five direct ST systems and segmentation techniques employed in our experiments, whose behavior is compared and extensively analyzed in Section 6.4. Finally, a solution that combines two segmentation techniques is presented in Section 6.5.

6.2 Segmentation and Linguistic Phenomena

Towards investigating the potential impact of segmentation on gender translation, we draw on existing studies highlighting how segmentation can interact with different linguistic phenomena and frequency distributions (6.2.1). Accordingly, we then foreground how gender is represented in the two language pairs and corpora used in our study (6.2.2).

[2] Example taken from our own set of BPE-segmented data. The '@@' symbol is used to indicate segmentation boundaries.

[3] As previously described in Section 2.3.2, since direct ST systems receive source audio input, *textual* segmentation is only applied to the target text in the decoder.

6.2.1 The Impact of Segmentation

As previously discussed in Section 2.3.2, although early attempts in neural MT employed word-level sequences (Sutskever et al., 2014; Bahdanau et al., 2015), the need for open-vocabulary systems able to translate rare/unseen words led to the definition of several word segmentation techniques. Currently, the statistically motivated approach based on byte-pair encoding (BPE) by Sennrich et al. (2016) represents the *de facto* standard in MT. Recently, its superiority to character-segmentation (Costa-jussà and Fonollosa, 2016; Chung et al., 2016) – i.e., splitting at the level of single characters – has been also proved in the context of ST (Di Gangi et al., 2020).

However, depending on the languages involved in the translation task, the data conditions, and the linguistic properties taken into account, BPE greedy procedures can be suboptimal. By breaking the surface of words into plausible semantic units, linguistically motivated segmentations (Smit et al., 2014; Ataman et al., 2017) were proven more effective for low-resource and morphologically-rich languages (e.g., agglutinative languages like Turkish), which often have a high level of sparsity in the lexical distribution due to their numerous derivational and inflectional variants. Moreover, fine-grained analyses comparing the grammaticality of character, morpheme and BPE-based models exhibited different capabilities. Sennrich (2017) and Ataman et al. (2019) show the syntactic advantage of BPE in managing several agreement phenomena in German, a language that requires resolving long range dependencies. In contrast, Belinkov et al. (2020) demonstrate that while subword units better capture semantic information, character-level representations perform best at generalizing morphology, thus being more robust in handling unknown and low-frequency words.

Indeed, as the above-mentioned studies show, using different atomic units can affect models' ability to handle specific linguistic phenomena. However, whether low gender translation accuracy can be to a certain extent considered a by-product of certain compression algorithms is still unknown.[4]

6.2.2 Gendered Morphology and Language Data

As just discussed, the effect of segmentation strategies can vary depending on language typology (Ponti et al., 2019) and data conditions. To inspect the interaction between word segmentation and gender expressions, we thus first clarify some relevant aspects of grammatical gender in the two languages of our interest: French and Italian. Then, we verify their representation in the datasets used for our experiments.

[4] Rather, in our first work on the comparison of cascade and direct ST systems (Sec. 5.3.1), these models were built, respectively, with BPE and character-level segmentations. In such a work, we assumed the different segmentation methods would not affect gender phenomena.

Gendered morphology. As the reader is already aware of, the extent to which information about the gender of referents is grammatically encoded varies across languages (Hellinger and Motschenbacher, 2015; Gygax et al., 2019) (see Sec. 2.2). Unlike English – whose gender distinction is chiefly displayed via personal pronouns (e.g., *he/she*) – grammatical gender languages like French and Italian systematically articulate such gender distinctions on several parts of speech (Hockett, 1958; Corbett, 1991, 2013c). Accordingly, many lexical items exist in both feminine and masculine variants, overtly marked through morphology (e.g., en: *the tired kid sat down*; it: *il bimbo stanco si è seduto* MASC. vs. *la bimba stanca si è seduta* FEM.). As the example shows, the word forms are distinguished by two morphemes (*–o*, *–a*), which respectively represent the most common inflections for Italian (singular) masculine and feminine markings.[5] In French, the morphological mechanism is slightly different (Schafroth, 2003), as it relies on an additive suffixation on top of masculine words to express feminine gender (e.g., en: *an expert is gone*, fr: *un expert est allé* M vs. *une experte est allée* F). Hence, feminine French forms require an additional morpheme. Similarly, another productive strategy – typical for a set of personal nouns – is the derivation of feminine words via specific affixes for both French (*e.g., –eure, –euse*)[6] and Italian (*–essa, –ina, –trice*) (Schafroth, 2003; Chini, 1995), whose residual evidence is still found in some English forms (e.g., *heroine, actress*) (Umera-Okeke, 2012).

In light of the above, translating gender from English into French and Italian poses several challenges to automatic models. The first – which we have already largely explored also in the previous chapters – is that gender translation from English into Italian and French does not allow for one-to-one mapping between source and target words. Second, the richer morphology of the target languages increases the number of variants and thus *data sparsity*, i.e., the higher number of possible inflections for any given lexical entry – or lemma – leads to a lower average frequency for each of its variants in a given corpus.[7] Hereby, the question is whether – and to what extent – statistical word segmentation differently treats the less frequent variants. Also, considering the morphological complexity of some feminine forms, we speculate whether linguistically unaware splitting may disadvantage their translation. To test these hypotheses, we explore below if such conditions are represented in the ST datasets used in our study: the MuST-SHE test set and MuST-C training data.

Gender in used Datasets MuST-SHE contains ~2,000 pairs of (correct/wrong) gendered forms (e.g., en: *tired*, fr: *fatiguée* vs. *fatigué*) that we compare in terms of *i)* length, and *ii)*

[5] In a fusional language like Italian, one single morpheme can denote several properties as, in this case, gender and singular number (the plural forms would be *bimbi* vs. *bimbe*).

[6] French also requires additional modification on feminine forms due to phonological rules (e.g. en: *chef/spy*, fr: *cheffe/espionne* vs. *chef/espion*).

[7] From a NMT perspective, this entails that models have fewer linguistic "examples" for each variant to learn from.

frequency in the MuST-C training set. As regards frequency, we assess that, for both language pairs, feminine variants are less frequent than their masculine counterpart in over 86% of the cases. Among the exceptions, we mostly find words that are linked to controversial or socially connoted activities (e.g., *raped, nurses*). Looking at words' length, 15% of Italian feminine forms result to be longer than their masculine counterparts, whereas in French this percentage amounts to almost 95%. These scores confirm that MuST-SHE reflects the typological features described above.

6.3 Experimental Settings

All the **direct ST systems** used in our experiments are built in the same fashion within a controlled environment, so to keep the effect of different word segmentations as the only variable. Accordingly, all our models are based on the same ST-adapation of the Transformer (Vaswani et al., 2017) as described in the previous Section 5.4.2. Except for SpecAugment (Park et al., 2019; Bahar et al., 2019b)[8], for the sake of hypothesis-testing here we do not rely on additional data augmentation or transfer-learning technique to boost model's performance. Rather, we train all of our models on the fairly uniform MuST-C corpus only, which contains 492 hours of speech for en-fr and 465 for en-it. MuST-C target Italian and French text was tokenized with Moses.[9]

So to avoid rewarding models' potentially biased behaviour, as a validation set we rely on the MuST-C gender-balanced dev set (Sec. 5.4.1). The models are optimized using label smoothed cross entropy (Szegedy et al., 2016) and their trainings are stopped after 5 epochs without improvements on the validation loss. Our final models are created by averaging the 5 model checkpoints around the best one on the validation set.

6.3.1 Segmentation Techniques

To allow for a comprehensive comparison of word segmentation's impact on gender bias in ST, we identified three substantially different categories of splitting techniques to which the five candidates selected for our experiments belong.

Character Segmentation. Splitting words at their maximal level of granularity, **character-based** solutions have been first proposed by Ling et al. (2015) and Costa-jussà and Fonollosa

[8] Such a procedure is fundamental for the development of current direct systems.

[9] `https://github.com/moses-smt/mosesdecoder`. Tokenization is a preprocessing step where unstructured textual data are broke down into meaningful elements, e.g., words. Sometimes used as a synonym for word segmentation (Mielke et al., 2021), we here avoid terminological overlaps and use *tokenization* for splitting into words (*tokens*), whereas we refer to (word) *segmentation* for splitting into subword *units*.

(2016). This technique proves simple and particularly effective at generalizing over unseen words. On the other hand, character splitting results in longer sequences[10] that increase the memory footprint, and thus slow down a model in both training and inference phases (Libovický et al., 2022). We perform our segmentation by appending "@@ " to all characters but the last of each word.

Statistical Segmentation. This family comprises data-driven algorithms that generate statistically significant subwords units. The most popular one is **BPE** (Sennrich et al., 2016),[11] which proceeds by merging the most frequently co-occurring characters or character sequences. Recently, He et al. (2020) introduced the Dynamic Programming Encoding **(DPE)** algorithm, which performs competitively with respect to BPE, and was claimed to accidentally produce more linguistically-plausible subwords: according to He et al. (2020) this might result from the fact that DPE conditions the segmentation of target words on the source sentence (e.g., src-de: *Wagen*, tgt-en BPE: *car* + *ts* vs. DPE *cart* + *s*). DPE is obtained by training a mixed character-subword model. As such, the computational cost of a DPE-based ST model is around twice that of a BPE-based one. We trained the DPE segmentation on the transcripts and the target translations of the MuST-C training set, using the same settings of the original paper.[12]

Morphological Segmentation. A third possibility is linguistically-guided tokenization that follows morpheme boundaries. Among the unsupervised approaches,[13] one of the most widespread tools is **Morfessor** (Creutz and Lagus, 2005), which was extended by Ataman et al. (2017) to control the size of the output vocabulary, giving birth to the **LMVR** – i.e., Linguistically Motivated Vocabulary Reduction – segmentation method. These techniques have outperformed other approaches when dealing with low-resource and/or morphologically-rich languages (Ataman and Federico, 2018). In other languages, they are not as effective, so they are not widely adopted. Both Morfessor and LMVR have been trained on the MuST-C training set.[14]

As explained in Section 2.3.2, we need to set the desired vocabulary size by determining the

[10] To exemplify, consider the lenght of *descritto* in Italian, split into characters (*d@@e@@s@@c@@r@@i@@t@@t@@o*) vs. subwords (*des@@critto*). As discussed by Libovický et al. (2022), this comes at the expense of higher computation costs, and might be a deterring factor to the adoption of character-level methods.

[11] We use SentencePiece (Kudo and Richardson, 2018): `https://github.com/google/sentencepiece`.

[12] See `https://github.com/xlhex/dpe`.

[13] Note that these approaches are still data-driven, and not purely rule-based. As explained in Mielke et al. (2021), they are informed by and leverage distributional regularities and morphosyntactic constraints of a given language to find and generate segmentations with high correlations to morpheme boundaries.

[14] We used the parameters and commands suggested in `https://github.com/d-ataman/lmvr/blob/master/examples/example-train-segment.sh`

[15] Segmentation is case-sensitive, and include punctuation, symbols and numbers, too.

	en-fr	**en-it**
# tokens	5.4M	4.6M
# types	96K	118K
BPE	8,048	8,064
Char[15]	304	256
DPE	7,864	8,008
Morfessor	26,728	24,048
LMVR	21,632	19,264

Table 6.1: MuST-C types, tokens, and resulting dictionary sizes per each segmentation method.

number of merge rules. For fair comparison, we chose the optimal vocabulary size for each method (when applicable). Following Di Gangi et al. (2020), we employed 8k merge rules for BPE and DPE. In LMVR, instead, the desired target dimension is actually only an upper bound for the vocabulary size. We tested 32k and 16k, but we only report the results with 32k as it proved to be the best configuration.[16] Finally, character-level segmentation and Morfessor do not allow to determine the vocabulary size. In table 6.1, we show the original distribution of types (unique words) and tokens (all words) in MuST-C data and the size of the resulting dictionaries per each method.

6.3.2 Evaluation

This work is motivated by the intent to shed light on whether issues in the generation of feminine forms are also a by-product of current segmentation algorithms. In our view, architectural improvements of ST systems should also account for the relation, or trade-offs, between overall translation quality and gender representation. Thus, to explore the impact of word segmentation on gender bias, as well as how it relates to generic performance, we are interested in measuring both *i)* the overall translation quality obtained by different segmentation techniques, and *ii)* the correct generation of gender forms.

Accordingly, we evaluate translation quality on both the MuST-C *tst-COMMON* set (2,574 sentences for en-it and 2,632 for en-fr) and MuST-SHE (§6.2.2), using SacreBLEU (Post, 2018a).[17] For fine-grained analysis on gender translation, we rely on gender accuracy as described in Section 5.4.3. We report gender accuracy for the two categories of phenomena in MuST-SHE (see Sec. 4.2). To recap, Category (1) contains first-person references (e.g., *I'm a student*), which are contextually ambiguous.[18] Gender phenomena of Category (2), instead,

[16] As per preliminary analyses carried out with the Spacy analyzer (Honnibal et al., 2020), the 32k bound was ideal with respect to the distribution of morphological tags in MuST-C.

[17] `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3.`

[18] In category 1, we let ST models leverage the correlation between speakers' vocal characteristics as a gender cue to infer gender translation and gender forms. Although potentially harmful, here we do not investigate methods to

	en-fr			en-it		
	MuST-C	MuST-SHE	Avg.	MuST-C	MuST-SHE	Avg.
BPE	**30.7**	25.9	**28.3**	21.4	**21.8**	21.6
Char	29.5	24.2	26.9	21.3	20.7	21.0
DPE	29.8	25.3	27.6	21.9	21.7	**21.8**
Morfessor	29.7	25.7	27.7	21.7	21.4	21.6
LMVR	30.3	**26.0**	28.2	**22.0**	21.5	**21.8**

Table 6.2: SacreBLEU scores on MuST-C tst-COMMON (MuST-C) and MuST-SHE for en-fr and en-it.

shall be translated in concordance with other gender information in the sentence (e.g., *she/he is a student).*

6.4 Comparison of Segmentation Methods

Table 6.2 shows the overall **translation quality** of our ST systems trained with distinct segmentation techniques: BPE, Char, DPE, Morfessor, and LMVR. By looking at these results, BPE comes out as competitive as LMVR for both language pairs. On averaged results, it exhibits a small gap (0.2 BLEU) also with DPE on en-it, while it achieves the best performance on en-fr. The disparities are small though: they range within 0.5 BLEU, apart from Char standing ~1 BLEU below. Overall, in light of the trade-off between computational cost (LMVR and DPE require a dedicated training phase for data segmentation) and average performance (BPE achieves winning scores on en-fr and competitive for en-it), **we hold BPE as the best segmentation strategy in terms of general translation quality** for direct ST.

Turning to **gender translation**, the gender accuracy scores presented in Table 6.3 exhibit that all ST models are clearly biased, with masculine forms (M) disproportionately produced across language pairs and categories. However, we intend to pinpoint the relative gains and losses among segmenting methods. Focusing on overall accuracy (ALL), we see that – despite its lowest performance in terms of BLEU score – **Char emerges as the favourite segmentation for gender translation.** For French, however, DPE is only slightly behind. Looking at morphological methods, they surprisingly do not outperform the statistical ones. The greatest variations are detected for feminine forms of Category 1 (1F), where none of the segmentation techniques reaches 50% of accuracy, meaning that they are all worse than a random choice when the speaker should be addressed by feminine expressions. Char appears close to such threshold,

control gender translation, as we already did this in the previous section. Rather, we experimented with unmodified models for the sake of hypothesis testing without adding variability. As discussed in our results, in this way we unveiled the potential of certain word segmentation techniques to better capture correlations from the received input, a capability that could be exploited to redirect ST attention away from speakers' vocal characteristics by means of other information provided.

	en-fr				
	ALL	1F	1M	2F	2M
`BPE`	65.18	37.17	75.44	61.20	80.80
`Char`	**68.85**	48.21	74.78	65.89	**81.03**
`DPE`	68.55	**49.12**	70.29	**66.22**	80.90
`Morfessor`	67.05	42.73	75.11	63.02	80.98
`LMVR`	65.38	32.89	**76.96**	61.87	79.95
	en-it				
`BPE`	67.47	33.17	88.50	60.26	81.82
`Char`	**71.69**	**48.33**	85.07	**64.65**	**84.33**
`DPE`	68.86	44.83	81.58	59.32	82.62
`Morfessor`	65.46	36.61	81.04	56.94	79.61
`LMVR`	69.77	39.64	**89.00**	63.85	83.03

Table 6.3: Gender accuracy (%) for MuST-SHE Overall (ALL), as well as across masculine (M) and feminine (F) forms for Category 1 and 2 on en-fr and en-it.

while the others (apart from `DPE` in French) are significantly lower.

These results illustrate that target segmentation is a relevant parameter for gender translation. In the comparison between `Char` and `BPE`, the advantage of the former model emerges for feminine translation across both categories of phenomena, although the larger gap concerns category 1. In particular, this might suggest that character-level segmentation improves models' ability to learn correlations between the received input and gender forms in the reference translations. Although in this experiment models rely only on speakers' vocal characteristics to infer gender – which we discourage as a cue for gender translation for real-world deployment – such ability shows a potential advantage for Char, which could be better redirected toward learning correlations with reliable gender meta-information included in the input. For instance, in a scenario in which meta-information (e.g., a gender tag as in Sec. 5.4) is added to the input to support gender translation, a `Char` model might better exploit this information. Our question then becomes: What makes character segmentation less biased? What are the segmentation features determining a better/worse ability in generating the correct gender forms?

6.4.1 Lexical Diversity

We posit that the limited generation of feminine forms can be framed as an issue of data sparsity, whereas the advantage of characters-based segmentation ensues from its ability to handle less frequent and unseen words (Belinkov et al., 2020). In fact, Vanmassenhove et al. (2018); Roberts et al. (2020) link the MT reduction of *linguistic diversity* (i.e., the range of lexical items used in a text) with the overfitted distribution of masculine forms in its outputs. To explore such an hypothesis, we compare the lexical diversity (LD) of our models' translations and MuST-SHE

	en-fr		**en-it**	
	TTR	MATTR	TTR	MATTR
M-SHE Ref	*16.12*	*41.39*	*19.11*	*46.36*
`BPE`	14.53	39.69	17.46	44.86
`Char`	**14.97**	**40.60**	17.75	**45.65**
`DPE`	14.83	40.02	**18.07**	45.12
`Morfessor`	14.38	39.88	16.31	44.90
`LMVR`	13.87	39.98	16.33	44.71

Table 6.4: Lexical diversity scores on en-fr and en-it. As a term of comparison, we also include LD scores for MuST-SHE (human) reference translations.

references. To this aim, we rely on two measure for LD:

- Type/Token ratio **(TTR)** (Chotlos, 1944; Templin, 1957), a simple and widely used metric, yet potentially affected by fluctuations due to text length (Tweedie and Baayen, 1998). It computes the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language (translation output).

- The more robust Moving Average Type/Token ratio **(MATTR)** (Covington and McFall, 2010), which first calculates TTRs for successive non-overlapping segments of tokens within a given length (i.e., window size), and then provides the mean value of the estimated TTRs.[19]

As we can see in Table 6.4, character-based models exhibit the highest LD (the only exception is `DPE` en-it, but on the less reliable TTR metric). However, we cannot corroborate the hypothesis formulated in the above-cited studies, as LD scores do not strictly correlate with gender accuracy (Table 6.3). For instance, `LMVR` is the second-best in terms of gender accuracy on en-it, but shows a very low lexical diversity (the worst according to MATTR and second-worst according to TTR).

6.4.2 Sequence Lenght

As discussed in section 6.2.2, we know that Italian and French feminine forms are, although to a different extent, longer and less frequent than their masculine counterparts. In light of such conditions, we expected that the statistically-driven BPE segmentation would leave feminine forms unmerged at a higher rate, and thus add uncertainty to their generation. To verify if this is the actual case – explaining `BPE`'s lower gender accuracy – we check whether the number of units (characters or subwords) of a segmented feminine word is higher than that of the corresponding masculine form. We exploit the coupled wrong and correct MuST-SHE references in MuST-

[19] Metrics computed with software available at: `https://github.com/LSYS/LexicalRichness`. We set 1,000 as `window_size` for MATTR.

	en-fr (%)	**en-it** (%)
BPE	1.04	0.88
Char	1.37	0.38
DPE	2.11	0.77
Morfessor	1.62	0.45
LMVR	1.43	0.33

Table 6.5: Percentage increase of unit sequence's length for feminine words over masculine ones.

SHE, and compute the average percentage of additional units found in the feminine segmented sentences[20] over the masculine ones. Results are reported in Table 6.5.

At a first look, we observe opposite trends: BPE segmentation leads to the highest increment of units for feminine words in Italian, but to the lowest one in French. Also, DPE exhibits the highest increment in French, whereas it actually performs slightly better than Char on feminine gender translation (see Table 6.3). Hence, even the increase in sequence length does not seem to be an issue on its own for gender translation. Nonetheless, these apparently contradictory results encourage our last exploration: *How* are gender forms actually split?

6.4.3 Gender Isolation

By means of further manual analysis on 50 output sentences per each of the 5 systems, we inquire if longer sequences for feminine words can be explained in light of the different characteristics and gender productive mechanisms of the two target languages (Sec. 6.2.2). Table 6.6 reports selected instances of coupled feminine/masculine segmented words, with their respective frequency in the MuST-C training set.

Starting with Italian, we find that BPE sequence length increment indeed ensues from greedy splitting that, as we can see from examples **(a)** and **(c)**, ignores meaningful affix boundaries for both same length and different-length gender pairs, respectively. Conversely, on the French set – with 95% of feminine words longer than their masculine counterparts – BPE's low increment is precisely due to its loss of semantic units. For instance, as shown in **(e)**, BPE does not preserve the verb root (*adopt*), nor isolates the additional unit (*-e*) responsible for the feminine form, thus resulting into two words with the same sequence length (2 units). Instead DPE, which achieved the highest accuracy results for en-fr feminine translation (Table 6.3), treats the feminine additional character as a unit *per se* (**f**).

Based on such patterns, our hypothesis is that the proper splitting of the morpheme-encoded gender information as a distinct token favours gender translation, as models learn to productively generalize on it. Considering the high increment of DPE tokens for Italian in spite of the limited

[20] As such references only vary for gender-marked words, we can isolate the difference relative to gender units.

	English	Segmentation	Fem	Masc	Freq. Fem/Masc
a)	asked	BPE	chie–sta	chiesto	36/884
b)		DPE	chie–sta	chiesto	36/884
c)	friends	BPE	a–miche	amici	49/1094
d)		DPE	a–miche	amici	49/1094
e)	adopted	BPE	adop–tée	adop–té	30/103
f)		DPE	adop–t–é–e	adop–t–é	30/103
g)	sure	Morfessor	si–cura	sicuro	258/818
h)	grown up	LMVR	cresci–uta	cresci–uto	229/272
i)	celebrated	LMVR	célébr–ées	célébr–és	3/7

Table 6.6: Examples of word segmentation. The segmentation boundary is identified by "–".

number of longer feminine forms (15%), our analysis confirms that DPE is unlikely to isolate gender morphemes on the en-it language pair. As a matter of fact, it produces the same kind of coarse splitting as BPE (see **(b)** and **(d)**).

Finally, we attest that the two morphological techniques are not equally valid. Morfessor occasionally generates morphologically incorrect subwords for feminine forms by breaking the word stem (see example **(g)** where the correct stem is *sicur*). Such behavior also explains Morfessor's higher token increment with respect to LMVR. Instead, although LMVR (examples **(h)** and **(i)**) produces linguistically valid suffixes, it often condenses other grammatical categories (e.g., tense and number) with gender. As suggested above, if the pinpointed split of morpheme-encoded gender is a key factor for gender translation, LMVR's lower level of granularity explains its reduced gender accuracy. Working on character' sequences, instead, the isolation of the gender unit is always attained.

Overall, this final analysis has indeed allowed us to explain the seemingly inconsistent results obtained by measuring sequence length increase in the previous sections. Also, it raised the hypothesis of gender isolation as a factor to explain better gender translation, which ought to be further verified by means of more focused inquiries.

6.5 The Quality-Gender Trade-off

Informed by our experiments and analysis (Sec. 6.4), we conclude this study by proposing a model that combines BPE overall translation quality and Char ability to translate gender. Such a system represents a first step towards addressing the trade-off between generic performance and gender representation. To this aim, we train a multi-decoder approach that exploits both segmentations to draw on their corresponding advantages.

In the context of ST, several multi-decoder architectures have been proposed, usually to

	en-fr				
	ALL	1F	1M	2F	2M
BPE	65.18	37.17	**75.44**	61.20	80.80
Char	**68.85**	**48.21**	74.78	65.89	81.03
BPE&Char	68.04	40.61	75.11	**67.01**	**81.45**
	en-it				
BPE	67.47	33.17	**88.50**	60.26	81.82
Char	**71.69**	48.33	85.07	**64.65**	**84.33**
BPE&Char	70.05	**52.23**	84.19	59.60	81.37

Table 6.7: Gender accuracy (%) for MuST-SHE Overall (ALL), as well as across masculine (M) and feminine (F) forms for Category 1 and 2 on en-fr and en-it.

	en-fr			**en-it**		
	MuST-C	MuST-SHE	Avg.	MuST-C	MuST-SHE	Avg.
BPE	**30.7**	25.9	28.3	21.4	21.8	21.6
Char	29.5	24.2	26.9	21.3	20.7	21.0
BPE&Char	30.4	**26.5**	**28.5**	**22.1**	**22.6**	**22.3**

Table 6.8: SacreBLEU scores on MuST-C tst-COMMON (MuST-C) and MuST-SHE on en-fr and en-it.

jointly produce both transcripts and translations with a single model, i.e., a decoder is devoted to each task. Among those in which both decoders access the encoder output, here we consider the best performing architectures according to Sperber et al. (2020). As such, we consider: *i) Multitask direct*, a model with one encoder and two decoders, both exclusively attending the encoder output as proposed by Weiss et al. (2017), and *ii)* the *Triangle* model (Anastasopoulos and Chiang, 2018), in which the second decoder attends the output of both the encoder and the first decoder.

For the *Triangle* model, we used a first BPE-based decoder and a second Char decoder. With this order, we aimed to enrich BPE high quality translation with a refinement for gender translation, performed by the Char-based decoder. However, the results were negative: the second decoder seems to excessively rely on the output of the first one, thus suffering from a severe *exposure bias* (Ranzato et al., 2016) at inference time.[21] Hence, we do not report the results of these experiments.

Instead, the *Multitask direct* has two separate decoders: one BPE-based and one Char-based decoder. The system requires a limited training time increase of 10% and 20% compared to, respectively, Char and BPE models. At inference phase, instead, running time and size are the same of a BPE model. We report overall translation quality (Table 6.8) and gender accuracy

[21] Bias is here statistically intended. *Exposure bias* (Ranzato et al., 2016) refers to a discrepancy between the distributions observed by the model at training (source and reference data) and test time (model's predictions). In fact, at test time, models predict one word at the time, and rely on the generated word to predict the next one. In the case of a wrong prediction, errors can thus quickly accumulate in the output translation.

(Table 6.7) of the BPE-decoder output of the multitask model, hence called **BPE&Char**.[22] Starting with gender accuracy, the Multitask model's overall gender translation ability (ALL) is still lower, although very close, to that of the Char model. Nevertheless, feminine translation improvements with respect to BPE are present on both Category 2F for en-fr and – with a larger gain – on 1F for en-it.

Note that in the *Multi-task* model, the two decoders are disjoint and do not directly interact with each other. Thus, we believe that the presence of the Char-based decoder is beneficial to influence and capture gender information into the encoder output, which is then also exploited by the BPE-based decoder. As the encoder outputs are richer, overall translation quality is also slightly improved (Table 6.8). This finding is in line with other work (Costa-jussà et al., 2022c), which proved a strict relation between gender accuracy and the amount of gender information retained in the intermediate representations (encoder outputs).

Our results thus disclose the importance of target segmentation, whose effect can be explored under different experimental conditions, e.g., in combination with other bias mitigating strategies. With this work, we have taken a step forward in ST for English-French and English-Italian, pointing at plenty of new ground to cover concerning *how to* split for different language typologies.

6.6 Conclusion

Still focusing on direct ST, in this Chapter we studied factors concurring to the emergence of gender bias beside *data*. Accordingly, we explored whether technical choices can exacerbate gender bias by focusing on the influence of word segmentation on gender translation in ST. To this aim, we compared several word segmentation approaches on the target side of ST systems for English-French and English-Italian, in light of the linguistic gender features of the two target languages. Our results show that segmentation does affect gender translation, and that the higher BLEU scores of state-of-the-art BPE-based models come at the cost of lower gender accuracy. Moreover, first analyses on the behaviour of segmentation techniques found that improved generation of gender forms could be linked to the proper isolation of the morpheme that encodes gender information, a feature which is attained by character-level split. At last, we proposed a multi-decoder approach to leverage the qualities of both BPE and character splitting, so to balance overall translation quality in terms of BLEU score and gender representation in translation outputs, while keeping computational costs under control. Rather than being a definitive solution, I view this combined ST systems as a proof of concept

[22] Score obtained on the Char-decoder output are not reported, as they are not enhanced compared to the base Char encoder-decoder model.

towards the integration of gender representation concerns for systems' development, rather than accounting for computational efficiency/cost and generic performance, only.

In light of our finding, the next Chapter 7 enriches the evaluation of BPE-based and char-based models. Unlike this Chapter, where we focused on gender in relation to morphology, in the upcoming one we account for gender bias across different part-of-speech and morphosyntactic agreement.

7

Gender Evaluation under the Morphosyntactic Lens

Gender bias in language technologies can surface differently depending on the languages involved, thus motivating studies that foreground the specific features of the accounted languages. Grammatical gender languages are characterized by morphosyntactic chains of gender agreement, marked on a variety of lexical items and parts-of-speech (POS). However, current evaluation practices adopt a word level approach, mostly focusing on a narrow set of occupational nouns under synthetic conditions. But by remaining at the word level, are missing out and overlooking the properties of gender agreement? Also, to what extent are other POS beyond occupational nouns impacted by gender bias? To answer these questions, in this chapter we present the enrichment of the natural, gender-sensitive MuST-SHE corpus with two new linguistic annotation layers (POS and agreement chains) and explore to what extent different lexical categories and agreement phenomena are impacted by gender skews. Focusing on ST, we conduct a multifaceted evaluation on three language directions (English-French/Italian/Spanish), with models trained on varying amounts of data and different word segmentation techniques, i.e., character-level and BPE. We are able to pinpoint which lexical items are more involved in gender bias, shed light on models' behavior, and emphasize the value of disaggregated analysis.

7.1 Introduction

In previous chapters, we have examined how ST is affected by gender bias in relation to audio cues (Chapter 5) and algorithmic decisions related to word segmentation (Chapter 6). In those same chapters, we have also explored, respectively, how to alleviate contributing factors to bias by means of data-driven mitigating strategies, and by implementing changes to systems components. For all these studies, the availability of the MuST-SHE test set has been crucial not only for uncovering the problem, but also for assessing the effectiveness of mitigation efforts. Indeed, the diagnostic role of dedicated benchmarks cannot be overstated, as results obtained on test sets directly inform the direction of future research. In this chapter, while still considering a comparison of systems built with BPE and character-level segmentations, we reflect on and further develop these benchmarking practices to better account for the specific features of grammatical gender languages.

As previously discussed in Section 3.4.3, most of current evaluation practices for assessing gender bias in automatic translation commonly inspect such concerning behaviours by focusing only on a restricted set of occupational nouns (e.g., *nurse*, *doctor*), and on synthetic benchmarks (Stanovsky et al., 2019; Escudé Font and Costa-jussà, 2019; Renduchintala et al., 2021). Also, even by relying on lexically richer natural benchmarks (Alhafni et al., 2022), including MuST-SHE, the designed metrics still work at the word level, treating all gender-marked words indiscriminately (Alhafni et al., 2020; Bentivogli et al., 2020). Accordingly, current test sets and protocols: *i)* do not allow us to inspect if and to what extent different word categories participate in gender bias, *ii)* overlook the underlying morphosyntactic nature of grammatical gender on agreement chains, which cannot be monitored on single isolated words (e.g., *en*: a strange friend; *it*: una/o strana/o amica/o). In fact, to be grammatically correct, each word in the chain has to be inflected with the same (masculine or feminine) gender form.[1]

We believe that fine-grained evaluations including the analysis of gender agreement across different parts of speech (POS) are relevant not only to gain a deeper understanding of bias in grammatical gender languages, but also to inform mitigating strategies and data curation procedures. Toward these goals, our contributions in this chapter are as follows. **(1)** As a new resource, we enrich MuST-SHE with two layers of linguistic information: POS and agreement chains.[2] **(2)** To further explore how model design and overall perfomance interplay with gender bias (Roberts et al., 2020; Gaido et al., 2021), we rely on our manually curated resource to compare three ST models, which are trained on varying amounts of data, and built with different

[1] For an analogy, consider the case of (lack of) number agreement in the following: “*a dogs barks”.

[2] The annotation layers are an extension of MuST-SHE and are freely downloadable at: ict.fbk.eu/must-she/ under the same MuST-SHE licence (CC BY NC ND 4.0)

segmentation techniques: character and byte-pair-encoding (BPE) (Sennrich et al., 2016).

We carry out a multifaceted evaluation that includes automatic and extensive manual analyses on three language pairs (English-French/Italian/Spanish) and we consistently find that: *i)* not all POS are equally impacted by gender bias; *ii)* systems are nearly perfect at translating words in agreement; *iii)* ST systems produce a considerable amount of neutral rewordings in lieu of gender-marked expressions, which current binary benchmarks fail to recognize. Finally, in line with our findings in Chapter 6, we confirm that *iv)* character-based systems have an edge on translating gender phenomena, by favouring morphological and lexical diversity.

7.2 Foregrounding Linguistic Specifity

Towards a more holistic comprehension of bias, alongside technical interventions, mounting focus has been given to bias analysis in models' innards and outputs (Vig et al., 2020; Costa-jussà et al., 2022c), and to ascertain the validity of bias measurement practices (Blodgett et al., 2021; Antoniak and Mimno, 2021; Goldfarb-Tarrant et al., 2021). Complementary evidence suggests that – rather than striving for generalizations – gender bias detection can benefit from incorporating contextual and linguistic specificity (González et al., 2020; Ciora et al., 2021; Matthews et al., 2021; Malik et al., 2021; Kurpicz-Briki and Leoni, 2021), which however receives little attention due to a heavy focus on English NLP (Bender and Friedman, 2018). Purported language-agnostic approaches and evaluations (Bender, 2009), however, might not really be fully language-independent, and can prevent sensible conclusions and mitigating recommendations from being drawn, as attested by monolingual studies on grammatical gender languages (Zhou et al., 2019; Gonen et al., 2019; Zmigrod et al., 2019) and in automatic translation scenarios (Vanmassenhove et al., 2018).

Indeed, as already largely discussed in Sections 2.2.1.1 and 4.2.1, grammatical gender languages exhibit an elaborate morphological and syntactic system, where gender is overtly marked on numerous POS (e.g., verbs, determiners, nouns), and related words have to agree on the same gender features. Still, current corpora and evaluation practices do not fully foreground systems' behaviour on such grammatical constraints.[3] WinoMT (Stanovsky et al., 2019) represents the standard corpus to evaluate gender bias in MT within an English-to-grammatical gender language scenario. This corpus is "targetless": it only contains monolingual source English data, whose translation is evaluated with automatic procedures without comparison against a human reference translation. WinoMT has been progressively enriched with new features (Saunders et al., 2020; Kocmi et al., 2020), and adapted for ST (Costa-jussà et al., 2022a). While this re-

[3] More details on all corpora described below have been given in Sec. 3.4

source can be useful to diagnose gender stereotyping at scale, it excludes languages' peculiarities since it is built on the concatenation of two English corpora designed for English monolingual tasks[4]– WinoGender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018b) – which consist of synthetic sentences with the same structure and a pre-selected occupational lexicon (e.g., "The lawyer yelled at the **hairdresser** because *he* did a bad job"). To increase variability, Troles and Schmid (2021) extend WinoBias by accompanying occupations with highly gender-stereotypical verbs and adjectives. Their evaluation, though, still only considers the translated professions as to verify if the co-occuring words might skew the models' assumptions. However, gender-marking involves also several others, so far less accounted POS categories, but if they are just as problematic is not clear yet. Existing bilingual (Alhafni et al., 2021), and multilingual natural benchmarks like MuST-SHE, instead, are curated as to identify a variety of gender phenomena specifically modeled on the accounted languages. As a result, they maximize lexical and contextual variability to inspect whether translation models yield feminine under-representation in real-world-like scenarios. However, since this variability is not mapped into fine-grained linguistic information, evaluations on such corpora do not single out which instances may be more responsible for gender bias. Finally, by considering each word in isolation, they neglect the underlying features of gender agreement, which determine the grammatical acceptability of the translation.

To the best of our knowledge, only one prior work has currently interplayed issues of syntactic agreement and gender bias. Renduchintala and Williams (2021) designed a set of English sentences involving a syntactic construction that requires to translate an occupational term according to its unequivocal "gender trigger" (e.g., that *nurse* is a funny man). While they find that MT struggles even in such a simple setting, they only inspect the translation of a single disambiguated word (*nurse*) rather than a whole target group of words in agreement. In section 5.4.5, instead, we have analyzed the output of different ST systems and noted that models seem to wrongly pick divergent gender inflections for unrelated words in the same sentence (e.g., *en*: As a researcher, professor; *fr*: En tant que chercheuse$_F$, professeur$_M$) but not for dependency-related ones (e.g., *en*: The classic Asian student; *it*: [La classica studentessa asiatica]$_F$). Such analysis, however, was restricted in scope, and starting from this observation from a small output sample, we hereby intended to explore whether such a behavior holds systematically and study how POS, agreement and gender bias intersect.

[4] The Anglo-centric frame incorporated in WinoMT also emerges in its design of sentences with pro- and anti-stereotypical associations. These are based on U.S. labor market statistics, which according to González et al. (2020) are not always in line with other national gender statistics, thus potentially imposing culture-specific frames for the detection of bias in other language scenarios.

7.3 MuST-SHE Enrichment

In light of the above, a fine-grained evaluation of bias focused on POS and gender agreement requires the creation of a new dedicated resource. Towards this goal, we add two annotation layers to our TED-based MuST-SHE benchmark (Chapter 4), which is available for three language pairs: en-es, en-fr, and en-it.[5] As the only multilingual *GBET* (see Sec. 3.4.3) for both MT/ST built on natural language data, MuST-SHE is a corpus that permits the identification of numerous and qualitatively different grammatical gender instances under authentic conditions. Furthermore, the target languages covered in MuST-SHE (es, fr, it) are particularly suitable to focus on linguistic specificity. In fact, as Gygax et al. (2019) suggest, accounting for gender in languages with similar typological features allows for proper comparison. In the following, we describe the phenomena of our interest and their annotation in the reference translations of the test set (Sec. 4.3.2).

7.3.1 POS and Agreement Categorization

Parts-Of-Speech. We annotate each target gender-marked word in MuST-SHE with POS information. As shown in Table 7.1 (*a-c*), we differentiate among six POS categories:[6] ***i)*** articles, ***ii)*** pronouns, ***iii)*** nouns, ***iv)*** verbs. For adjectives, we further distinguish ***v)*** limiting adjectives with minor semantic import that determine e.g., possession, quantity, space (*my*, *some*, *this*); and ***vi)*** descriptive adjectives that convey attributes and qualities, e.g., *glad*, *exhausted*. This distinction enables to neatly sort our POS categories into the closed class of function words, or into the open one of content words (Schachter and Shopen, 2007). Since words from these two classes differ substantially in terms of variability, frequency, and semantics, we reckon they represent a relevant variable to account for in the evaluation of gender bias.

Agreement. We also enrich MuST-SHE with linguistic information that is relevant to investigate the morphosyntactic nature of grammatical gender agreement (see Sec. 2.2.1.1). Gender agreement, or *concord* (Corbett, 2006; Comrie, 1999), requires that related words match the same gender form, as in the case of *phrases*, i.e., groups of words that constitute a single linguistic unit.[7] Thus, as shown in Table 7.1, we identify and annotate as agreement chains gender-marked words that constitute a phrase, such as a noun plus its modifiers (*d*), and verb phrases for compound tenses (*e*). Also, structures that involve a gender-marked (semi-) copula

[5] We hereby also take advantage of the later released English-Spanish portion of the test set.

[6] Other POS categories (e.g., conjunctions, adverbs) are not considered since they are not inflected for gender.

[7] If agreement is not respected, the unit becomes ungrammatical e.g., *es*: *el$_{Masc}$ buen$_{Masc}$ niña$_{Fem}$ (the good kid). As we will discuss in more detail, higher range dependencies such as *subject-verb* agreement are not included in our released annotations (due to their low frequency in the corpus), yet we inspected them with a manual analysis of systems outputs in Section 7.6.2.

		PARTS-OF-SPEECH
(a)	SRC	As *one* of the *first* women...
	REF$_{fr}$	En tant que l'**une**$_{Pron}$ des **premières**$_{Adj-det}$ femmes..
(b)	SRC	As a *child growing up* in Nigeria...
	REF$_{it}$	Da **bambino**$_{Noun}$ **cresciuto**$_{Verb}$ in Nigeria.
(c)	SRC	Then *an amazing* colleague...
	REF$_{es}$	Luego **una**$_{Art}$ **asombrosa**$_{Adj-des}$ colega...
		AGREEMENT
(d)	SRC	I was *the first Muslim* homecoming queen, *the first* Somali student *senator*...
	REF$_{es}$	Fui [**la primera** reina **musulmana**] del baile, [**la primera senadora**] somalí estudiantil...
(e)	SRC	She's also *been interested* in research.
	REF$_{it}$	E' [**stata** anche **attratta**] dalla ricerca .
(f)	SRC	I also *became a* high school *teacher*.
	REF$_{fr}$	Je suis aussi [**devenu un professeur**] de lycée.

Table 7.1: MuST-SHE target **gender-marked words** annotated per $_{POS}$ and [agreement chains].

verb and its predicative complement are annotated as chains (*f*), although in such cases the agreement constraint is "weaker".[8] As example (*e*), shows, dependency chains do not need to be constituted of linearly adjacent words.[9] This annotation let us verify whether a model consistently picks the same gender paradigm for all words in the chain, enabling the assessment of its syntagmatic behaviour.

7.3.2 Manual Annotation

POS and agreement annotation was manually carried out by 6 annotators (2 per language pair) undergoing a linguistics/translation studies MA degree, and with native/excellent proficiency in the assigned target language. To ensure data quality, the two layers of linguistic information have been added: *i)* in the course of two separate annotation rounds, *ii)* independently by each annotator, and *iii)* based on detailed guidelines. I personally wrote the first version of the guidelines based on a preliminary manual analysis of a MuST-SHE sample. Successively, such guidelines have been refined and improved by means of discussions with the annotators, who had carried out a trail annotation round to get acquainted with MuST-SHE language data.[10]

Once the whole corpus had been annotated, we computed inter-annotator agreement (IAA). In the case of POS, annotators had been given a set of six labels – i.e., the POS categories described

[8] Such structure, due to the semantics of some linking verbs, can enable more flexibility. E.g., in French, *Elle est devenue*$_F$ *un*$_M$ *canard*$_M$ (*She became a duck*) is grammatical, although *un canard* (a duck) is formally MASC.

[9] The released annotations also distinguish between *continous* and *discontinous* agreement chains. As this distinction did not prove significant in our evaluations, we do not include it here.

[10] The final version of the annotation guidelines is made available at: https://bit.ly/3CdU50s

in Sec. 7.3.1 – to choose from to label each target gender-marked word already annotated in MuST-SHE. Hence, we computed IAA on label assignment with the kappa coefficient (in Scott's π formulation) (Scott, 1955), which measures the consensus between two raters who each classify N items into C mutually exclusive categories, while also accounting for consensus occurring by chance. The resulting values of 0.92 (en-es), 0.94 (en-fr) and 0.96 (en-it) correspond to "almost perfect" agreement according to its standard interpretation (Landis and Koch, 1977). Overall, across all three target languages, the few cases of lack of consensus included the identification of nominalized adjectives within noun phrases consisting of several adjectives, e.g., whether the Spanish "una feliz joven **australiana**" (a happy young Australian) is to be annotated as *adj-des* or *noun*. Still consistently across languages, the distinction between *adj-des* and *verbs* in the form of past participles initially emerged as troubling, e.g., *fr*: Vera était **morte** (Vera was dead). These cases, however, were preemptively identified in the trial annotation round, and defined distributionally (Schachter and Shopen, 2007), i.e., depending on syntactic contexts in which past participles and adjectives typically appear. Such distributional properties are inherently language-specific (more details in the above-referenced annotation guidelines).

In the case of gender agreement, annotators had to identify (when occurring) agreement chains among MuST-SHE gender-marked target words. Hence, in this case, IAA was calculated using the Dice coefficient (Dice, 1945) (previously described in Sec. 4.3.2) on the exact match of complete agreement chains, i.e., perfectly overlapping chains by the two annotators. The resulting Dice coefficients (Dice, 1945) of 89.23% (en-es), 93.0% (en-fr), and 94.34% (en-it), can be considered highly satisfactory given the more complex nature of this latter task. Except for few liminal cases that were excluded from the dataset, all disagreements were reconciled. Such liminal cases regard structures that, depending on the linguistic theory, might be viewed as constituting a syntactic dependency at the phrase-level or not. Such is the case of appositions, which were eventually not included as chains of agreement consisting of a noun and modifier, e.g., *en*: His name was Dr. Pizzutillo, an Italian American...; *it*: Il suo nome era **Dottor** Pizzutillo, **un Italoamericano**.

7.3.3 Statistics

We show the final annotation statistics in Table 7.2. Overall, we see variations across languages due to inherent cross-lingual differences. For instance, such is the case of verbs: Spanish relies less than French or Italian on the gender-enforcing ***to be*** auxiliary for intransitive constructions, resulting in less gender-marked verbs (*en*: he/she is gone; *fr*: **est** parti/ie; *it*: **è** partita/o; *es*: se ha ido). Also in the case of passive constructions (e.g., *en*: he/she was called), gender-marked verbs are found more in Italian (è **stato/a** chiamato/a), than French (a été appelé/ée) and

	en-es	**en-fr**	**en-it**	**M-SHE All**
POS (tot)	1977	1823	1942	5742
Art	450	298	382	1130
Pronoun	87	57	46	190
Adj-det	112	103	140	355
Adj-des	661	566	439	1720
Noun	563	307	319	1189
Verb	104	492	616	1212
AGR-CHAINS	420	293	421	1080

Table 7.2: Distribution of POS and agreement chains per each language and in the whole MuST-SHE.

Spanish (ha sido llamado/a), as only Italian auxiliary carry gender-inflections, thus explaining its comparatively higher number of verbs. Also, the consistently low representation of pronouns across all languages is to be explained in light of both grammatical structures and design choices. Indeed, both Spanish and Italian are both pro-drop languages, where the subject can be omitted, thus leading to a few instances of personal pronouns. Differently, French requires expressing pronouns before verbs; and yet, in light of our design choice to only include in MuST-SHE sentences featuring a gender unmarked English word that corresponded to a gender-marked one in the target languages (Sec 4.3), many gender-marked third-person pronouns in French (*Il/Elle*) were excluded since they mapped to *he/she* in the source. Finally, by looking at agreement chains, in our qualitative inspection we hypothesize that the lower French distribution can be due to some linguistic differences: *i)* definite plural articles do not mark gender (e.g., *en*: **The** friends; *it*: **[Gli/le** amici/che]; *es*: **[Los/as** amigos/as]; *fr*: Les amis/ies); *ii)* and neither do some highly frequent limiting adjectives, such as *these*, *our*, *yours* (e.g., *en*: **our** friend; [il/la **nostro/a** amico/a]; *es*: **[nuestro/a** amigo/a]; *fr*: notre ami/e).[11] All in all, while a detailed analysis of such aspects is beyond the primary scope of the chapter, these figures underscore the often unaccounted variability that exist across – even highly comparable – languages.

		en-es				**en-fr**				**en-it**			
		1977				1823				1942			
		Fem		**Masc**		**Fem**		**Masc**		**Fem**		**Masc**	
		950		1027		898		925		898		1044	
		1F	**2F**	**1M**	**2M**	**1F**	**2F**	**1M**	**2M**	**1F**	**2F**	**1M**	**2M**
		392	558	419	608	424	474	410	515	401	497	415	629
Open	*noun*	121	106	151	185	58	62	75	112	48	62	71	138
	adj-des	191	190	139	141	177	153	129	107	118	119	92	110
	verb	19	36	12	37	156	90	141	105	178	133	176	129
Closed	*article*	35	147	75	193	29	89	61	119	41	105	59	177
	pronoun	5	33	26	23	1	28	3	25	3	20	6	17
	adj-det	21	46	16	29	3	52	1	47	13	58	11	58

Table 7.3: Word-level statistics for all MuST-SHE dimensions on each language pairs: *i)* Feminine and Masculine gender forms, *ii)* Categories 1 and 2, *iii)* Open/Closed Class and POS.

[11] This is corroborated by the comparatively lower frequency of articles and adj-det in French.

For a full account of all the orginal and present annotation in MuST-SHE, in Table 7.3 we also report the word-level statistics for POS and Class across gender forms and categories.

7.4 Experimental Settings

7.4.1 Speech Translation Models

In this investigation, we further explore the comparison of models built with character-level and BPE segmentation approaches. Concurrently, in our experiments we draw on studies exploring the relation between overall system performance, model/data size and gender bias. Vig et al. (2020) posit that bias increases with model size as larger systems better emulate biased training data. Working on WinoMT/ST, (Kocmi et al., 2020) correlates higher BLEU scores and gender stereotyping, whereas (Costa-jussà et al., 2022a) shows that systems with lower performance tend to produce fewer feminine translations for occupations, but rely less on stereotypical cues. To account for these findings and inspect the behavior of different models under natural conditions, we experiment with three direct ST solutions, namely: `large-BPE`, `small-BPE`, and `small-Char`.

Developed to achieve state-of-the-art performance, **`large-BPE`** models rely on Transformer (Vaswani et al., 2017) and are trained in rich data conditions (1.25M ASR/ST utterances) by applying BPE segmentation (Sennrich et al., 2016). To achieve high performance, these systems are built as the *base* models described in Section 5.4.2, by making use of: *i)* all the available ST training corpora for the languages addressed;[12] *ii)* consolidated data augmentation methods (Nguyen et al., 2020; Park et al., 2019; Jia et al., 2019); and *iii)* knowledge transfer techniques from ASR and MT, namely component pre-training and knowledge distillation (Weiss et al., 2017; Bansal et al., 2019). In terms of BLEU score – 34.12 on en-es, 40.3 on en-fr, 27.7 on en-it – our `large-BPE` models compare favorably with published results on MuST-C test data at the time of our experiments (Bentivogli et al. 2021[13] and Le et al. 2021[14]).

Also built with Transformer-based core technology, the other systems, **`small-BPE`** and **`small-Char`** replicate the training settings and controlled data environment (i.e., MuST-C data only) of the previous chapter (see Sec. 6.3). The systems are designed to study the interplay between gender bias and splitting methods, and allow for apples-to-apples comparison between the different capabilities of BPE and character-level segmentation, namely: *i)* the syntactic

[12] As discussed in Sections 5.4.1 and 5.4.2, we know that MuST-C is characterized by a majority (70%) of masculine speakers For the other training resources, comprehensive statistics are not available but we can safely consider them as similarly biased.

[13] 32.93 on en-es, 28.56 on en-it.

[14] 28.73 on en-es, 34.98 on en-fr, 24.96 on en-it.

advantage of BPE in managing several agreement phenomena (Sennrich, 2017; Ataman et al., 2019), and *ii)* the higher capability of character-level at generalizing morphology (Belinkov et al., 2020). Given the morphological and syntactic nature of gender – now captured in MuST-SHE new annotations layers – such differences make them enticing candidates for further analysis.

7.4.2 Evaluation

We employ the enriched MuST-SHE corpus to assess generic performance and gender translation at several levels of granularity. While evaluating gender translation under natural conditions grants the advantage of inspecting diverse phenomena, the intrinsic variability of natural language can defy automatic approaches based on reference translations.[15] Indeed, since language generation is an open-ended task, systems' outputs may not contain the exact gender-marked words annotated in MuST-SHE. In fact, in our MuST-SHE evaluation method (Sec. 5.4.3), we first compute dataset *coverage*, i.e., the proportion of gender-marked words annotated in MuST-SHE that are generated by the system, and on which gender translation is hence measurable. Then, we calculate *gender accuracy* as the proportion of words generated in the correct gender among the measurable ones. As a result, all the *out of coverage* words are necessarily left unevaluated.

In this chapter, for all **word-level** gender evaluations (Sections 7.5.1 and 7.5.2), we compute accuracy as in our above-described original script and include scores based on the POS annotations. Instead, for **chain-level** gender agreement evaluation (Section 7.6.1) we modified the original script to process full agreement chains instead of single words.[16]

Finally, since we aim at gaining qualitative insights into systems' behaviour, and at ensuring a sound and thorough multifaceted evaluation, we overcome the described coverage limitation of the automatic evaluation by complementing it with a manual analysis of *all* the gender-marked words (Section 7.5.3) and agreement chains (Section 7.6.2) that remained out of coverage. This extensive manual evaluation was carried out via a systematic annotation of systems' outputs, performed by the same linguists who enriched MuST-SHE, who provided the appropriate knowledge of both the resource and the evaluation task. Accordingly, we manage to make this study completely exhaustive by covering every gender-marked instance of MuST-SHE. Also, such additional manual evaluation serves as a proof-of-concept to ensure the validity of the employed automatic evaluation metrics.

[15] This applies for all reference-based methods, as previously described also in Section 2.3.4.

[16] The scripts are released together with the MuST-SHE annotated extensions.

		BLEU	All-Cov	All-Acc	F-Acc	M-Acc
en-es	small-BPE	27.6	65.0	64.1	45.8	79.6
	small-Char	26.5	64.2	67.3	**52.8**	79.6
	large-BPE	**34.1**	**72.0**	**69.1**	**52.8**	**83.6**
en-fr	small-BPE	25.9	55.7	64.9	50.3	78.1
	small-Char	24.2	55.9	68.5	**57.7**	78.2
	large-BPE	**34.3**	**64.3**	**70.9**	57.1	**83.4**
en-it	small-BPE	21.0	53.1	67.7	52.3	80.3
	small-Char	20.7	52.6	**71.6**	**57.2**	83.9
	large-BPE	**27.5**	**59.2**	69.1	52.2	**85.4**

Table 7.4: BLEU, coverage, and gender accuracy scores computed on MuST-SHE.

7.5 Word-level Evaluation

7.5.1 Overall Quality and Gender Translation

Table 7.4 presents SacreBLEU (Post, 2018a),[17] coverage, and gender accuracy scores on the MuST-SHE test sets. All language directions exhibit a consistent trend: `large-BPE` systems unsurprisingly achieve by far the highest overall translation quality. Also, in line with our analyses in the previous chapter (Sec. 6.4) `small-BPE` models outperform the `Char` ones by ~1 BLEU point. The higher overall translation quality of `large-BPE` models is also reflected by the coverage scores (All-Cov), showing that they generate the highest number of MuST-SHE gender-marked words for all language pairs.

By turning to overall gender accuracy (All-Acc) though, the edge previously assessed for the bigger state-of-the-art systems ceases to be clear-cut. For en-es and en-fr, `large-BPE` systems outperform the concurring `small-Char` by ~2 points only – a slim advantage compared to the large gap observed on BLEU score. Moreover, for en-it, `small-Char` proves the best at translating gender.

We further zoom into the comparison of gender translation for FEM (F-Acc) and MASC (M-Acc) forms, where we can immediately assess that all ST models are skewed toward a disproportionate production of MASC forms (on average, 53.1% for F vs. 81.3% for M). However, focusing on `large-BPE` models, we discover that their higher global gender accuracy (All-Acc) is actually due to the higher generation of masculine forms, while they do not compare favorably when it comes to feminine translation. In fact, in spite of achieving the lowest generic translation quality, `small-Char` prove on par (for en-es) or even better (for en-it and en-fr) than `large-BPE` models at handling feminine gender translation.

In light of the above, our results reiterate the importance of dedicated evaluations that, unlike

[17] `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3`

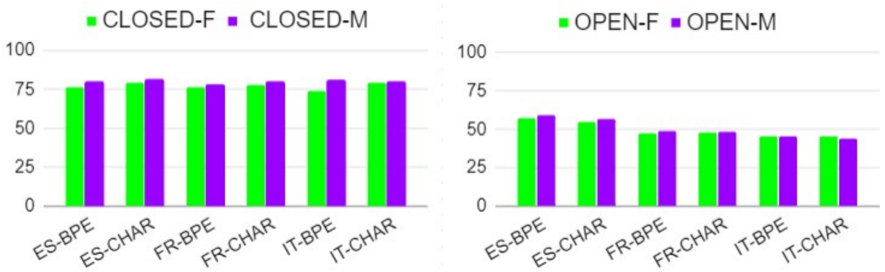


Figure 7.1: Feminine *vs.* Masculine **coverage** scores for closed and open class words.

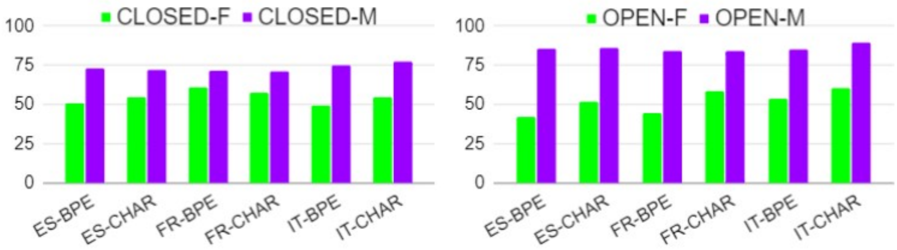


Figure 7.2: Feminine *vs.* Masculine **accuracy** scores for closed and open class words.

holistic metrics, are able to disentangle gender phenomena. As such, we can confirm that higher generic performance does not entail a superior capacity of producing feminine gender. This does not only emerge, as in the previous chapter (Sec. 6.4) in the comparison of (small) BPE- and char-based ST models. Rather, even for stronger systems, we attest how profiting from a wealth of – uncurated and synthetic (Bender et al., 2021) – data does not grant advantages to address gender bias. This motivates us to continue our multifaceted evaluation by taking into account only small models – henceforth `Char` and `BPE` – that, being trained on the same MuST-C data, allow for a sound and transparent comparison.

7.5.2 Word Classes and Parts-of-speech

At a finer level of granularity, we use our extension of MuST-SHE to inspect gender bias across open and closed class words. Their coverage ranges between 74-81% for function words, but it shrinks to 44-59% for content words (see Figure 7.1). This is expected given the limited variability and high frequency of functional items in language. Instead, the coverage of FEM and MASC forms is on par within each class for all systems, thus allowing us to evaluate gender accuracy on a comparable proportion of generated words. A bird's-eye view of Figure 7.2 attests that, although MASC forms are always disproportionately produced, the gender accuracy gap is amplified on the open class words. The consistency of such a behaviour across languages and systems suggests that content words are involved to a greater extent in gender bias. We hence analyse this more problematic class by looking into a breakdown of the results per POS.

Table 7.5 presents results for *verbs*, *nouns* and *descriptive adjectives*. First, in terms of system

		Verbs		Nouns		Adj-des	
		F-Acc	M-Acc	F-Acc	M-Acc	F-Acc	M-Acc
en-es	BPE	44.4	**93.8**	21.1	89.0	57.4	**80.0**
	Char	**60.0**	84.2	**37.4**	**89.7**	**61.2**	79.7
en-fr	BPE	51.3	**79.8**	16.4	93.5	50.6	78.6
	Char	**68.4**	75.0	**27.4**	**95.3**	**63.0**	**81.4**
en-it	BPE	63.7	83.7	28.6	92.2	62.0	76.7
	Char	**66.7**	**89.2**	**33.3**	**94.3**	**70.6**	**84.5**

Table 7.5: Feminine *vs.* Masculine accuracy scores per open class POS.

capability, Char still consistently emerge as the favorite models for feminine translation. What we find notable, though, is that even within the same class we observe evident fluctuations, where nouns come forth as the most biased POS with a huge divide between MASC and FEM accuracy (52–77 points). Specifically, scores below 50% indicate that FEM forms are generated with a probability that is below random choice, thus signalling an extremely strong bias.

In light of this finding, we hypothesize that semantic and distributional features might be a factor to interpret words' gender skew. Specifically, occupational lexicon (e.g., lawyer, professor) makes up for most of the nouns represented in MuST-SHE (~70%). This high rate of professions in TED data is not surprising *per se*: given that TED talks are held by field experts, references to education and titles are common (MacKrill et al., 2021). However, it singles out that professions may actually represent a category where systems largely rely on spurious cues to perform gender translation, even within natural conditions that do not ambiguously prompt stereotyping. We exclude basic token frequency by POS as a key factor to interpret our results, as MuST-SHE FEM nouns do not consistently appear as the POS with the lowest number of occurrences, nor do they have the highest M:F ratio within MuST-C training data.[18] Accordingly, we believe that our breakdown per POS is informative inasmuch as it unveils qualitative considerations, useful towards pursuing gender bias mitigation in corpora and models (Czarnowska et al., 2021; Doughman et al., 2021).

7.5.3 Manual Analysis

We manually inspect Char and BPE system's output on the out-of-coverage (OOC) words that could not be automatically evaluated (see "All-Cov" column in Table 7.4), which amount to more than 5,000 instances. As shown in Table 7.6, our analysis discerns between OOC words due to *i)* translation *errors* (Err), and *ii)* adequate *alternative* translations (i.e., meaning equivalent) for the expected gender-marked words. Such alternatives comprise instances in which word omission is

[18] MuST-C M:F ratio of verbs (es 1,87; fr 3,51, it 3,12) adjectives-des (es 1,85, fr 2,04; it 1,96) and nouns (es 2,11; fr 1,68; it 2,17).

		Errors
	SRC	Robert became **fearful** and **withdrawn**.
	REF$_{it}$	Robert divenne **timoroso** e **riservato**.
	OUT$_{it}$	Robert diventò **timore** e **con John**.
		(*Robert became fear and with John*)
		Acceptable Alternatives
Alt-O	Src	He was **an** artist.
	REF$_{fr}$	C'était **un** artiste.
	OUT$_{fr}$	C'était (__) artiste.
Alt-C	SRC	These girls [...], they are so **excited**...
	REF$_{es}$	Estas niñas [...], están **emocionadas**...
	OUT$_{es}$	Estas chicas [...], están **entusiasmadas**...
Alt-W	SRC	Mom [...] **became manager...**
	REF$_{it}$	Mamma [...] venne **messa** a capo di...
	OUT$_{it}$	La madre [...] diventò **capo** di...
Alt-N	SRC	I **felt** really good.
	REF$_{fr}$	Je me suis **senti** vraiment bien.
	OUT$_{fr}$	Je me **sentais** vraiment bien.

Table 7.6: Classification of OOC words.

acceptable (Alt-O) (Baker, 1992), and rewordings through synonyms or paraphrases. Since our focus remains on gender translation, we distinguish when such rewordings are generated with correct (Alt-C) or wrong (Alt-W) gender inflections, as well as neutral expressions devoid of gender-marking (Alt-N). Indeed, especially with respect to English (Cao and Daumé III, 2020; Vanmassenhove et al., 2021a; Sun et al., 2021) overcoming the structural pervasiveness of gender specifications in grammatical gender languages is extremely challenging (Gabriel et al., 2018), but some rewordings are *de facto* indirect neutral forms – also defined as indirect non-binary language (INL).[19]

The results of the analysis are shown in Figure 7.3. Surprisingly, we find that BPE models – in spite of their higher BLEU scores – accumulate more translation errors than their Char counterparts.[20] Conversely, Char models generate an overall higher proportion of alternatives and, more importantly, alternatives whose gender translation is acceptable by being correct (-C) or neutral (-N). This suggests that Char output is characterized by a favorable variability that conveys both lexical meaning and gender realization better than BPE. Also, note that the outcome of the manual analyses reiterates the results obtained with the automatic evaluation based on

[19] INL relies on generic expressions rather than gender-specific ones (e.g., *service* vs. *waiter/tress*), and is being recommended and/or implemented, especially for institutional communcation (Papadimoulis, 2018; Attanasio et al., 2021). INL strategies are distinct, and not a replacement for emerging non-binary linguistic innovations such as neo-pronouns and -morphemes (López, 2020).

[20] In the manual analysis, we noticed that Char's lower translation quality may have to do more with fluency rather than lexical issues.

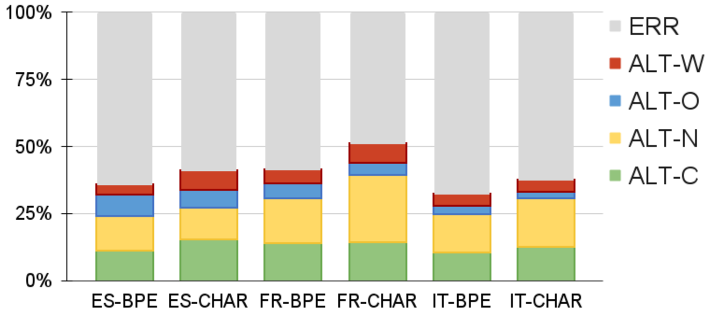


Figure 7.3: Proportion of OOC words due to translation errors and alternative translations per system.

accuracy at the word-level, thus confirming its reliability.

As a final remark, we find that all systems produce a considerable amount of neutral alternatives in their outputs. To gain insight into such neutralizations, we audit on which POS they are realized. Accordingly, we find that neutralizations of adjectives and nouns are quite limited, and concern the production of epicene synonyms (e.g., *en*: happy; *es-ref*: content$\underline{o}$/$\underline{a}$; *es-out*: feliz). Verbs, instead, are largely implicated in the phenomenon, since inflectional changes in tense and aspect paradigms (e.g., present, imperfective) that do not convey gender distinctions are feasible (see the -N example in Table 7.6). Such range of alternatives for verbs is in fact also reflected by its lowest coverage among all POS (as low as ∼32%). Finally, the most frequent way to neutralize POS in the output consists of paraphrases based on verbs, (e.g., *en*: the more I drive, the *scareder* I get; *es-ref*: y cuanto más conduzco, más *asustad* $\underline{a/o}_{Adj-des}$ me siento; *es-out*: y cuanto más manejo, más me *asusta*$_{Verb}$). Since such neutral expressions are suitable, or even preferable, for several scenarios (e.g., to substitute masculine generics, to avoid making unlicensed gender assumptions – See Sec. 3.3.1, 4.2.1) our finding encourages the creation of test sets accounting for such a third viable direction, and can shed light on systems' potential to produce INL alternatives.

7.6 Gender Agreement Evaluation

7.6.1 Automatic Analysis

The final step in our multifaceted analysis goes beyond the word level to inspect agreement chains in translation. To this aim, we define *coverage* as the proportion of generated chains matching with those annotated in MuST-SHE. Then, the *accuracy* of the generated chains accounts for 3 different cases where: *i)* agreement is respected, and with the correct gender (C); *ii)* agreement is respected, but with the wrong gender (W); and *iii)* both FEM and MASC gender inflections occur together, and thus agreement is not respected (NO).

		All			Feminine			Masculine		
		C	W	NO	C	W	NO	C	W	NO
en-es	BPE	74.3	**24.6**	**1.2**	33.9	**64.4**	**1.7**	95.5	**3.6**	0.9
	Char	**78.4**	21.0	0.6	**42.4**	57.6	0.0	**96.6**	2.6	0.9
en-fr	BPE	67.9	**31.0**	**1.2**	54.1	**45.9**	0.0	78.7	**19.1**	**2.1**
	Char	**76.7**	22.3	1.0	**57.5**	40.0	**2.5**	**88.9**	11.1	0.0
en-it	BPE	71.7	**27.5**	0.7	47.4	**50.9**	**1.8**	88.9	**11.1**	0.0
	Char	**78.5**	20.0	**1.5**	**54.2**	44.1	1.7	**97.4**	1.3	**1.3**

Table 7.7: Agreement results for All chains matched in MuST-SHE, and split into FEM and MASC chains. Accuracy scores are given for agreement respected with correct gender (C), agreement respected with wrong gender (W), agreement not respected (NO).

Table 7.7 shows accuracy scores for all MuST-SHE agreement chains (All), also split into FEM (F) and MASC (M) chains. The overall results are promising: we find very few instances (literally 1 or 2) in which ST systems produce an ungrammatical output that breaks gender agreement (NO). In fact, both systems tend to be consistent with one picked gender for the whole dependency group. Thus, in spite of previous MT studies concluding that character-based segmentation results in poorer syntactic capability (Belinkov et al., 2020), respecting concord does not appear as an issue for any of our small ST models. For the sake of comparability, however, we note that our evaluation focuses on potentially "easier" agreement chains at the phrase level, and involves language pairs that do not widely resort to long-range dependencies; this may contribute to explaining why Char better handles correct gender agreement.

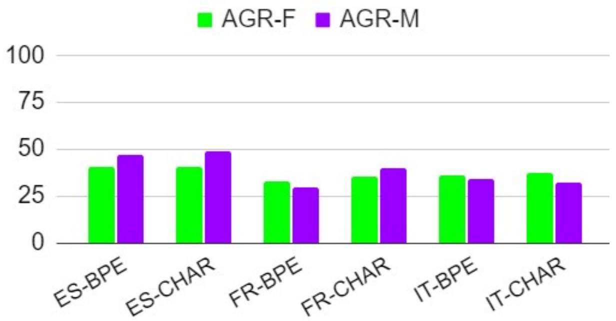


Figure 7.4: Feminine *vs* Masculine coverage scores for agreement chains.

Overall, agreement translation was measured on a lower *coverage* (30-50%) – presented in Figure 7.4 – than the word-level one (Section 7.4). While this is expected given the strict requirement of generating full chains with several words, we recover such a loss by means of the comprehensive manual evaluation discussed below, where we also include an assessment of *subject-verb* gender agreement.

7.6.2 Manual Analysis

Our manual inspection recovers a total of ∼1,200 OOC agreement chains from `Char` and `BPE` output. Similarly to the approach employed for single words (Section 7.5.3), we discern between OOC chains due to: *i)* translation *errors* (Err), and *ii) alternative* translations preserving the source meaning. We distinguish different types of alternatives. First, alternatives that do not exhibit a morphosyntactic agreement phenomenon to be judged, as in the case of neutral paraphrases or rewordings consisting of a single word (NO-chain). Instead, when the generated alternative chain exhibits gender markings, we distinguish if the chosen gender is correct (C), wrong (W), or if the system produces a chain that does not respect gender agreement because it combines both feminine and masculine gender inflections (NO).

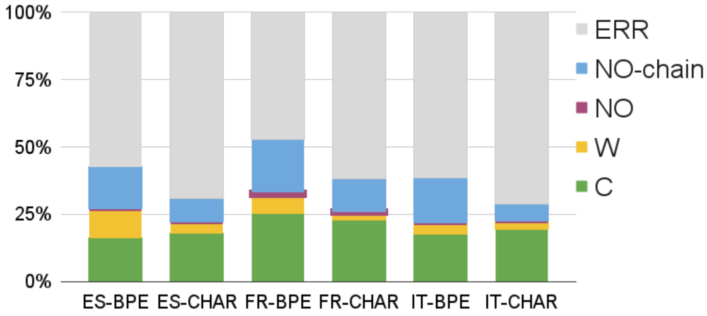


Figure 7.5: Proportion of OOC chains due to translation errors or alternative agreement translations per system.

The outcome of such OOC chains categorization is presented in Figure 7.5. Interestingly, such results are only partially corroborating previous analyses. On the one hand, unlike the OOC words' results discussed in Section 7.5.3, we attest that `Char` models produce the highest proportion of translation errors. Thus, it seems that `Char` capability in producing adequate alternatives is confined to the single-word level, whereas it exhibits a higher failure rate on longer sequences. On the other hand, by looking at alternative chains, `Char` still emerges as the best at properly translating gender agreement, with the highest proportion of chains with correct gender (C), and the lowest one with wrong gender (W).

Finally, again in line with our automatic evaluation (Table 7.7), we confirm that respecting agreement is not an issue for our ST models: we identify only 3 cases (2 for en-fr `BPE`, 1 for en-fr `Char`) where concord is broken (NO). Given the rarity of such instances, we are not able to draw definitive conclusions on the nature of these outliers. Nonetheless, we check the instances in which agreement was not respected (both in and out of coverage). We see that cases of broken concord also concern extremely simple phrases, consisting of a noun and its modifier (e.g. en: *talking to [this inventor],...because he*; fr: *parler à* ***[cette***$_F$ ***inventeur***$_M$***]****..., parce qu' il*). However, the most common type among these outliers are constructions with semi-copula verbs (e.g. en: *She... [became a vet]*; it: *...E' [***diventata**$_F$ **un**$_M$ **veterinatrio**$_M$*]*), which – as discussed in Section 4.2.1 – exhibit a weaker agreement constraint.

Subject-predicate agreement Considering long-range dependencies that go beyond the phrase level, a relevant case is also that of *subject-verb* gender agreement, which might stretch across longer sequences of text: To account for such longer spans, we considered all MuST-SHE sentences where both *i)* a word (or chain) functioning as a subject, and *ii)* its referring verb or predicative complement are annotated as gender-marked words in the reference translations; as shown in the following:

SRC **A** young **scientist** that I was working with ..., Rob, **was**..
REF$_{it}$ [[**Un** giovane **scienziato**] con cui lavoravo ..., Rob, [è **stato**]]..

We identified 55 sentences for en-es, 54 for en-fr and 41 for en-it, and we manually analyzed all the corresponding systems' outputs. We found that, even in the case of dependencies within a longer range, systems largely respect agreement in translation and consistently pick the same gender form for all co-related words. In fact, we identified only 4 cases where concord is broken: 1 case each for `BPE` and `Char` for en-es and en-it, and none for en-fr. Among these 4 cases, we found the above-discussed weaker gender-enforcing structures – i.e., (semi-)copula verbs and their predicative complements – while also detecting what appears as *agreement attraction errors* (Linzen et al., 2016). Namely, the model does not produce a verb or complement in agreement with its actual (but distant) subject, as other words intervene in the sentence and agreement is conditioned by the verb/complement's preceding word; hence, it is not respected. The following (very long) sentence is an example of such an attraction error, where the complement *desperate*, rather than agreeing in gender with the referring subject (*the nurse*), is inflected in the same masculine and plural form as its just linearly preceding noun.

SRC I watched in horror heartbreaking footage of **the head nurse**, Malak, in the aftermath of the bombing, grabbing premature babies out of their incubators, **desperate** to get them to safety, before she broke down in tears.

OUT$_{es}$ Vi una imagen horrible desgarradora de **la enfermera** (FEM, sing.) mi laguna, en los ratones después del bombardeo, agarrando a los bebés permaturos fuera de de sus incubadores (MASC, pl.) **desesperados** (MASC, pl.) por hacerlos, antes de que...

Such kind of agreement issues have more to do with overall syntactic capacity of ST models, rather than being implicated with gender bias. We can thus conclude that, even taking into account longer dependencies, agreement still does not emerge as an issue entrenched with gender bias.

7.7 Conclusion

The complex system of grammatical gender languages entails several morphosyntactic implications for different lexical categories. In this chapter, we underscored such implications and explored how different POS and grammatical agreement are involved in gender bias. To this aim, we enriched the MuST-SHE benchmark with new linguistic information, and carried out an extensive evaluation on the behaviour of ST models built with different segmentation techniques and data quantities. On three language pairs (English-French/Italian/Spanish), our study shows that, while all POS are subject to masculine skews, they are not impacted to the same extent. Respecting gender agreement for the translation of related words, instead, is not an issue for current ST models. We also find that ST generates a considerable amount of neutral expressions, suitable to replace gender-inflected ones, which however current test sets do not recognize. Overall, this multifaceted evaluation reiterates the importance of dedicated analyses that, unlike holistic metrics, can single out systems' behaviour on gender phenomena. Accordingly, our results are in line with those presented in Chapter 6 showing that, in spite of lower generic performance, character-based segmentation exhibits a better capability at handling feminine translation at different levels of granularity.

While the main focus of this chapter lies in the analysis of systems' behaviour, our insights prompt broader considerations. Specifically, our investigation on the relation between different models/segmentation techniques and gender bias provides initial cues on which models and components to audit and implement toward the goal of reducing gender bias. This, in particular, may be informative to define the path for emerging direct ST technologies. Then, our results disaggregated by POS invite reflections on how to intend and mitigate bias by means of interventions on the training data. In fact, while we have previously unveiled that the MuST-C corpus (Cattoni et al., 2021) used for training comprises a majority of masculine speakers (70%, see Section 5.4.1) the fact that certain lexical categories are more biased than others suggests that, on top of more coarse-grained quantitative attempts at gender balancing (Costa-jussà and de Jorge, 2020), data curation ought to account for more sensitive, nuanced, and qualitative asymmetries. These also imply *how*, rather than only *how often*, gender groups are represented (Wagner et al., 2015; Devinney et al., 2020). Also, although nouns come forth as the most problematic POS, current practices of data augmentation based on a pre-defined occupational lexicon may address occupational stereotyping (Saunders and Byrne, 2020), but do not increase the production of other nonetheless skewed lexical categories. Overall, our enriched resource can be useful to monitor the validity of different technical interventions in both ST and MT. Armed with this MuST-SHE enrichment, in the next and last experimental Chapter 8 of this thesis, we adopt a shift in perspective to unveil gender learning over the course of training in ST systems.

8

On Gender Learning

Due to the complexity of bias and the opaque nature of current neural approaches, auditing practices have gained traction to shed light on the correlations and (spurious) features that models rely on to perform a task. In this chapter, we contribute to such a line of inquiry by exploring the emergence of gender bias in ST. As a new perspective, rather than focusing on the final systems only, we examine their evolution over the course of training. In this way, we can account for different variables related to the dynamics of gender translation learning and investigate when and how gender divides emerge in ST. Accordingly, for three language pairs (en-es, en-fr, en-it) we compare how ST systems behave for MASC and FEM translation across training epochs and at several levels of granularity. We find that gender learning curves are dissimilar, with the feminine one being characterized by a more erratic behavior and late improvements over the course of training. Also, depending on the considered category phenomena, the trends of masculine and feminine learning curves can be either antiphase or parallel, thus suggesting that models are not always able to concurrently acquire both genders. Overall, we show how such a progressive analysis can inform on the reliability and time-wise acquisition of gender, which is concealed by static evaluations.

8.1 Introduction

In this chapter, which represents the final experimental portion of this thesis, we offer a last set of analyses aimed at grasping a better understanding of bias and its emergence in automatic translation models. Indeed, in Chapters 5, 6, and 7, we have respectively *i)* unveiled the relation between audio features and gender in ST; *ii)* evaluated in light of gender disparities the algorithmic choices underpinning the construction of current models; *iii)* and finally pinpointed the impact of gender bias on different linguistic phenomena and POS. On the wake of our previous findings, and towards a better comprehension of the complex nature of both neural approaches and bias, we hereby proceed with additional focused analyses that adopt a different perspective. Namely, over the whole course of ST training.

To the best of our knowledge – with the recent exception of the English monolingual study by Van Der Wal et al. (2022) – all current studies have adopted a static approach, which exclusively focuses on systems' biased behaviors once their training is completed. Rather than fully treating training as a black box, we explore the evolution of gender (in)capabilities across systems' training process by analyzing their output across each training round. On this basis, we address several questions that still stand unanswered. When does this gender gap emerge? How does gender bias relate to progress in terms of generic performance? To what extent is gender learning altered by the chosen system's components?

We keep the comparison between ST systems built with two segmentation techniques: character and BPE. For three language pairs (en-es/fr/it), we examine their gender learning curves for FEM and MASC translation at several levels of granularity. Overall, our contributions can be summarized as follows: **(1)** We conduct the first study that explores the dynamic emergence of gender bias in translation technologies; **(2)** By considering the trend and stability of the gender evolution, we find that *(i)* unlike overall translation quality, FEM gender translation emerges more prominently in the late training stages, and does not reach a plateau within the number of training epochs necessary to maximize systems generic performance. Such trend is however concealed by static assessments on the single, final model, and unaccounted when stopping the training of the systems. *(ii)* For contextually unambiguous gender-phenomena, masculine and feminine show a generally parallel and upwards trend, except for nouns. Characterized by flat trends and a huge gender divide, their learning dynamics suggests that ST systems confidently rely on spurious cues and generalize masculine from the very early stages of training onwards.

8.2 Studies on the Learning Process

After a strong initial wave of studies tackling gender bias by means of newly introduced debiasing and mitigating techniques (see Sec. 3.5), many studies have taken a step back to conduct *post-hoc* examinations, increasingly putting under scrutiny existing datasets (Hitti et al., 2019), models (Vig et al., 2020; Silva et al., 2021; Bau et al., 2019) and evaluation practices (Goldfarb-Tarrant et al., 2021). Along this line, Basta and Costa-jussà (2021) have underscored how "*interpreting and analyzing current data and algorithms*" represents a key path forward to address gender bias, an endeavor which is indeed complicated by the opaque nature of current neural approaches to language technologies.

As previously discussed in Section 2.1.2, the training process underlying neural systems is a black box. Thus, it is not transparent how models learn to solve a specific task, or if they rather rely on easy-to-learn shortcuts (Levesque, 2014; Geirhos et al., 2020) based on spurious correlations. The observation of models' training evolution by means of dedicated techniques, however, represents a viable method to make systems more explainable to human analyses (Van Der Wal et al., 2022).

Observing the learning dynamics of NLP models is not a new approach. It has been adopted for interpretability analysis to probe when and how linguistic capabilities emerge within language models (Saphra and Lopez, 2018, 2019), or inspect which features may be "harder/easier" to learn (Swayamdipta et al., 2020). With respect to analyses on a single (final) snapshot, a diachronic perspective has the advantage of accounting for the evolution of NLP capabilities, making them more transparent based on trends' observation. Such an understanding can then be turned into actionable improvements. Accordingly, Voita et al. (2021) looked at the time-wise development of different linguistic abilities in MT, so to inform knowledge distillation practices (Sec. 2.3.3) aimed to improve the performance of their systems. Additionally, the studies by Voita et al. (2019a,b) on the learning dynamics of extra-sentential phenomena highlighted how stopping criteria based on BLEU (Papineni et al., 2002) are unreliable for context-aware MT. Finally, Stadler et al. (2021) observed the evolution of different linguistic phenomena in systems' output, noting how the ability to handle certain phenomena seem to actually worsen across training iterations.

Overall, as Stadler et al. (2021) noted, not much effort has been put into investigating how the training process evolves with regard to measurable factors of translation quality, such as linguistic criteria (grammar, syntax, semantics). We aim to fill this gap by evaluating gender translation of ST systems at different points of their training.

8.3 Experimental Settings

8.3.1 Speech Translation Models

Our direct ST models adopt two different target segmentation: byte-pair encoding (**BPE**)[1] and (Sennrich et al., 2016) and characters (**Char**). To keep the effect of different word segmentations as the only variable and favor progress analyses as transparent as possible, our systems are built as in Sections 6.4 and 7.4. Namely, with the same core technology (Vaswani et al., 2017), within a controlled data environment (Cattoni et al., 2021), and avoiding additional procedures for boosting performance that could introduce noise.[2] Since the focus of this chapter lies on the progress of gender learning over the course of training, below we describe in more detail the training procedures that our models underwent.

Training procedure. As per standard procedure, the encoder of our ST systems is initialized with the weights of an ASR model (Bahar et al., 2019a) trained on MuST-C *audio-transcript* pairs. In our ST training – over *audio-translation* data – we use our MuST-C gender-balanced validation set to avoid rewarding systems' biased predictions (Sec. 5.4.1). As previously described in Section 2.3.1, training is basically the process of adjusting the model (i.e., minimise the error between the predicted output and reference translation via gradient descent and back-propagation) through a series of iterations over the training data. To compute the error, we adopt the label smoothed cross-entropy loss function (Szegedy et al., 2016).[3] To allow for faster computations, convergence and better generalizations, we rely on the Adam optimization algorithm (Kingma and Ba, 2015) , which is a commonly used variant of gradient descent[4] where model parameters are updated often, over smaller subsets of training data instead of the entire dataset at once. Such smaller subsets, called *mini-batches*, are of fixed size. In our setting, each mini-batch consists of 8 subsets of training data, and we set the parameters' update frequency to 8, meaning that the model's parameters are updated each time the model has seen 8 mini-batches.[5]

For each iteration over the whole training set – called *epoch* – we record 538 parameters'

[1] Using SentencePiece (Kudo and Richardson, 2018).

[2] Open source code publicly available at: `https://github.com/mgaido91/FBK-fairseq-ST`.

[3] When system performance over training is measured with label-smoothed cross entropy, instead of assigning 100% probability to the correct target word, a small amount of probability mass is distributed among all other target words. This adds a small amount of uncertainty to the target labels to make the model more robust and reduce overfitting. The overall effect is a small reduction in accuracy but improved generalization. We use 0.1 as a smoothing factor.

[4] For Adam, we set the parameters β_1=0.9, β_2=0.98, and the learning rate decays with the inverse square root policy, after increasing for the initial 4.000 updates up to 5×10^{-3}.

[5] We train on 4 GPUs.

updates for en-es, 555 for en-fr, and 512 for en-it.[6] Given the similar number of updates across languages, as a point of reference we decide to compare the progress of our systems after each full pass over the entire training set. Thus, at the end of each epoch, we save the checkpoint of our models (herein ckp), i.e., we save the state of the model and use it to translate the test set thus allowing the model's evaluation.

All models reach their best generic performance, or best ckp, within 42 epochs, with a tendency of `BPE` to converge faster than `Char`. Specifically, they respectively stop improving after 33/42 epochs (en-es), 25/29 epochs (en-fr), and 29/32 epochs (en-it). As a stopping criterion, we finish our trainings and identify the best ckp when the loss on the validation set (Sec. 2.3.2) does not show improvements for 5 consecutive epochs. Note that a common practice, which we also used in the previous experimental settings described in Chapters 5 and 6, is that of creating the ST final model by averaging the best ckp with the 2 previous and 2 following ones. Here, for the sake of our analysis across epochs, we do not generate such an averaged model and rather treat the best ckp itself as the final model. Finally, to inspect the stability of the best model results, in our analysis we also include the subsequent 5 epochs with their corresponding 5 models' ckps.

8.3.2 Evaluation

Test set and metrics. To study the evolution of gender translation over the course of training and how it relates to generic performance, we keep on using the MuST-SHE benchmark (Chapter 4) and its annotated extension (Sec. 7.3) for en-es/fr/it. To measure overall translation quality, we employ SacreBLEU (Post, 2018b).[7] To inspect gender translation, we compute word-level accuracy and coverage (Sec. 5.4.3) across several MuST-SHE dimensions: *i)* MASC and FEM gender forms; *ii)* Categories 1 and 2; and *iii)* POS from the Open class (verb, noun, descriptive adjective) and Closed class (article, pronoun, limiting adjective).[8] Comprehensive statistics for all MuST-SHE dimensions were previously given in Table 7.3.

Setup. Since we aim to observe the learning curves of our ST models, we evaluate both overall and gender translation quality after each epoch of their training process. As explained in Section 8.3.1, training includes also the 5 epochs that follow the best system ckp. To investigate systems' behaviour, we are particularly interested in the two following aspects of the learning curves: *i)* **training trend**: is gender accuracy raising across epochs, does it reach a plateau or can it

[6] The different number of updates is due to the different amount of training data (and thus resulting number of mini-batches) for the different language pairs.

[7] BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3.

[8] As our results in Section 7.6.1 yielded no significant relation between bias and gender agreement, we do not include the evaluation of agreement chains.

		BLEU	All-Cov	All-Acc	F-Acc	M-Acc
en-es	`BPE`	27.4	64.0	66.0	49.0	80.7
	`Char`	27.2	64.0	**70.5**	**58.9**	80.5
en-fr	`BPE`	24.0	53.7	65.4	51.7	77.2
	`Char`	23.5	53.1	**69.7**	**64.0**	74.9
en-it	`BPE`	20.4	48.7	65.6	49.9	79.0
	`Char`	19.1	51.2	**71.2**	**52.9**	86.7

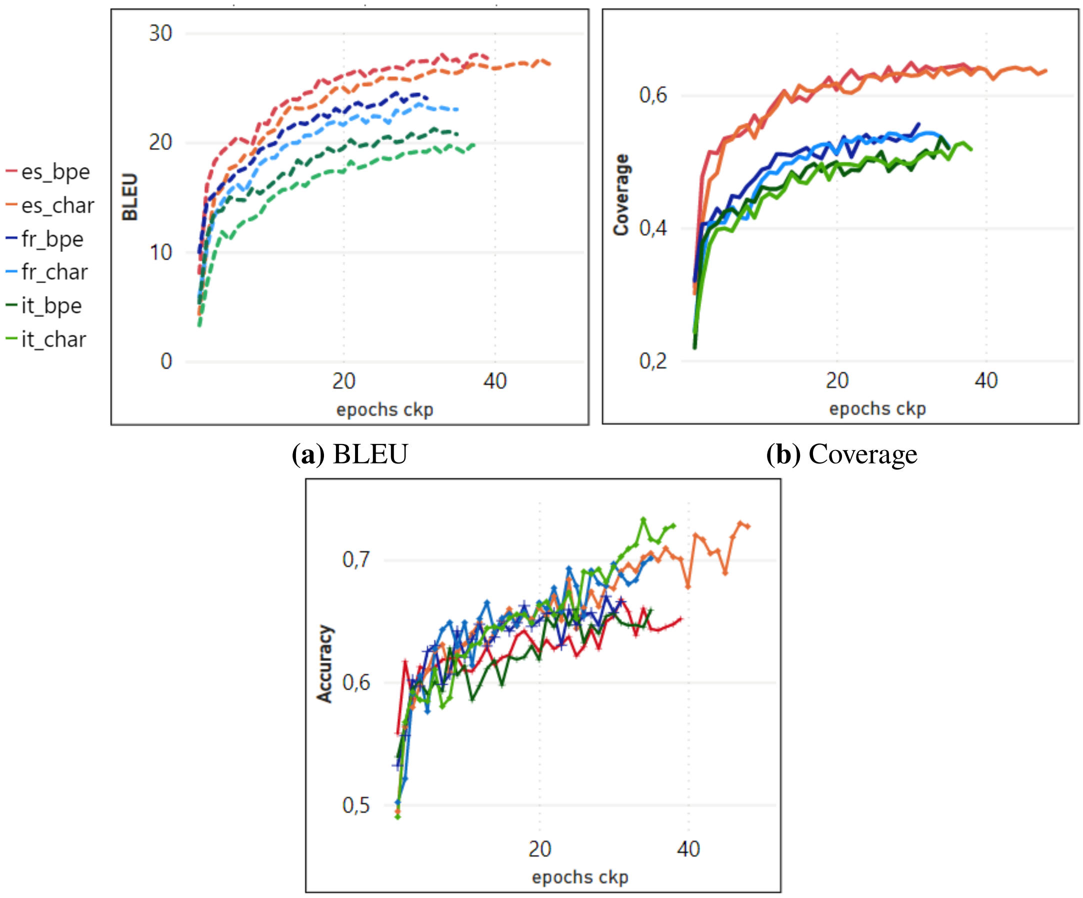
Table 8.1: BLEU, coverage, and gender accuracy scores of system's best ckp computed on MuST-SHE.

actually worsen across iterations?; *ii)* **training stability**: is gender learning steady or erratic across epochs?
Depending on the aspect addressed, we present results with different visualizations, reporting either the actual scores obtained at each ckp (more suitable to detect small fluctuations) or aggregated scores calculated with moving average over 3 ckp (more suitable to highlight general trends). Note that, since the total number of epochs differs for each system, to allow for a proper comparison we also plot results at different percentages of the training progress, where each progress point represents a 5% advancement (i.e., 5%, 10%, 15% etc.).

With this in mind, we proceed in our analyses comparing overall performance across metrics (Sec. 8.4.1), and inspecting FEM and MASC gender translation (Sec. 8.4.2) at several levels of granularity (Sec 8.4.3 and 8.4.4). For any addressed aspect, we compare `Char` and `BPE` models across language pairs.

8.4 Results and Discussion

First of all, in Table 8.1 we provide a snapshot of the results obtained by our ST models on their best ckp. As expected, the accuracy scores clearly exhibit a strong bias favoring MASC forms in translation (M-acc>F-acc), with FEM forms being generated with a probability close to a random guess for most systems for en-it. Moreover, these results are once again in line with our previous analyses in Sections 6.4 and 7.5.1, showing that `Char` has an edge in gender translation (All-Acc), which is largely ascribed to better treatment of FEM gender.[9] Thus, we confirm a previously verified behaviour, which we now further inquiry in terms of its dynamic evolution.



(a) BLEU

(b) Coverage

(c) Accuracy

Figure 8.1: Results for every ckp of each model: BLEU (a), Coverage (b), and gender Accuracy (c).

8.4.1 Overall Results

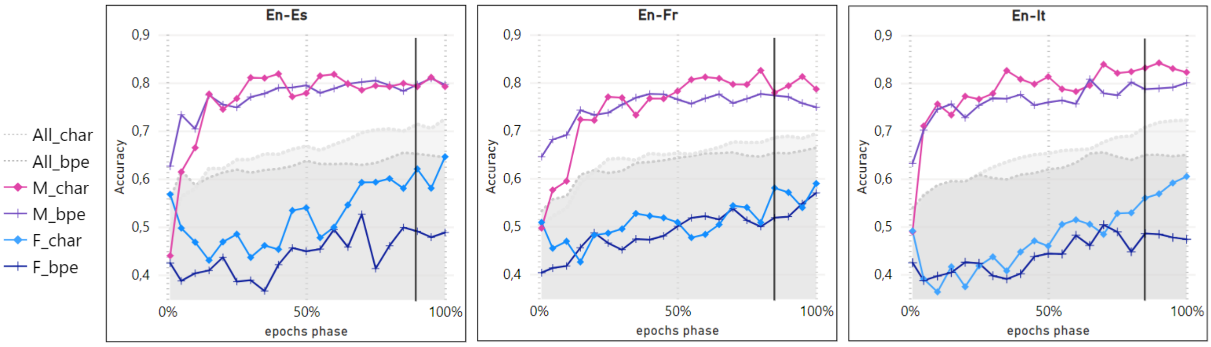
Here, we start by looking at the evolution of models' performance assessed in terms of BLEU, coverage, and accuracy (Figure 8.1) to inspect the time of emergence of the different capabilities captured by such metrics. For a bird's-eye view, we present the actual scores per each ckp.

The evolution of both overall translation performance and gender translation is positive, but dissimilar in time and quality. By looking at Figure 8.1, we observe that the gender accuracy learning curve (8.1c) immediately stands out. Indeed, the curves for both BLEU (8.1a) and gender coverage (8.1b) have a rapid and steady initial increase,[10] which starts to level off around the 20th ckp.[11] Also, the BLEU trends reveal a divide across models (BPE>Char) that remains visible over the whole course of training. In terms of coverage, the boundaries

[9] As mentioned in Sec. 8.3.1, the hereby presented systems are not created by averaging the 5 models around the best ckp. Thus, with respect to the ST systems from Chapter 6 and 7, they compare less favorably in terms of BLEU score, also reducing the performance gap bewteen *de facto* standard BPE and Char.

[10] Computed as a binary task, gender accuracy starts at ~50-55% in the first ckp. Such scores reflect that correct gender is assigned randomly at the beginning of the training process, i.e., the model is generating the wrong gender form in half of the cases (Sec. 5.4.3).

[11] The plateau is particularly visible for en-es Char due to its longer training.



(a) en-es **(b)** en-fr **(c)** en-it

Figure 8.2: (F)eminine vs M(asculine) and over(all) accuracy scores for `Char` and `BPE` in en-es (8.2a), en-fr (8.2b), and en-it (8.2c). For better comparability across systems and trend visibility, results are shown at different percentages of the training progress (increasing by 5%), and scores at each progress point are calculated with moving average over 3 ckp. The first ckp (0%) is the actual score of the first epoch. The vertical line indicates the average score for the best ckp.

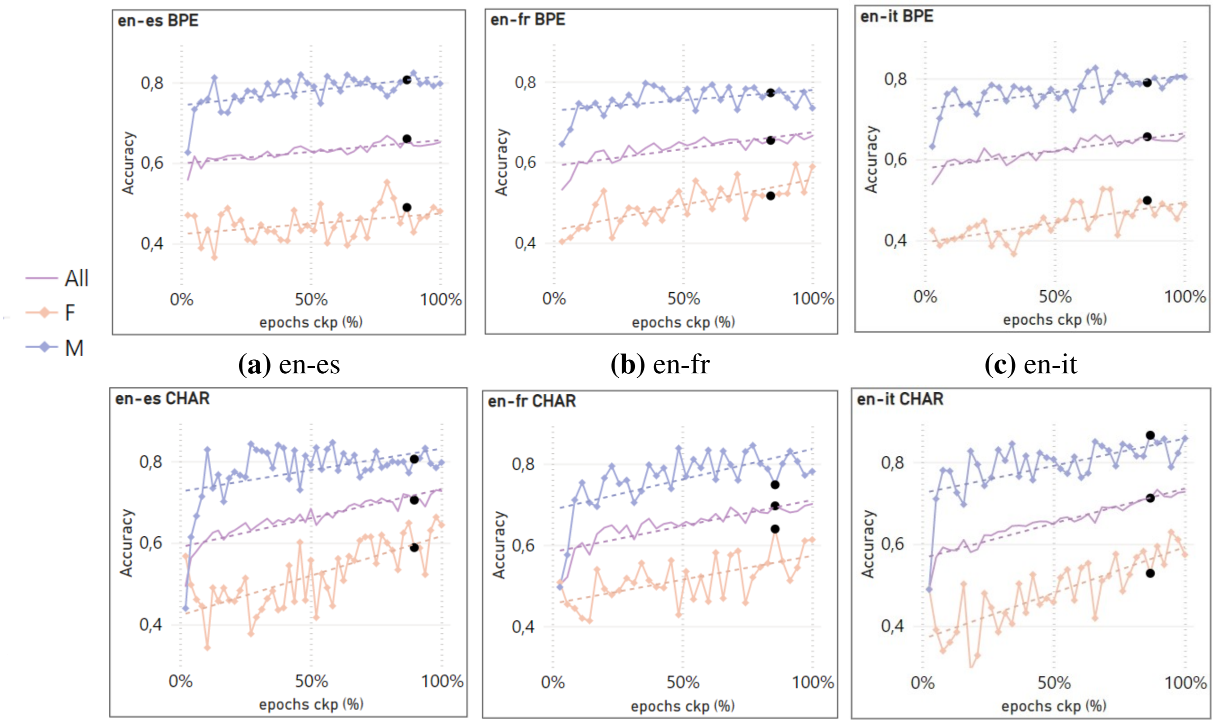
between types of models are more blurred, but correlate with BLEU scores for all language pairs. Conversely, by looking at the gender accuracy curves (8.1c) we assess that, while the overall trends show a general improvement across epochs, **gender learning *i)* proceeds with notable fluctuations, unlike the smoother BLEU and coverage curves; *ii)* emerges especially in the final iterations.** In particular, it is interesting to note that by epoch 30 (roughly 80% of the training process), *all* `Char` models handle gender translation better than *all* the `BPE` ones, regardless of the lower overall quality of the former group. Notably, the en-it `Char` system - with the lowest BLEU – exhibits the steepest increase in gender capabilities.

Takeaways. Generic translation quality improves more prominently in the initial training stages, while gender is learned later. Thus, standard quality metrics conceal and are inadequate to consider gender refinements in the learning process.

8.4.2 Masculine and Feminine Gender

Moving onto a deeper level of analysis, we compare the learning dynamics that undergo FEM (F) and MASC (M) gender in terms of accuracy. To give better visibility of their *trends* and comparisons across models, in Figure 8.2 we plot the averaged results. As a complementary view into training stability, Figure 8.3 shows the actual accuracy scores.

Masculine forms are largely and consistently acquired since the very first iterations. As shown in Figure 8.2, MASC gender (M) is basically already learnt at 15% of the training process. Henceforth, its accuracy remains high and stable within 70-80% average scores for all models. As an exception, we notice a slightly decreasing trend in the iterations that follow the best ckp for en-fr `BPE` (8.2b). **Instead, feminine translation exhibits an overall upward trend that emerges later in the training process.** In Table 8.1, we already attested `Char`'s advantage in



(a) en-es **(b)** en-fr **(c)** en-it

(d) en-es **(e)** en-fr **(f)** en-it

Figure 8.3: overAll, (F)eminine vs (M)asculine actual accuracy scores per each ckp of BPE (top row) and Char (bottom row). Black dots indicate the best ckp.

dealing with FEM translation. Here, we are able to verify how such a capability is developed over the whole course of training. Specifically, Char gains a clear advantage over BPE in the last training phases, in particular for en-es (8.2a) and en-it (8.2c). Moreover, the overall rising F trend for Char models does not seem to dwindle: even after systems have reached their best ckp, FEM translation shows potential for further improvement.

Unlike Char systems, BPE disproportionately favours masculine forms since the first ckp. In the first ckp of the training, we notice an interesting difference between BPE and Char. Namely, the former models are biased since the very beginning of their training with an evident gender divide: ∼65% accuracy for M and only ∼40% for F forms.[12] Conversely, accuracy scores for both F and M forms in Char systems present about the same accuracy: both around 50% for en-it and en-fr, whereas the en-es model notably presents lower scores on the M set. From such behaviours, we infer that Char systems *i)* are initially less prone towards masculine generalization, which is instead a by-product of further training; *ii)* promptly acquire the ability to generate both M and F inflections, although they initially assign them randomly. As we further discuss in Sec. 8.5, they occasionally acquire target morphology even before its lexicon, thus generating English source words inflected as per the morphological rules of the target language,

[12] As outlined in Sec. 5.4.3, 40% accuracy for F means that in the remaining 60% of the cases systems generate a MASC inflection instead of the expected FEM one.

e.g., en: *sister*; es: *sistera* (hermana). We regard this finding as evidence of the already attested capabilities of character-level segmentation to better handle morphology (Belinkov et al., 2020), which may explain the higher capability of `Char` models at generating feminine forms.

Despite a common upward trend, F and M gender curves progress with antiphase fluctuations. In Figure 8.3, we see how this applies to `Char` in particular. Far from being monotonic, the progress of gender translation underlies a great level of instability with notable spikes and dips in antiphase for F and M - although eventually resulting in gains for F. Interestingly, it thus seems that systems become better at enhancing F translation by partially suppressing the representation of the other gender form.

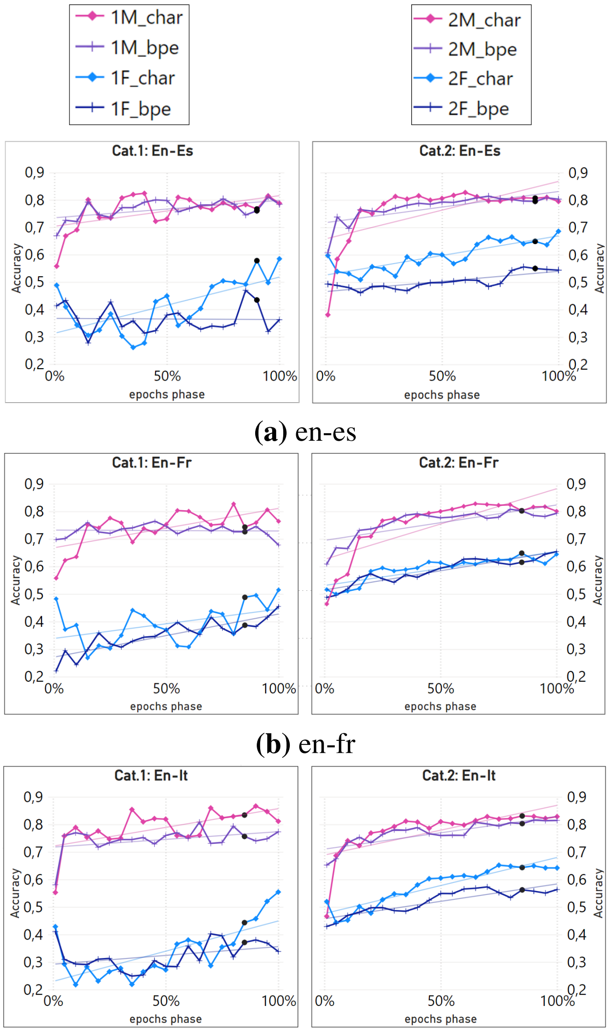
Takeaways. The insights are more fine-grained: *i)* FEM is the actual gender form that is learned late in the training process; *ii)* the progress of gender translation involves unstable antiphase fluctuations for F and M; *iii)* there is still room for improvements for FEM gender, especially for `Char` models. Overall, these findings – beside reiterating the inappropriateness of standard metrics for diagnosing gender bias (see 6.4, 7.5.1, 8.4.1) – make us also question the use of the loss function as a stopping criterion. Along this line, previous work has foregrounded that even when a model has converged in terms of BLEU, it continues to improve for context-aware phenomena (Voita et al., 2019a). Hereby, although we find a good (inverse) correlation between loss and BLEU, we attest that they seem to be unable to properly account for gender bias and the evolution of feminine capabilities. Looking at both Figures 8.2 and 8.3, we question whether a longer training would have facilitated an improvement in gender translation and, in light of F and M antiphase relation, if it would lead to a suppression of M by favoring F. If that were the case, such type of diversity could be leveraged to create more representative models.[13]

8.4.3 Gender Category

We now examine the learning curves for the translation of *i)* contextually ambiguous references to the speaker (CAT1), and *ii)* references disambiguated by a contextual cue (CAT2). Figure 8.4 shows the comparison of FEM (1F, 2F) and MASC (1M, 2M) forms.

Compared to the extremely unstable learning of CAT1, feminine and masculine curves from the unambiguous CAT2 exhibit a smooth upward parallel trend. In Figure 8.4, the differences across categories fully emerge, and are consistent across languages and models. On the one hand, F and M curves from CAT2 show a steady trend which, despite a ∼10-20% accuracy gap across genders, suggests an increasing ability to model gender cues and translate

[13] Since more ckps would be needed to investigate this point, we discuss at the end of this chapter the potential to examine this point in future work.



(a) en-es

(b) en-fr

(c) en-it

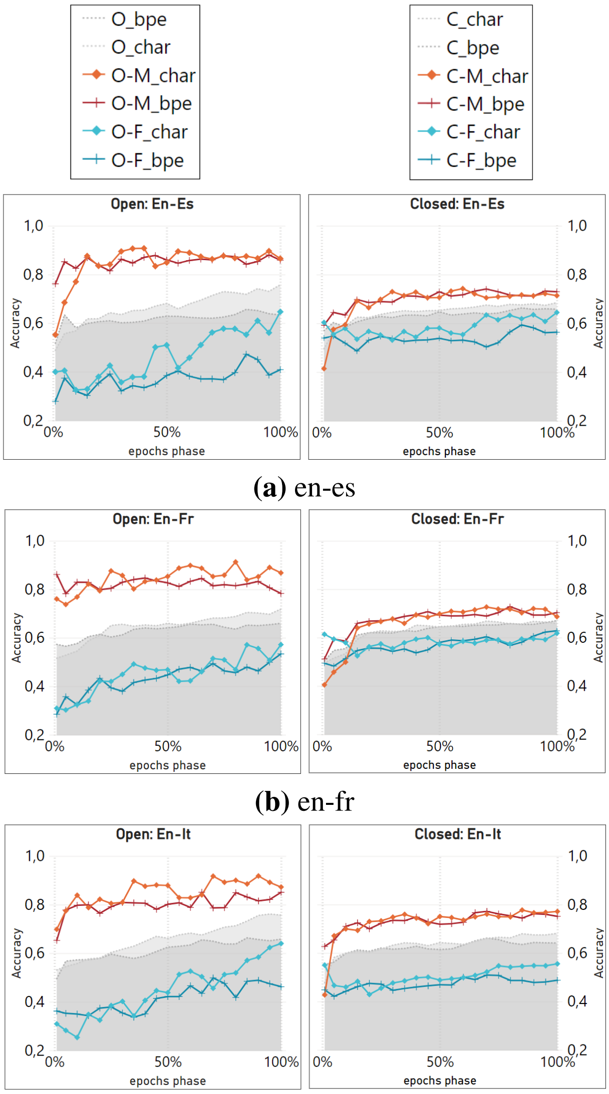
Figure 8.4: F vs M accuracy for CAT1 and CAT2. Scores are averaged over 3 ckps, and reported for each training phase. Dots indicate averaged scores for best ckp.

accordingly. On the other hand, CAT1 proves to be largely responsible for the extreme instability and antiphase behavior discussed for Figure 8.2, which is so strong to be evident even over the presented averaged scores.[14] Overall, we recognize a moderately increasing trend of 1F curves for all the Char models and the en-fr BPE. However, it barely raises above a random prediction, i.e., ~50-57% accuracy meaning that a wrong masculine form is generated in ~50-43% of the instances.

In light of the above, we obtain a new perspective on the use of audio information as a gender cue in ST (See Chapter 5), and it brings us to reflect upon the extent to which direct ST models may actually use this information to translate gender. One possible explanation for systems' behaviour on CAT1 is that – although highly undesirable – ST *does* leverage speaker's voice as

[14] For instance, the actual scores for 1F accuracy for en-it Char plummets as low as 11% at ckp9, and rockets at 60% at ckp36.

a gender cue, but finds the association "hard to learn". By looking in particular at BPE model for en-es and en-it, however, the lack of even moderate progressive increase in the F learning trend seem to suggest otherwise. Namely, that such models do not leverage audio information and deal with CAT1 as gender ambiguous input, and simply opt more frequently for masculine output in this scenario. Towards the development of ST technology that does not rely on biometric features, both observations are overall promising, and invite finding new ways to ensure the bypass of audio-gender associations in ST.



(a) en-es

(b) en-fr

(c) en-it

Figure 8.5: Open (left) vs Closed (right) classes accuracy scores per F and M gender. Scores are averaged over 3 ckps and reported for training phase.

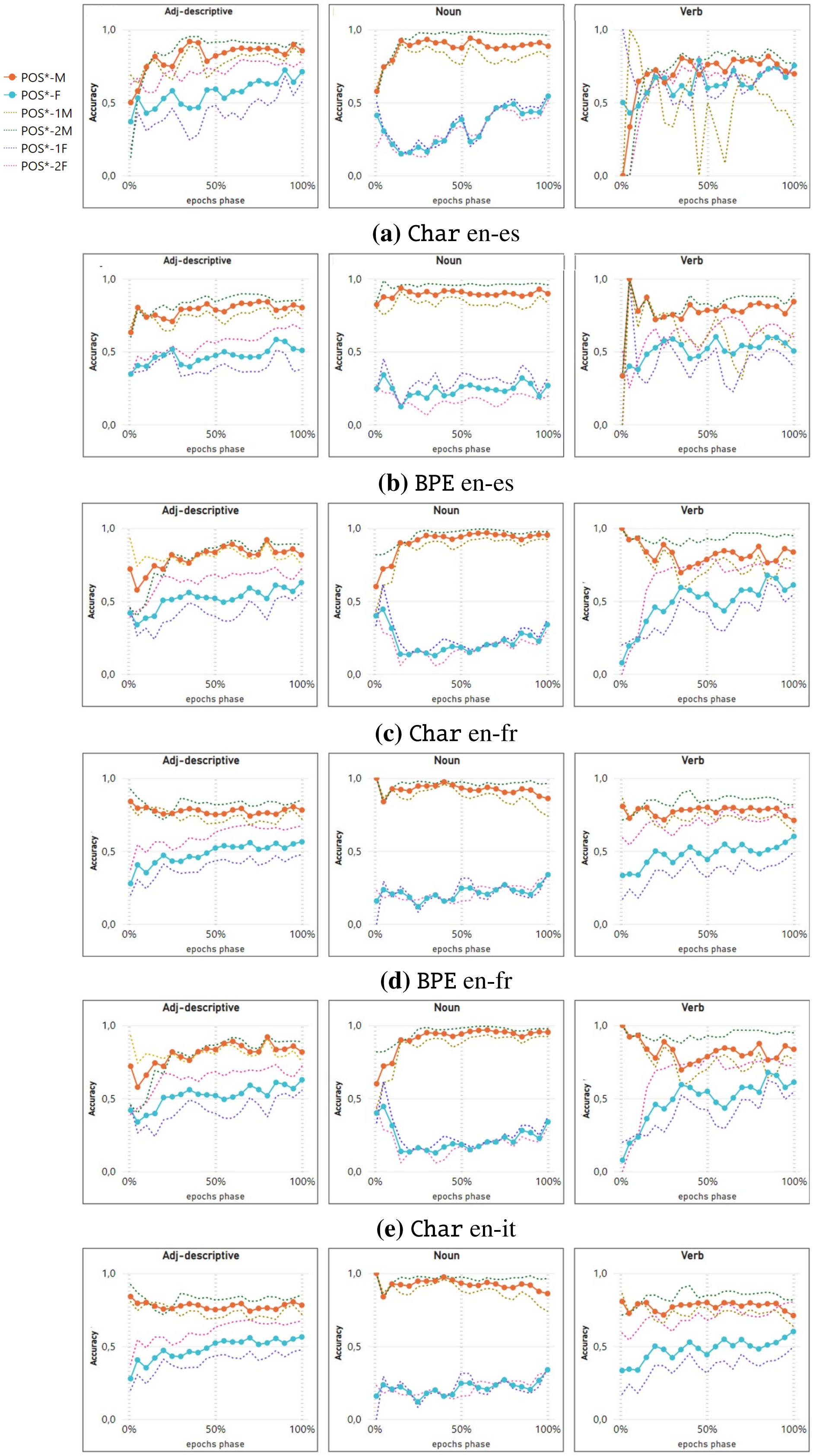
8.4.4 Class and POS

In Figure 8.5, we compare the gender curves for open and closed class words, which – as previously discussed in Sec. 7.3 – differ substantially in terms of frequency of use, variability and semantic content.

Both F and M curves of the closed class change very little over the course of training. In Figure 8.5, the closed class exhibits a stable trend with minimal increases, and a small F *vs.* M gap compared to the open class. We hypothesize this may be due to simple source constructions involving articles next to a gendered word, which are learned since the very beginning (e.g., the mum; fr: la mère). **Open class words, instead, show an unstable upward trend for F, opposed to the steady and early-learned M translation.** Consistently, the M curve starts off with unprecedented high scores (i.e., ~80% accuracy within the first 20% of the training process) which further increases to 90% accuracy scores for `Char`. The F curve is progressively improving and – once again – with more significant gains late in training. This also implies that the M/F gap is reduced over the epochs. In light of the evident bias and distinct behavior of F and M learning progress for the open class, we now turn to examine how each POS in this group evolves over training.

Nouns are outliers, being the only POS that exhibits low variability in its learning curves, with limited to almost no room for improving F translation. Consistently across languages and models, this claim can be verified in Figure 8.6. M nouns are basically fluctuation-free and reach almost perfect accuracy since the early ckps. Conversely, the F curve presents extremely low scores throughout the training process, signalling the strongest bias attested so far (i.e., the highest accuracy scores for F-nouns is around 40% for both `Char` and `BPE` in en-it and en-fr). F learning curves exhibit little to no real improvement; the only partial exception is the en-es `Char` model (8.6a), which exhibit a steeper upward trend, but still it remains just slightly above 50% accuracy. Oddly enough, unlike adjectives and verbs, nouns learning dynamics do not even reflect the different trends assessed for CAT1 and the "easier" CAT2 (Sec. 8.4.3). Namely, despite the presence of a gender cue, the translation of feminine nouns from CAT2 (2F) does not benefit from such a disambiguation information. In fact, the accuracy for 2F nouns is basically on par (or even worse) with the performance of F nouns of CAT1 (1F), whereas for any other POS – and even M-nouns – the subset of CAT2 always exhibits a more positive learning trend.

Takeaways. Overall, our remarks are in line with the findings formulated in the previous chapter (Sec. 7.5.2): nouns emerge as the lexical category that is most impacted by gender bias, arguably because systems tend to rely more on stereotypical, spurious cues for the translation of professional nouns (e.g., *scientist*, *professor*). By examining their training progress however, we additionally unveil that *i)* biased associations influence noun translation more than unambiguous and relevant information, which is available for CAT2; *ii)* ST models rely on such patterns so confidently that they never really adjust their trend over the epochs.



(a) Char en-es

(b) BPE en-es

(c) Char en-fr

(d) BPE en-fr

(e) Char en-it

(f) BPE en-it

Figure 8.6: Accuracy per each open class POS for en-es (8.6a 8.6b), en-fr (8.6c, 8.6d), and en-it (8.6e, 8.6f) Char and BPE models. The graph shows F vs M scores, also at the level of CAT1/2. Scores are provided for training phases, and calculated as the average over 3 ckps.

2F	SRC	My **older** sister Claire, she became **a** young mother, and **a master** at getting things done
CHAR$_{Es}$	(A) INI	Mi madre la *sistería sistencia*, encontró **una** madre joven **una** madre, y **una** madre, y una cosa que estaba cosas.
	(B) INI	Mi madre *sistera* de clara se convirtió en un joven jóvenes jóvenes jóvenes, y **un maestro** de cosas que hicieron.
	(C) MID	Mi hermana más **antigua** clara, se convirtió en **una** madre joven y **una maestra** que hacía cosas.
	(D) MID	Mi **antigua** hermana Clare, se convirtió en **una** madre joven y **una maestra** que hice las cosas.
	(E) FIN	Mi hermana mayor Clare, se convirtió en **una** madre joven, y **una maestra** de hacer las cosas.
	(F) FIN	Mi hermana mayor, Clare, se convirtió en **una** joven madre, y **una maestra** por hacer las cosas.
BPE$_{Es}$	(G) INI	Mi hija de la Tierra se convirtió en **un** joven joven, ella me convertí en **una** madre, y **un** hombre que hizo cosas.
	(H) INI	Mi hermana mayor claridad se convirtió en **una** madre joven, y **una maestra**, lo hice.
	(I) MID	Mi hermana mayor se volvió a ser **una** joven madre, y una maestría que hice.
	(J) MID	Mi hermana mayor declaró, se convirtió en **una** madre joven, y **un maestro** que se está haciendo.
	(K) FIN	Mi hermana mayor declaró que se convirtió en **una** madre joven, y **un amo** logrando hacer las cosas.
	(L) FIN	Mi hermana mayor Clare se convirtió en **una** madre joven, y **un dueño** de hacer que las cosas se hicieran.

Table 8.2: En-es outputs at initial, middle, and final epochs. The source sentence contains **neutral words** to be translated according to the available gender cues. In the outputs, we indicate correct feminine gender translation vs masculine. We also signal repetitions and *copied source lemma*+target morphology combinations.

8.5 Qualitative Insights

We conclude our analysis with a manual inspection of the outputs of our ST systems at two *i)* initial, *ii)* middle, and *ii)* final ckps of their training process. To this aim, we opt for the en-es language pair – for which we observed the highest BLEU and gender coverage scores (Table 8.1) – to minimize the amount of low-quality translations that could be hard to analyze. Table 8.2 presents an example sentence from CAT2, translated by both Char and BPE, which backs up some of the quantitative observations formulated in Sec. 8.4. The source sentence contains neutral words (*older*, *a*, *a master*) occurring together with gender-marked words that disambiguate the correct gender (*sister*, *she*, *mother*). Given the presence of these gender cues, the neutral words should be fairly easy to translate.

In the first two ckps of both models, the output has very low quality.[15] It is characterized by extensive repetitions of frequent words, like *mother* (*madre* in A) or *young* (*joven*, *jóvenes* in B-G). Also, whereas functional words from the closed class are already appropriately employed and inflected with the correct gender (e.g., ***a*** *mother* → ***una*** *madre*, A,G), the noun *master* is not learned yet and remains out of coverage; notably, BPE generates the word *hombre* (i.e., *man*) instead. Interestingly, if we also look at the gender cue noun *sister*, it maps to another kinship term *daughter* for BPE (G), whereas Char generalizes target morphology over English lexicon at these stages (*sistería* in A, *sistera* in B).

Such lexical issues are all refined by the middle ckps, where the systems have acquired both *sister/hermana* (with the FEM inflected adjective *antigua*[16] in C) and *master/maestr**, which in this case undergoes an interesting gender evolution across systems. For Char, we assist to an

[15] Still, we believe that it is to a certain degree intelligible thanks to the ASR initialization, see Sec. 8.3.1.

[16] The most fluent choice would be the neutral *mayor*. Here, however, we just focus on gender realizations.

adjustment from MASC inflection (B), to a FEM one (C onwards) that remains stable until the end of training, even after the best ckp. Instead, BPE has a reversed trend. In H, the output *una maestra* reveals that this system can and has learned to generate feminine forms. However, as the training progresses it switches to MASC inflections and never rebounds to the FEM ones. Rather, in the last epochs K-L it produces alternative synonyms, but always with the wrong gender (*amo, dueño*).[17] Overall, this case exemplifies how, in spite of *i)* having F morphological capabilities, and *ii)* the presence of a close cue disambiguating gender, the BPE system confidently relies on spurious and irrelevant patterns for gender translation.

8.6 Conclusion

The understanding of bias in language technologies is hindered by the complex and opaque nature of current neural approaches. To shed light on the emergence of bias in ST, in this chapter we focused on gender learning dynamics over the course of ST training. In this way, we adopted a new perspective that accounts for the time-wise appearance of gender capabilities, and examine their stability, reliability, and course of development. For three language pairs (en-es, en-fr, and en-it) we inspected the learning curves of FEM and MASC gender translation *i)* at several levels of granularity; *ii)* with respect to progress in terms of overall translation quality; *iii)* on the output of ST systems trained on target data segmented as either character or BPE sub-words units. In our diachronic analysis, we unveiled that – although gender refinements are concealed by standard evaluation metrics – *i)* feminine gender is learned late over the course of training, *ii)* it never reaches a plateau within the number of iterations required for model convergence at training time. This observation is unfortunately limited by the inclusion of only 5 epochs after the best validation loss, and future work is needed to confirm whether and to what extent FEM learning keeps improving after the best epoch. Such a direction, however, could inform studies on how to leverage diversified output to alleviate gender bias in our models, as well as with regard to gender-sensitive stopping criteria. Finally, by looking at the stability vs. fluctuations of the explored trends, we identified under which circumstances ST models seem to progressively acquire feminine and masculine translation, and when instead their erratic, antiphase behavior reflects unreliable choices made by the systems. In this way, we found that *nouns* – the lexical category most impacted by gender bias – present a firm and huge gender divide over the whole training, where ST systems do not rely on relevant information to support feminine translation and never really adjust its generation.

[17] Although inappropriate in this context, both *amo* and *dueño* are valid mapping to the word *master*.

9

Conclusion

When I started working on gender bias in automatic translation in 2019, the first focused studies on the topic in NLP represented an inspiring but limited niche. Three years later, the topic has gathered a vibrant multidisciplinary community of researchers, with a growing number of studies putting forward novel practices, methods, and resources to address the issue, as well as ways to frame and think about bias in machines. This work contributes to such a community, and marks a step ahead to rise awareness on the need for careful deployment of more equitable technologies. With this chapter, I conclude the thesis. After summarizing the main results and contributions, I discuss the future directions that I envision for the field.

9.1 Results and Contributions

This thesis represents my set of contributions towards the understanding, assessing and mitigating of bias in automatic translation. The contributions and results can be summarized as follows:

- As a **theoretical contribution**, the thesis reviews the state of research on gender bias in automatic translation. I found that the topic had largely remained within the focus of text-to-text MT and a strictly technical perspective. Instead, I underscored that gender bias is a sociotechnical problem. Armed with insights and knowledge from the field of the social sciences and (socio)linguistics, I have explored the relation between language and the extra-linguistic reality of gender, as well as the implications that this relation has on the perception and representation of individuals and communities. On this basis, I have discussed competing – or only partially overlapping – notions of bias, systematized its different conceptualizations in MT, and presented a taxonomy of harms that can ensue. Informed by such taxonomy, I have presented and mapped disparate studies on assessing and mitigating bias to their addressed conceptualization of bias and harm. Also, I put forward the argument that boiling bias down to generically invoked "training data bias" is misguided. Accordingly, I bring forth a classification of different sources of bias in MT, by foregrounding the societal context in which both models and language data are generated and deployed, and examined concurring technical factors for bias, such as annotation practices, algorithmic components and testing procedures. These insights have guided the creation of the resources presented in the thesis, as well as the discussed experiments. It is in my hope that they can provide guidance for other researchers, too.

- As **resource and methodological contributions**, the thesis presents several, manually annotated, training and test data, which are necessary to foster the evaluation and gender-aware development of MT and ST systems:
  - The MuST-SHE benchmark with its POS and agreement extension. We found that generic test sets and metrics are inadequate to monitor gender translation. Existing bias-aware test sets, however, were built on simple, artificial sentence templates, which risks being overfitted and do not reflect the variability of natural language. Hence, I created the multilingual MuST-SHE corpus, built on natural language from TED talks data. MuST-SHE presents several annotation layers to account for a diversified range of gender phenomena. Also, by means of such annotations, MuST-SHE implements a new evaluation protocol, which conciliates i) pinpointed analyses on the correct translation of gendered words, and ii) and overall translation quality assessment, where global scores are made informative in terms of bias by allowing to isolate gender from other factors that might affect overall performance.
  - MuST-Speakers and the Gender-Balanced Validation Set. Data-centric approaches to data mitigation require large datasets for which gender information are available. To supply for their lack, the existing TED-based MuST-C corpus (Cattoni et al., 2021) was enriched with speaker's gender information. Since commonly employed automatic method for gender classification and labeling based on voice detection or proper names risks introducing heteronormative notions of gender and reinforcing stereotypes, a different procedure was followed. Namely, speaker's gender information were

annotated based on the speaker's personal pronouns, which were manually identified and retrieved from their publicly available TED bios. With MuST-Speaker, I was able to obtain gender-annotated data to be used for training gender-aware models. Furthermore, I created the Gender-balanced Validation Set: a development set made of an equal number of TED talks by feminine and masculine speakers, so to avoid rewarding systems' biased behavior at training time.

These resources were made available for three language pairs: English-French, Italian and Spanish, and allowed for experimenting under different types of scenarios and hypotheses in the thesis. They are all released to the research community to promote further advances and investigations on gender bias.

- As **empirical contributions**, the thesis has put to use the newly introduced resources to conduct extensive experiments and evaluations by showing that:
  - Just like their MT counterparts, both cascade and direct ST solutions are affected by gender bias. A comparison of the two ST approaches, however, has unveiled that direct ST leverages speakers' voice as a gender cue to translate speaker-related gender phenomena. While this capability to exploit audio features grants direct ST an apparent advantage for gender translation, relying on biometric features to infer gender can be intrinsically problematic, making ST solutions unsuitable or harmful for a diverse range of users, such as children, vocally impaired people, and transgender individuals. Accordingly, two gender-aware solutions – enhanced with external information about speaker's gender – have been proposed: *i)* a *multi-gender* system, trained on language data annotated with speakers' gender, and *ii)* two *specialized* models, each trained, respectively, on masculine and feminine data only. Although their usefulness has been tested only with regard to speaker-related phenomena, the specialized models offer a viable solution, able to largely improve feminine translation and override speech cues to translate gender based on eternal information.
  - Rather than solely being a "data problem", gender bias can be exacerbated by algorithmic choices. I investigated how word segmentation, a seemingly neutral, and yet crucial component of current systems, can have an impact on gender bias. An extensive comparison of five different segmentation methods in ST has shown that state-of-the-art BPE splitting comes at the cost of higher bias in comparison to character-based segmentation. Accordingly, a hybrid solution leveraging the benefits of both BPE and character-level segmentations was proposed. While there is still plenty of ground to cover regarding *how to split*, this hybrid proposal marks a step toward the integration of

gender concerns for systems' development, rather than accounting for computational efficiency and generic performance, only.

- Larger models with higher generic performance do not grant an advantage for feminine gender translation. A multifaceted manual and automatic evaluation of different ST models at several levels of granularity has reiterated the value of dedicated analyses, able to disentangle gender bias from overall translation quality. Also, the study has consistently unveiled that, while respecting gender agreement is not an issue for ST systems, not all POS are equally impacted by gender bias. Rather, even within natural conditions that are not designed to prompt stereotyping, nouns emerged as the most biased lexical category. Furthermore, ST systems generate a considerable amount of neutral wordings in lieu of gender-marked expressions, which current binary benchmarks fail to recognize. I believe these findings can inform the qualitative design of future data curation practices, mitigating approaches and benchmarks, and pave the way for research on alternative translation strategies beside binary feminine and masculine forms.

- Standard training procedures optimized for overall translation quality are suboptimal for feminine gender translation, which still show potential for improvement when the systems' training is stopped. I carry out for the first time a dynamic analysis that accounts for the emergence of gender translation capabilities over the whole course of ST training. This new perspective examines the trend and stability of gender evolution for different phenomena, showing that unlike overall translation quality, feminine translation emerges more prominently in the late training stages, and does not reach a plateau when the training is stopped. Such trend is however concealed by static evaluations, and and unaccounted for in current training procedures. It remains open to future explorations the extent to which feminine learning has the potential to keep improving. Also, for gender phenomena that can be easily be disambiguated from the sentence context, masculine and feminine curves show a generally parallel and upwards trend, except for nouns. Characterized by flat trends and a huge gender divide, their learning dynamics suggests that ST systems confidently rely on spurious cues and generalize masculine from the very early stages of training onwards. While these analyses are conducted on ST outputs, they nonetheless contribute to shedding light on the opaque ST training process, and can be paired with explainability and probing approaches on systems' inner mechanisms to verify their compatibility with our findings.

9.2 Future Directions

One of the main points made in this thesis is that gender bias in automatic translation should be framed as a sociotechnical issue. In fact, the three-tiered categorization of sources of bias – *pre-existing, technical, and emergent* – emphasizes both social and technical implications (Sec. 3.3). As a result, I acknowledged and started from the assumption that gender bias reflects and is confronted with the historical dynamics and language practices that technologies have come to crystallize. Then, the experimental sections of the thesis, were primarily focused on what I have defined as *technical* sources of bias (e.g., data, annotations, model design, and evaluation) Both social and technical factors have a significant impact on the problem's ongoing evolution, and make the study of gender bias in language technologies, including this work, a snapshot of a moving target. That is because the issue is intrinsically cultural and ever-evolving, and technological advances in the field are happening at a rapid pace.

In the last three years, the emerging direct ST technology has reached the translation quality of the traditional cascaded ST pipeline (Bentivogli et al., 2021), and ever-growing pre-trained models are increasingly taking over development procedures and ancillary research practices (Brown et al., 2020; Le Scao et al., 2022; Lewis et al., 2020; Costa-jussà et al., 2022b). Furthermore, technological advances resulting from massive industrial efforts are incessant, with the conversational model ChatGPT[1] being among the most recent, which displays (multilingual) abilities that have surprised both the research community and the general public. If and how new models emerging day by day can be used to foster bias research is the object of ongoing discoveries, as is their impact on gender and translation. Thus, a clear outline of technical aspects to be addressed for future studies on gender bias in translation technologies is difficult to foresee, as it is contingent and extremely sensitive to what will be the most recent introductions to the field. For this reason, I avoid predictions based on technical aspects and models. Rather, I envision two possible research directions that, I hope, will stand the test of time and retain their value in the face of rapid technological change.

Beyond binary distinctions. Beside a few notable exceptions for English NLP tasks (Cao and Daumé III, 2020; Sun et al., 2021; Vanmassenhove et al., 2021a), and one in MT (Saunders et al., 2020), the study of gender bias has been centered around the binary masculine/feminine dichotomy. Indeed, bias examination has focused on systems' ability to correctly generate feminine and masculine forms in translation, inspected stereotypical representations of men and women, and has also framed bias as the differential between masculine and feminine linguistic representation. This research has been without any doubt crucial to unveil the weaknesses of

[1] Launched by OpenAI in November 2022: https://openai.com/blog/chatgpt/

current systems and to question, perhaps retroactively, the decisions typically incorporated in the design of language technology. As we start to think about *proactive* measures toward the development of more inclusive technology, however, the dichotomous masculine/feminine approach appears limiting. For one, it neglects the recognition of non-binary identities and individuals that do not resort to the masculine/feminine linguistic repertoire (Dev et al., 2021). Additionally, such a binary repertoire does not offer straightforward solutions to confront ambiguous input, or expressions that simply do not entail a single, correct translation.[2] Consider the example sentence "***Doctors** and **nurses** are needed*". In a grammatical gender such as Italian, assuming we do not want to resort to stereotypical associations (e.g., doctors as MASC and nurses as FEM), what would be a desirable translation? A possibility is the use of masculine generics, which are however increasingly considered unsatisfactory solutions (see Sec. 3.3). To confront these scenarios in MT, approaches based on rewriters that offer a double output for simple queries can be deployed (Alhafni et al., 2021), and are actually already integrated for a number of language pairs in commercial systems such as Google Translate (Johnson, 2020b). Yet, the portability of such rewrites is challenged if confronted with multiple gendered entities and thus combinations to be offered in the output (e.g., it-output1: *dottori* MASC *e infermieri* MASC *sono richiesti* MASC; it-output2: *dottoresse* FEM *e infermieri* MASC sono *richiesti* MASC; it-output3: *dottoresse* FEM *e infermiere* FEM sono *richieste* FEM, etc.).

Towards alternative inclusive translation strategies, it is from the realm of non-binary linguistic strategies that a path ahead can be found (López, 2020). On the one hand, Direct Non-binary Language (DNL) relies on neologisms and neomorphemes (Bradley et al., 2019; Papadopoulos, 2019; Knisely, 2020), which aim at increasing the visibility of non-binary individuals, but are also adopted to simply avoid gender-specifications (e.g., the Italian schwa as in *Buongiorno a tuttə*, and potentially for *dottorə e infermierə*). These innovative linguistic forms emerge from grassroot efforts and can be found in more informal communication channels such as social media (Comandini, 2021), but their grammatically and acceptability within standardized form is not yet acknowledged. Instead, Indirect Non-binary Language (INL), while remaining within standard grammar and forms, overcomes gender specifications (e.g., using *service, humankind* rather than *waiter/waitress* or *mankind*). These are largely top-down proposals, some of which have been recently implemented in NLP, such as by Microsoft text editors (Langston, 2020). Whilst more challenging, INL can be achieved also for grammatical gender languages (e.g., *Il personale medico e infermieristico é richiesto*) (Motschenbacher, 2014; Lindqvist et al., 2019), and is endorsed in the official documents of several institutions, such as the EU (Papadimoulis, 2018). Mindful that the acceptability and usefulness of these alternative strategies are inconsistent across languages and their speakers, and that their deployment has to be carefully weighted

[2] Namely, gender phenomena for unspecified referents, as those included in MuST-SHE category 4 in Sec. 4.2.1.

in light of both their scenarios of use and users, future work is needed to explore the integration of such alternative strategies within the cross-lingual task of automatic translation.

Human-in-the-loop. Language technologies are built for people. Yet, research on gender bias in automatic translation is still restricted to lab tests. In fact, unlike other areas that rely on participatory design (Turner et al., 2015; Cercas Curry et al., 2020; Liebling et al., 2020), the advancement of the field is not measured with people's experience in focus or in relation to specific deployment contexts. In particular, representational harms (Crawford, 2017) – such as the potential propagation of stereotypical associations (Sec. 3.2) – are intrinsically difficult to estimate, and available benchmarks only provide a rough idea of their extent. What is the cascaded effect of being exposed to an automatic system that consistently suggests that *top dancers* are men, but *sexy dancers* are women?[3] Can it contribute to the consolidation of prejudices? And how are translators and post-editors affected by biased translation output? Are they primed by biased suggestions, possibly contributing to the generation of future text and data that increasingly incorporate biased associations? In this sense, several studies focusing on stylistic priming in MT can be of guidance (Resende et al., 2020). Also, to the best of my knowledge, mitigating strategies are not being judged based on the diverse needs of a range of users. For instance, it is to be verified whether the above-mentioned rewriters can be of assistance or impediment to the workflow of post-editors, whereas their efficacy might be positively judged by monolingual users. Along the same line, in Chapter 6 of this thesis, I have discussed the trade-off between models based on state-of-the-art segmentation methods that grant higher translation quality, and solutions that instead show an edge for the generation of feminine forms in translation. This trade-off, however, is detected by means of automatic evaluations and framed as a difference in scores, which can be opaque. However, whether a human reader would perceive it as such – and thus motivate treating overall quality and gender representation as two necessarily concurring goals – remains to be ascertained.

Overall, I believe these are fundamental considerations to guide the field forward and to propel the creation of gender-inclusive technology, as Human-Computer Interaction studies show (Vorvoreanu et al., 2019; Keyes, 2018). This motivates focused research on the individual or aggregate effects of bias and technical interventions with dedicated case studies, which allow engaging with the people involved in MT pipelines at all stages.

[3] This example is obtained with commercial MT systems in date 23.02.23.

Bibliography

Emad A. S. Abu-Ayyash. Errors and Non-errors in English-Arabic Machine Translation of Gender-Bound Constructs in Technical Texts. *Procedia Computer Science*, 117:73–80, 2017. doi: https://doi.org/10.1016/j.procs.2017.10.095. URL https://www.sciencedirect.com/science/article/pii/S187705091732152X.

Lauren Ackerman. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1), 2019. doi: http://doi.org/10.5334/gjgl.721.

Alexandra Y. Aikhenvald. Gender meanings in grammar and lexicon. In *How Gender Shapes the World*. Oxford University Press, 08 2016. ISBN 9780198723752. doi: 10.1093/acprof:oso/9780198723752.003.0005. URL https://doi.org/10.1093/acprof:oso/9780198723752.003.0005.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.12.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses, 2021.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.199.

Khaled M. Alhawiti. Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications*, 5(12), 2014.

Tanel Alumäe and Mikko Kurimo. Efficient Estimation of Maximum Entropy Language Models with n-gram Features: an SRILM Extension. In *Proceedings of INTERSPEECH 2010*, pages 1820–1823, Chiba, Japan, 2010.

Antonios Anastasopoulos and David Chiang. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana, 2018.

Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, 2008. URL https://www.wired.com/2008/06/pb-theory/. Accessed: 2021-02-25.

Gavriel Y. Ansara and Peter Hegarty. Misgendering in English language contexts: Applying non-cisgenderist methods to feminist research. *International Journal of Multiple Research Approaches*, 7(2):160–177, 2013. doi: 10.5172/mra.2013.7.2.160. URL https://doi.org/10.5172/mra.2013.7.2.160.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.iwslt-1.1.

Maria Antoniak and David Mimno. Bad Seeds: Evaluating Lexical Methods for Bias Measurement. 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 1889–1904, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148.

Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W. Black. Intent transfer in speech-to-speech machine translation. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 153–158. IEEE, 2012.

Allison M.N. Archer and Cindy D. Kam. She is the chair(man): Gender, language, and leadership. *The Leadership Quarterly*, page 101610, 2022. ISSN 1048-9843. doi: https://doi.

org/10.1016/j.leaqua.2022.101610. URL https://www.sciencedirect.com/science/article/pii/S1048984322000133.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating Discrete Translation Lexicons into Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, 2016.

Duygu Ataman and Marcello Federico. An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/W18-1810.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342, 2017. URL http://ufal.mff.cuni.cz/pbml/108/art-ataman-negri-turchi-federico.pdf.

Duygu Ataman, Orhan Firat, Mattia A. Di Gangi, Marcello Federico, and Alexandra Birch. On the Importance of Word Boundaries in Character-level Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5619. URL https://www.aclweb.org/anthology/D19-5619.

Giuseppe Attanasio, Salvatore Greco, Moreno La Quatra, Luca Cagliero, Michela Tonti, Tania Cerquitelli, and Rachele Raus. E-mimic: Empowering multilingual inclusive communication. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4227–4234, 2021. doi: 10.1109/BigData52589.2021.9671868.

Jenny Audring. Gender, 07 2016. URL https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-43.

David Azul. On the varied and complex factors affecting gender diverse people's vocal situations: Implications for clinical practice. *Perspectives on Voice and Voice Disorders*, 25(2):75–86, 2015.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. A Comparative Study on End-to-end Speech to Text Translation. In *Proceedings of International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore, December 2019a.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. On Using SpecAugment for End-to-End Speech Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, China, November 2019b.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 44–54, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.3. URL https://aclanthology.org/2020.iwslt-1.3.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.

Mona Baker. *A Coursebook on Translation*. Routledge, 1992.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014. URL https://doi.org/10.1111/josl.12080.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2076.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on High-resource Speech Recognition Improves Low-resource Speech-to-text Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1006. URL https://www.aclweb.org/anthology/N19-1006.

Alexis T. Baria and Keith Cross. The brain is a computer is a brain: Neuroscience's internal debate and the social significance of the Computational Metaphor. *ArXiv*, abs/2107.14042, 2021.

Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Naomi S. Baron. A reanalysis of English grammatical gender. *Lingua*, 27:113–140, 1971.

Marco Baroni. Linguistic Generalization and Compositionality in Modern Artificial Neural Networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.

Christine Basta and Marta R. Costa-jussà. Impact of Gender Debiased Word Embeddings in Language Modeling. *CoRR*, abs/2105.00908, 2021. URL https://arxiv.org/abs/2105.00908.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.winlp-1.25. URL https://www.aclweb.org/anthology/2020.winlp-1.25.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1z-PsR5KX.

Rachel Bawden, Guillaume Wisniewski, and Hélène Maynard. Investigating Gender Adaptation for Speech Translation. In *Proceedings of the 23ème Conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 490–497, Paris, FR, July 2016. URL https://hal.archives-ouvertes.fr/hal-01353860.

Simone de Beauvoir. *Le deuxième sexe*. Paris: Gallimard, 1949.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*, 46(1):1–52, 2020. doi: 10.1162/coli_a_00367. URL https://doi.org/10.1162/coli_a_00367.

Olfa Belkahla Driss, Sehl Mellouli, and Zeineb Trabelsi. From citizens to government policy-makers: Social media data analysis. *Government Information Quarterly*, 36(3):560–570, 2019. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2019.05.002. URL https://www.sciencedirect.com/science/article/pii/S0740624X18302983.

Emily M. Bender. Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-0106.

Emily M. Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/, 2019a. Accessed: 2021-02-25.

Emily M. Bender. A Typology of Ethical Risks in Language Technology with an Eye Towards where Transparent Documentation might help. In *CRAASH. The future of Artificial Intelligence: Language, Ethics, Technology*, Cambridge, UK, 2019b. URL http://www.crassh.cam.ac.uk/gallery/video/emily-m.-bender-washington-a-typology-of-ethical-risks-in-language-technology.

Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://www.aclweb.org/anthology/Q18-1041.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009. ISSN 1935-8237. doi: 10.1561/2200000006. URL http://dx.doi.org/10.1561/2200000006.

Ruha Benjamin. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK, 2019.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1025. URL https://www.aclweb.org/anthology/D16-1025.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 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, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.619.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.224. URL https://aclanthology.org/2021.acl-long.224.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1015.

Victoria L. Bergvall, Janet M. Bing, and Alice F. Freed. *Rethinking Language and Gender Research: Theory and Practice*. Addison Wesley Longman, London, UK, 1996. URL https://doi.org/10.4324/9781315842745.

Camiel J. Beukeboom and Christian Burgers. How Stereotypes are shared through Language: A Review and Introduction of the Social Categories and Stereotypes Communication (SCSC) Framework. *Review of Communication Research*, 7:1–37, 2019. URL https://nbn-resolving.org/urn:nbn:de:0168-ssoar-61187-9.

Claudia Bianchi. Slurs and appropriation: An echoic account. *Journal of Pragmatics*, 66:35–44, 2014.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Underlying Values of Machine Learning Research. In *Resistance AI Workshop @ NeurIPS*, Online, December 2020. URL https://drive.google.com/file/d/1tjrm3Bf1hxV8iuPSiCcM1IazITGp-GZj/view.

Dagmar Bittner. Gender classification and the inflectional system of german nouns. *Gender in Grammar and Cognition*, 124:1–25, 2000.

Su Lin Blodgett. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Doctoral Dissertations. 2092., 2021. doi: https://doi.org/10.7275/20410631.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. 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 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

Leonard Bloomfield. Notes on the Fox language. *International Journal of American Linguistics*, 4(2/4):181–219, 1927.

Leonard Bloomfield. *Language*. Holt, Rinehart and Winston, New York, USA, 1933.

Gerd Bohner. Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, 40(4):515–529, 2001.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, October 2018.

Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, volume 29, pages 4349–4357, Barcelona, ES, 2016. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

David Bourguignon, Vincent Y. Yzerbyt, Catia P. Teixeira, and Ginette Herman. When does it hurt? Intergroup permeability moderates the link between discrimination and self-esteem. *European Journal of Social Psychology*, 45(1):3–9, 2 2015. ISSN 0046-2772. doi: 10.1002/ejsp.2083. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2083.

Evan D. Bradley, Julia Salkind, Ally Moore, and Sofi Teitsort. Singular 'they' and novel pronouns: gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America*, 4(1):36–1, 2019. URL https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4542.

Friederike Braun. *Geschlecht im Türkischen: Untersuchungen zum sprachlichen Umgang mit einer sozialen Kategorie*. Turcologica Series. Otto Harrassowitz Verlag, Wiesbaden, DE, 2000. URL https://www.harrassowitz-verlag.de/title_290.ahtml.

Richard Brooks. Google Translate Facts You Should Know. The Language Blog, 2016. URL https://www.k-international.com/blog/google-translate-facts/#:~:text=Google%20Translate%20translates%20more%20than,128%2C000%20Bibles%2C%20every%20single%20day. Accessed: 2022-02-10.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sheila Brownlow, Julie A. Rosamond, and Jennifer A. Parker. Gender-linked Linguistic Behavior in Television Interviews. *Sex Roles*, 49(3-4):121–132, August 2003. URL https://doi.org/10.1023/A:1024404812972.

Janina Brutt-Griffler and Sumi Kim. In their own voices: Development of English as a gender-neutral language: Does learning English promote gender equity among Asian international students? *English Today*, 34(1):12–19, 2018. doi: 10.1017/S0266078417000372.

Mary Bucholtz and Kira Hall. Language and identity. *A companion to linguistic anthropology*, 1:369–394, 2004.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, USA, February 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Judith Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, USA, March 1990. URL https://www.routledge.com/Gender-Trouble-Feminism-and-the-Subversion-of-Identity/Butler/p/book/9780415389556.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, December 2016.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics Derived Automatically from Language Corpora contain Human-like Biases. *Science*, 356(6334):183–186, 2017. URL https://science.sciencemag.org/content/356/6334/183.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E06-1032.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0718.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the third workshop on statistical machine translation*, pages 70–106, 2008.

Deborah Cameron. Gender Issues in Language Change. *Annual Review of Applied Linguistics*, 23:187–201, March 2003. doi: 10.1017/S0267190503000266. URL https://doi.org/10.1017/S0267190503000266.

Kathryn Campbell-Kibler. The Nature of Sociolinguistic Perception. *Language Variation and Change*, 21(1):135–156, 2009.

Alex Campolo, Madelyn R. Sanfilippo, Meredith Whittaker, and Kate Crawford. AI Now Report 2017. *New York: AI Now Institute*, 2017. URL https://experts.illinois.edu/en/publications/ai-now-2017-report.

Yang T. Cao and Hal Daumé III. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.418. URL https://www.aclweb.org/anthology/2020.acl-main.418.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

Michael Castelle. The Linguistic Ideologies of Deep Abusive Language Classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5120. URL https://aclanthology.org/W18-5120.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. Approaches to human and machine translation quality assessment. In *Translation Quality Assessment*, pages 9–38. Springer, 2018.

Sheila Castilho, Maja Popović, and Andy Way. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, FR, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.461.

Roldano Cattoni, Mattia A. Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155, 2021. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2020.101155.

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.7.

Lori Chamberlain. Gender and the Metaphorics of Translation. *Signs: Journal of Women in Culture and Society*, 13(3):454–472, 1988. URL https://doi.org/10.1086/494428.

Kai-Wei Chang. Bias and Fairness in Natural Language Processing, 2019. URL http://web.cs.ucla.edu/~kwchang/documents/slides/emnlp19-fairNLP-part2.pdf. Tutorial at the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. Charge-Based Prison Term Prediction with Deep Gating Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, 2019.

Jenny Cheshire. Sex and Gender in Variationist Research. *The Handbook of Language Variation and Change*, pages 423–443, 2004.

Marina Chini. *Genere grammaticale e acquisizione: Aspetti della morfologia nominale in italiano L2*. Franco Angeli, 1995.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 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, IT, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3824. URL https://www.aclweb.org/anthology/W19-3824.

John W. Chotlos. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75, 1944.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1160. URL https://www.aclweb.org/anthology/P16-1160.

Kenneth Church. A Pendulum Swung too Far. *Linguistic Issues in Language Technology*, 6, 2011.

Chloe Ciora, Nur Iren, and Malihe Alikhani. Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, 2021.

Aleksandra Cislak, Magdalena Formanowicz, and Tamar Saguy. Bias Against Research on Gender Bias. *Scientometrics*, 115(1):189–200, April 2018. URL https://doi.org/10.1007/s11192-018-2667-0.

Gloria Comandini. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. : Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, 23:43–64, dic. 2021. URL https://testoesenso.it/index.php/testoesenso/article/view/524.

Bernard Comrie. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago press, 1989.

Bernard Comrie. Grammatical Gender Systems: A Linguist's Assessment. *Journal of Psycholinguistic Research*, 28:457–466, September 1999. doi: https://doi.org/10.1023/A:1023212225540.

Bernard Comrie. Grammatical gender and personification. In *Perspectives on language and language development*, pages 105–114. Springer, 2005.

Kirby Conrod. *Pronouns Raising and Emerging*. PhD thesis, University of Washington, 2019.

Kirby Conrod. Pronouns and Gender in Language. *The Oxford Handbook of Language and Sexuality*, 2020. URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190212926.001.0001/oxfordhb-9780190212926-e-63.

Ellen Contini-Morava and Marcin Kilarski. Functions of nominal classification. *Language sciences*, 40:263–299, 2013.

Greville G. Corbett. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK, 1991. URL https://doi.org/10.1017/CBO9781139166119.

Greville G. Corbett. *Agreement*. Cambridge University Press, 2006.

Greville G. Corbett. *The Expression of Gender*. De Gruyter Mouton, Berlin, DE, 2013a. URL https://doi.org/10.1515/9783110307337.

Greville G. Corbett. Systems of gender assignment. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL https://wals.info/chapter/32.

Greville G. Corbett. *The Expression of Gender*. De Gruyter, 2013c. doi: doi.org/10.1515/9783110307337.

Greville G. Corbett. Hybrid nouns and their complexity. *Agreement from a diachronic perspective*, pages 191–214, 2015.

Guillem Cortès Sebastià. Towards Robust End-to-End Speech Translation. Master's thesis, Universitat Politècnica de Catalunya, 2020. URL https://upcommons.upc.edu/bitstream/handle/2117/330400/GuillemCortes-Towards_Robust_End_to_End_Speech_Translation.pdf.

Marta R. Costa-jussà. From feature to paradigm: deep learning in machine translation. *Journal of Artificial Intelligence Research*, 61:947–974, 2018.

Marta R. Costa-jussà. An analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1:495–496, 2019. URL https://www.nature.com/articles/s42256-019-0105-5.

Marta R. Costa-jussà and Adrià de Jorge. Fine-tuning Neural Machine Translation on Gender-Balanced Datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.3.

Marta R. Costa-jussà and José A. R. Fonollosa. Character-based Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2058. URL https://www.aclweb.org/anthology/P16-2058.

Marta R. Costa-jussà, Pau Li Lin, and Cristina España-Bonet. GeBioToolkit: Automatic Extraction of Gender-Balanced Multilingual Corpus of Wikipedia Biographies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4081–4088, Marseille, FR, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.502.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.230.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022b.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting Gender Bias in Neural Machine Translation: The Multilingual Architecture Matters. *Accepted in 36th AAAI Conference on Artificial Intelligence*, 2022c.

Nikolas Coupland. *Style: Language variation and identity*. Cambridge University Press, 2007.

Michael A. Covington and Joe D. McFall. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6:389–407, 2020.

Kate Crawford. The Trouble with Bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, California, dec 2017. URL https://www.youtube.com/watch?v=fMym_BKWQzk.

Kate Crawford, Mary L. Gray, and Kate Miltner. Critiquing Big Data: Politics, ethics, epistemology. *International Journal of Communication*, 8:10, 2014.

Mathias Creutz and Krista Lagus. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. *International Symposium on Computer and Information Sciences*, 2005.

Caroline Criado-Perez. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Penguin Random House, London, UK, 2019. ISBN 9781784742928. URL https://books.google.it/books?id=A9AEugEACAAJ.

Anna Currey, Maria Nadejde, Raghavendra Pappagari, Mia Mayer, Stanislas LAULY, Xing Niu, Benjamin Hsu, and Georgiana Dinu. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *EMNLP 2022*, 2022. URL https://www.amazon.science/publications/mt-geneval-a-counterfactual-and-contextual-dataset-for-evaluating-gender-accuracy-in-machine-translation.

Anne Curzan. *Gender Shifts in the History of English*. Cambridge University Press, Cambridge, UK, 2003. URL https://doi.org/10.1017/CBO9780511486913.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021. doi: 10.1162/tacl_a_00425. URL https://aclanthology.org/2021.tacl-1.74.

Östen Dahl. *The Growth and Maintenance of Linguistic Complexity*, volume 10. John Benjamins Amsterdam, 2004.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 2018. Accessed: 2021-02-25.

Steven Davis and Paul Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.150. URL https://aclanthology.org/2021.emnlp-main.150.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Online, dec 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, USA, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://www.aclweb.org/anthology/N19-1202.

Mattia A. Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. Robust Neural Machine Translation for Clean and Noisy Speech Transcripts. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2019b.

Mattia A. Di Gangi, Matteo Negri, Roldano Cattoni, Dessi Roberto, and Marco Turchi. Enhancing transformer for end-to-end speech-to-text translation. In *Machine Translation Summit XVII*, pages 21–31, Dublin, Ireland, August 2019c. European Association for Machine Translation.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. Adapting Transformer to End-to-end Spoken Language Translation. In *Proceedings of INTERSPEECH*, pages 1133–1137, Graz, Austria, September 2019d.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. One-To-Many Multilingual End-to-end Speech Translation. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592, Sentosa, Singapore, December 2019e.

Mattia A. Di Gangi, Marco Gaido, Matteo Negri, and Marco Turchi. On Target Segmentation for Direct Speech Translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 137–150, Online, October 2020. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/2020.amta-research.13.

Bruna Di Sabato and Antonio Perri. Grammatical gender and translation: A cross-linguistic overview. In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*. Routledge, New York, USA, 2020. URL https://www.taylorfrancis.com/chapters/grammatical-gender-translation-bruna-di-sabato-antonio-perri/e/10.4324/9781315158938-32?context=ubx&refId=e46af1d4-21f3-4914-9b9f-e7f9435adf45.

Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26 (3):297–302, 1945.

Catherine D'Ignazio and Lauren F. Klein. *Data feminism*. MIT Press, London, UK, 2020. URL https://mitpress.mit.edu/books/data-feminism.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 314–331, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://www.aclweb.org/anthology/2020.emnlp-main.23.

Robert M. W. Dixon. *'Where Have All the Adjectives Gone?' and other Essays in Semantics and Syntax*. Mouton de Gruyter, Berlin, 1982.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. Transferring Emphasis in Speech Translation Using Hard-Attentional Neural Network Models. In *INTERSPEECH*, pages 2533–2537, 2016.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

Stephen Doherty. Translation Technology Evaluation Research. In *The Routledge Handbook of translation and technology*, pages 339–353. Routledge, 2019.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.5. URL https://aclanthology.org/2021.gebnlp-1.5.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Matthias Eck and Chiori Hori. Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the 2nd International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, Pennsylvania, 2005.

Penelope Eckert. The problem with binaries: Coding for gender and sexuality. *Language and Linguistics Compass*, 8(11):529–535, 2014.

Penelope Eckert. Age as a sociolinguistic variable. *The handbook of sociolinguistics*, pages 151–167, 2017.

Penelope Eckert and Sally McConnell-Ginet. *Language and Gender*. Cambridge University Press, Cambridge, UK, 2013. URL https://doi.org/10.1017/CBO9781139245883.

Penelope Eckert and Etienne Wenger. Communities of practice in sociolinguistics: What is the role of power in sociolinguistic variation? *Journal of Sociolinguistics*, 9(4):582–589, 2005.

Mohamed Elaraby and Ahmed Zahran. A Character Level Convolutional BiLSTM for Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 274–278, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4636. URL https://www.aclweb.org/anthology/W19-4636.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. Gender-Aware Spoken Language Translation Applied to English-Arabic. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6, Algiers, Algeria, 2018. URL https://ieeexplore.ieee.org/document/8374387.

Carolyn Epple. Coming to Terms with Navajo Nádleehí: A Critique of Berdache, "Gay", "Alternate Gender", and "Two-spirit". *American Ethnologist*, 25(2):267–290, 1998. doi: https://doi.org/10.1525/ae.1998.25.2.267. URL https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/ae.1998.25.2.267.

Joel Escudé Font and Marta R. Costa-jussà. 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, IT, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3821. URL https://www.aclweb.org/anthology/W19-3821.

Sarah Evans, Nick Neave, Delia Wakelin, and Colin Hamilton. The relationship between testosterone and vocal frequencies in human males. *Physiology & Behavior*, 93(4-5):783–788, 2008.

Anne Fausto-Sterling. Gender/Sex, Sexual Orientation, and Identity Are in the Body: How Did They Get There? *The Journal of Sex Research*, 56(4-5):529–555, 2019. URL https://doi.org/10.1080/00224499.2019.1581883.

Andrew Feenberg. Ten paradoxes of technology. *Techné: Research in Philosophy and Technology*, 14(1):3–15, 2010. doi: 10.5840/techne20101412.

Christiane Fellbaum. Large-scale lexicography in the Digital Age. *International Journal of Lexicography*, 27(4):378–395, 2014.

Lorenzo Ferrone and Fabio Massimo Zanzotto. Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI*, 6:153, 2020.

John Rupert Firth. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, 1957.

James L. Fitch and Anthony Holbrook. Modal vocal fundamental frequency of young adults. *Archives of Otolaryngology*, 92(4):379–382, 1970.

Mary Flanagan. Error Classification for MT Evaluation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994.

Luciano Floridi. Big Data and their Epistemological Challenge. *Philosophy & Technology*, 25 (4):435–437, 2012.

Mikel L. Forcada. Making sense of neural machine translation. *Translation spaces*, 6(2): 291–309, 2017.

Anke Frank, Christiane Hoffmann, and Maria Strobel. Gender Issues in Machine Translation. *University of Bremen*, 2004. URL http://static.lingenio.de/Publikationen/GIST.pdf.

Francine Wattmann Frank. Language planning and sexual equality: Guidelines for non-sexist usage. In *Sprachwandel und feministische Sprachpolitik: Internationale Perspektiven*, pages 231–254. Springer, 1985.

Stella Frank, Desmond Elliott, and Lucia Specia. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413, 2018. URL http://eprints.whiterose.ac.uk/132085/.

Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be Guilty but References are not Innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, 2020.

Batya Friedman and Helen Nissenbaum. Bias in Computer Systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, July 1996. URL https://doi.org/10.1145/230538.230561.

Susanne Fuchs and Martine Toda. Do differences in male versus female/s/reflect biological or sociophonetic factors. *Turbulent sounds: An interdisciplinary guide*, 21:281–302, 2010.

Ute Gabriel, M. Pascal Gygax, Oriane Sarrasin, Alan Garnham, and Jane Oakhill. Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior research methods*, 40(1):206–212, 2008.

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

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, Mauro Cettolo, and Marco Turchi. Contextualized Translation of Automatically Segmented Speech. In *Proceedings of Interspeech 2020*, 2020a.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.8. URL https://www.aclweb.org/anthology/2020.iwslt-1.8.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Breeding Gender-aware Direct Speech Translation Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Online, December 2020c. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.350. URL https://www.aclweb.org/anthology/2020.coling-main.350.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. How to Split: the Effect of Word Segmentation on Gender Bias in Speech Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.313. URL https://aclanthology.org/2021.findings-acl.313.

Marco Gaido, Matteo Negri, and Marco Turchi. Direct Speech-to-Text Translation Models as Students of Text-to-Text Models. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1), 2022.

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

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-Of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, IT, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1339.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, page 3–9, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369176. doi: 10.1145/3347449.3357480. URL https://doi.org/10.1145/3347449.3357480.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender Representation in Open Source Speech Resources. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6599–6605, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.813.

Alan Garnham, Jane Oakhill, and David Reynolds. Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory & Cognition*, 30(3):439–446, 2002.

Timnit Gebru. Race and gender. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford Handbook Online, 2020. doi: 10.1093/oxfordhb/9780190067397.013.16. URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190067397.001.0001/oxfordhb-9780190067397-e-16.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. A Convolutional Encoder Model for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Marylou Pausewang Gelfer and Victoria A. Mikos. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of voice*, 19(4):544–554, 2005.

Marylou Pausewang Gelfer and Kevin J. Schofield. Comparison of acoustic and perceptual measures of voice in male-to-female transsexuals perceived as female versus those perceived as male. *Journal of voice*, 14(1):22–33, 2000.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, 2018.

Vera Gheno. *Femminili Singolari: il femminismo è nelle parole*. Effequ, 2019.

Jesús Giménez and Lluís Màrquez. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3):209–240, 2010.

Lisa Gitelman. *Raw data is an oxymoron*. MIT press, 2013. URL https://mitpress.mit.edu/books/raw-data-oxymoron.

GLAAD. Transgender people. In *Media Reference Guide*. 11th edition, 2007. URL https://www.glaad.org/reference/transgender.

Fiona Glen and Karen Hurrell. Measuring gender identity. https://www.equalityhumanrights.com/sites/default/files/technical_note_final.pdf, 2012. Accessed: 2021-02-25.

Bruce Glymour and Jonathan Herington. Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 269–278, New York, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287573. URL https://doi.org/10.1145/3287560.3287573.

Kevin Gold. Norvig vs. Chomsky and the Fight for the Future of AI, TOR, 2011.

Yoav Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias. 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 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

Kirsten Gomard. The (Un)equal Treatment of Women in Language: a Comparative Study of Danish, English, and German. *Working Papers on Language, Gender and Sexism*, 5(1):5–25, 1995.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://www.aclweb.org/anthology/N19-1061.

Hila Gonen and Kellie Webster. Automatically Identifying Gender Issues in Machine Translation using Perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.180. URL https://www.aclweb.org/anthology/2020.findings-emnlp.180.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67, 2019.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Type B Reflexivization as an Unambiguous Testbed for Multilingual Multi-Task Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, 2020.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Mark Graham. Big data and the end of theory? The Guardian, 2012. URL http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory. Accessed: 2021-02-25.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, 2013.

Joseph H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113, 1963.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of personality and social psychology*, 74(6):1464, 1998.

Jacob Grimm. Deutsche Grammatik. *Gütersloh: Druck und Verlag von C. Bertelsmann*, 1890.

Jeffrey Grogger. Speech Patterns and Racial Wage Inequality. *Journal of Human resources*, 46 (1):1–25, 2011.

Liane Guillou. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, FR, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-3001.

Deborah Günzburger. An acoustic analysis and some perceptual data concerning voice change in male-female trans-sexuals. *International Journal of Language & Communication Disorders*, 28(1):13–21, 1993.

Suchin Gururangan, Dallas Card, Sarah K. Drier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. *arXiv preprint arXiv:2201.10474*, 2022.

Steven Gutstein, Olac Fuentes, and Eric Freudenthal. Knowledge Transfer in Deep Convolutional Neural Nets. *International Journal on Artificial Intelligence Tools*, 17(03):555–567, 2008. doi: 10.1142/S0218213008004059.

Pascal M. Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. Generically Intended, but Specifically Interpreted: When Beauticians, Musicians and Mechanics are all Men. *Language and Cognitive Processes*, 23:464–485, April 2008. doi: 10.1080/01690960701702035.

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 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, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01604. URL https://www.frontiersin.org/article/10.3389/fpsyg.2019.01604.

Nizar Habash, Houda Bouamor, and Christine Chung. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3822. URL https://www.aclweb.org/anthology/W19-3822.

Philipp Hacker. Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law. *Common market law review*, 55(4): 1143–1185, 2018. URL https://ssrn.com/abstract=3164973.

Kira Hall and Veronica O'Donovan. Shifting gender positions among Hindi-speaking hijras. *Rethinking language and gender research: Theory and practice*, pages 228–266, 2014. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-85077615829&partnerID=MN8TOARS.

Foad Hamidi, Morgan K. Scheuerman, and Stacy M. Branham. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173582. URL https://doi.org/10.1145/3173574.3173582.

Mykol C. Hamilton. Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex roles*, 19(11-12):785–799, 1988. URL https://doi.org/10.1007/BF00288993.

Mykol C. Hamilton. Masculine Bias in the Attribution of Personhood: People = Male, Male = People. *Psychology of Women Quarterly*, 15(3):393–402, 1991. URL https://doi.org/10.1111/j.1471-6402.1991.tb00415.x.

Charlie Hancock. Meta's AI Chatbot Repeats Election and Anti-Semitic Conspiracies, 2022. URL https://www.bloomberg.com/news/articles/2022-08-08/meta-s-ai-chatbot-repeats-election-and-anti-semitic-conspiracies. Accessed: 2022-08-13.

Alex Hanna, Andrew Smart, Ben Hutchinson, Christina Greer, Emily Denton, Margaret Mitchell, Oddur Kjartansson, and Parker Barnes. Towards Accountability for Machine Learning Datasets. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 560–575, Online, March 2021. ACM. URL https://dl.acm.org/doi/10.1145/3442188.3445918.

Christian Hardmeier and Marcello Federico. Modelling Pronominal Anaphora in Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289, Paris, FR, 2010. URL https://www.isca-speech.org/archive/iwslt_10/slta_283.html.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. How to Write a Bias Statement: Recommendations for Submissions to the Workshop on Gender Bias in NLP. *arXiv preprint arXiv:2104.03026*, 2021. URL https://arxiv.org/abs/2104.03026.

Thomas Wadleigh Harvey. *A practical grammar of the English language*. American Book Company, 1878.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.275. URL https://www.aclweb.org/anthology/2020.acl-main.275.

Jeffrey Heath. Some Functional Relationships in Grammar. *Language*, 51(1):89–104, 1975. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/413151.

Kevin Heffernan. Evidence from hnr that/s/is a social marker of gender. *Toronto Working Papers in Linguistics*, 23, 2004.

Marlis Hellinger and Hadumond Bußman. *Gender across Languages: The linguistic representation of women and men*, volume 1. John Benjamins Publishing, Amsterdam, NL, 2001. URL https://doi.org/10.1075/impact.9.

Marlis Hellinger and Hadumond Bußman. *Gender across Languages: The linguistic representation of women and men*, volume 2. John Benjamins Publishing, Amsterdam, NL, 2002. URL https://doi.org/10.1075/impact.10.

Marlis Hellinger and Hadumond Bußman. *Gender across Languages: The linguistic representation of women and men*, volume 3. John Benjamins Publishing, Amsterdam, NL, 2003. URL https://doi.org/10.1075/impact.11.

Marlis Hellinger and Heiko Motschenbacher. *Gender Across Languages. The Linguistic Representation of Women and Men*, volume 4. John Benjamins, Amsterdam, NL, 2015. URL https://doi.org/10.1075/impact.36.

Anita Louise Henderson. *Is your money where your mouth is? Hiring managers' attitudes toward African-American Vernacular English*. University of Pennsylvania, 2001.

Lisa A. Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Model. In *Proceedings of the European*

Conference on Computer Vision (ECCV), pages 740–755, Munich, DE, 2018. doi: https://doi.org/10.1007/978-3-030-01219-9_47.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In *Proceedings of the Speech and Computer - 20th International Conference (SPECOM)*, pages 198–208, Leipzig, Germany, September 2018. Springer International Publishing. ISBN 9783319995793. URL http://dx.doi.org/10.1007/978-3-319-99579-3_21.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *Proceedings of NIPS Deep Learning and Representation Learning Workshop*, Montréal, Canada, December 2015. URL http://arxiv.org/abs/1503.02531.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3802. URL https://aclanthology.org/W19-3802.

Charles F. Hockett. *A Course in Modern Linguistics*. Macmillan, New York,NY, US, 1958.

Janet Holmes and Miriam Meyerhoff. *The Handbook of Language and Gender*. Blackwell Publishing Ltd, Malden, USA, 2003. URL https://doi.org/10.1002/9780470756942.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Levi C. R. Hord. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 3(1): 4, 2016. URL https://ir.lib.uwo.ca/wpl_clw/vol3/iss1/4.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.11. URL https://aclanthology.org/2020.nlpcovid19-2.11.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in Natural Language Processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

Dirk Hovy and Shannon L. Spruit. 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, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://www.aclweb.org/anthology/P16-2096.

Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL https://aclanthology.org/2021.naacl-main.49.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User Review Sites as a Resource for Large-Scale Sociolinguistic Studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 452–461, Geneva, CH, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741141. URL https://doi.org/10.1145/2736277.2741141.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing"*, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-16. URL https://aclanthology.org/W17-1600.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.154. URL https://www.aclweb.org/anthology/2020.acl-main.154.

Eduard Hovy, Magaret King, and Andrei Popescu-Belis. An introduction to MT evaluation. In *Proceedings of Machine Translation Evaluation: Human Evaluators meet Automated Metrics. Workshop at the LREC 2002 Conference. Las Palmas, Spain*, pages 1–7, 2002.

Ieuan A. Hughes, John D. Davies, Trevor I. Bunch, Vickie Pasterski, Kiki Mastroyannopoulou, and Jane MacDougall. Androgen Insensitivity Syndrome. *The Lancet*, 9851(380):1419–1428, 2012. doi: https://doi.org/10.1016/S0140-6736(12)60071-3.

John W. Hutchins. Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445. Elsevier, 1995.

Janet S. Hyde. The Gender Similarities Hypothesis. *American psychologist*, 60(6):581–592, 2005. URL https://doi.org/10.1037/0003-066X.60.6.581.

Muhammad Hasan Ibrahim. *Grammatical Gender. Its Origin and Development.* De Gruyter Mouton, The Hague, 1973. ISBN 9783110905397. doi: doi:10.1515/9783110905397. URL https://doi.org/10.1515/9783110905397.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual End-to-End Speech Translation. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577, December 2019.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. ESPnet-ST: All-in-One Speech Translation Toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.34. URL https://www.aclweb.org/anthology/2020.acl-demos.34.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. Data Efficient Direct Speech-to-Text Translation with Modality Agnostic Meta-Learning. *arXiv preprint arXiv:1911.04283*, 2019.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7904–7908. IEEE, 2020.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Giménez. Adrià, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, Barcelona, Spain, May 2020. URL https://ieeexplore.ieee.org/document/9054626.

Julia Ive, Pranava Madhyastha, and Lucia Specia. Distilling Translations with Visual Awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, IT, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1653. URL https://www.aclweb.org/anthology/P19-1653.

Abigail Z. Jacobs. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 375–385, New York, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Abigail Z. Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. The Meaning and Measurement of Bias: Lessons from Natural Language Processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 706, New York, USA, 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3351095.3375671.

Daiane Dordete Steckert Jacobs. Vocal Body, gender and performance. *Revista Brasileira de Estudos da Presença*, 7:359–381, 2017.

Roman Jakobson. On Linguistic Aspects of Translation. In Reuben A. Brower, editor, *On translation*, pages 232–239. Harvard University Press, Cambridge, USA, 1959. URL https://web.stanford.edu/~eckert/PDF/jakobson.pdf.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China, July 2015a. Association for Computational Linguistics. doi: 10.3115/v1/P15-1001. URL https://aclanthology.org/P15-1001.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September 2015b. Association for Computational Linguistics. doi: 10.18653/v1/W15-3014. URL https://aclanthology.org/W15-3014.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184, Brighton, UK, May 2019.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.273. URL https://aclanthology.org/2021.findings-acl.273.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual Syntactic Variation over Age and Gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, CN, 2015. URL https://www.aclweb.org/anthology/K15-1011.

Joy L. Johnson, Robin Repta, John L. Oliffe, and Lorraine Greaves. Designing and conducting gender, sex and health research. *Thousand Oaks*, 18(6-7):345–355, 2012.

Kari Johnson. AI Weekly: A deep learning pioneer's teachable moment on AI bias. https://venturebeat.com/2020/06/26/ai-weekly-a-deep-learning-pioneers-teachable-moment-on-ai-bias/, 2020a. Accessed: 2021-02-25.

Melvin Johnson. A Scalable Approach to Reducing Gender Bias in Google Translate. https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html, 2020b. URL https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html. Accessed: 2021-02-25.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL https://www.aclweb.org/anthology/Q17-1024.

Karen Sparck Jones. *Natural Language Processing: A Historical Review*, pages 3–16. Springer Netherlands, Dordrecht, 1994. ISBN 978-0-585-35958-8. doi: 10.1007/978-0-585-35958-8_1. URL https://doi.org/10.1007/978-0-585-35958-8_1.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2009. URL https://aclanthology.org/P17-2009.

Lukasz Kaiser, Aidan N. Gomez, and Francois Chollet. Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*, 2017.

Nal Kalchbrenner and Phil Blunsom. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–

1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1176.

Cora Kaplan and David Glover. *Genders*. Routledge, London, UK, 2001.

Alina Karakanta, Sara Papi, Matteo Negri, and Marco Turchi. Simultaneous Speech Translation for Live Subtitling: from Delay to Display. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, pages 35–48, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-asltrw.4.

Anne Karpf. *The human voice: How this extraordinary instrument reveals essential clues about who we are*. Bloomsbury Publishing USA, 2006.

Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), November 2018. doi: 10.1145/3274357. URL https://doi.org/10.1145/3274357.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimed Tools Appl*, 2022. doi: https://doi.org/10.1007/s11042-022-13428-4.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Scott F. Kiesling. *Language, Gender, and Sexuality: An introduction*. Routledge, 2019.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. When and Why is Document-level Context Useful in Neural Machine Translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6503. URL https://aclanthology.org/D19-6503.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations,* *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Svetlana Kiritchenko and Saif Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, 2018. doi: 10.18653/v1/S18-2005. URL https://www.aclweb.org/anthology/S18-2005.

Kris Aric Knisely. Le français non-binaire: Linguistic forms used by non-binary speakers of French. *Foreign Language Annals*, 53(4):850–876, 2020. doi: https://doi.org/10.1111/flan.12500.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan, May 2018. URL https://www.aclweb.org/anthology/L18-1001.

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

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86, Phuket, TH, 2005. AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *First Workshop on Neural Machine Translation*, pages 28–39. Association for Computational Linguistics, 2017.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial Disparities in Automated Speech Recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689, 2020.

Toshi Konishi. The Semantics of Grammatical Gender: A Cross-Cultural Study. *Journal of psycholinguistic research*, 22(5):519–534, 1993.

Corina Koolen and Andreas van Cranenburgh. These are not the Stereotypes You are Looking for: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, ES, 2017.

Association for Computational Linguistics. doi: 10.18653/v1/W17-1602. URL https://www.aclweb.org/anthology/W17-1602.

Cheris Kramarae and Paula A. Treichler. *A Feminist Dictionary*. Pandora Press, London, UK, 1985.

Michał Krawczyk. Are all Researchers Male? Gender Misattributions in Citations. *Scientometrics*, 110(3):1397–1402, 2017. URL https://doi.org/10.1007/s11192-016-2192-y.

Jody Kreiman and Diana Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

Hamutal Kreiner, Patrick Sturt, and Simon Garrod. Processing Definitional and Stereotypical Gender in Reference Resolution: Evidence from Eye-Movements. *Journal of Memory and Language*, 58:239–261, 02 2008. doi: 10.1016/j.jml.2007.09.003.

James Kuczmarski. Reducing Gender Bias in Google Translate. https://www.blog.google/products/translate/reducing-gender-bias-google-translate/, 2018. Accessed: 2021-02-25.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Mascha Kurpicz-Briki and Tomaso Leoni. A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. *Frontiers in big Data*, 4:20, 2021.

William Labov. *The social stratification of English in New York City*. Cambridge University Press, 1966.

William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, 1972.

Robin Lakoff. Language and woman's place. *Language in society*, 2(1):45–79, 1972.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

Jennifer Langston. New AI tools help writers be more clear, concise and inclusive in Office and across the Web. https://blogs.microsoft.com/ai/microsoft-365-ai-tools/, 2020. Accessed: 2021-02-25.

Hugo Larochelle and Geoffrey E. Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. *Advances in neural information processing systems*, 23, 2010.

Meghan Beth Larose. *Trans* Vocal: Documenting Gender Subjectivity Through Changing Vocality*. PhD thesis, Carleton University, 2022.

Brian Larson. Gender as a variable in Natural-PLanguage Processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://www.aclweb.org/anthology/W17-1601.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics, 2018.

Anne Lauscher, Archie Crowley, and Dirk Hovy. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.105.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.103. URL https://aclanthology.org/2021.acl-short.103.

Ronan Le Nagard and Philipp Koehn. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, pages 252–261, Uppsala, SE, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-1737.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A

176b-parameter open-access multilingual language model. *arXiv e-prints*, pages arXiv–2211, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Peter Lee. Learning from Tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00000gjdpwwcfcus11t6oo6dw79gw, 2016. Accessed: 2022-08-24.

Winfred P. Lehmann. On earlier stages of the Indo-European nominal inflection. *Language*, 34 (2):179–202, 1958.

Elisabeth Leiss. Gender in Old High German. In *Gender in Grammar and Cognition*, pages 237–258. De Gruyter Mouton, 2000.

Yeptain Leung, Jennifer Oates, Siew-Pang Chan, and Viktória Papp. Associations between speaking fundamental frequency, vowel formant frequencies, and listener perceptions of speaker gender and vocal femininity–masculinity. *Journal of Speech, Language, and Hearing Research*, 64(7):2600–2622, 2021.

Hector J. Levesque. On Our Best Behaviour. *Artificial Intelligence*, 212(1):27–35, July 2014. ISSN 0004-3702. doi: 10.1016/j.artint.2014.03.007. URL https://doi.org/10.1016/j.artint.2014.03.007.

Roger J. R. Levesque. *Sex Roles and Gender Roles*, pages 2622–2623. Springer New York, New York, NY, 2011. ISBN 978-1-4419-1695-2. doi: 10.1007/978-1-4419-1695-2_602. URL https://doi.org/10.1007/978-1-4419-1695-2_602.

Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore. Automatic identification of gender from speech. In *In Proocedings of Speech Prosody 2016*, pages 84–88, Boston, Massachusetts, May-June 2016. doi: 10.21437/SpeechProsody.2016-18. URL http://dx.doi.org/10.21437/SpeechProsody.2016-18.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. *arXiv preprint arXiv:2109.03858*, pages 2470–2480, November 2021. doi: 10.18653/v1/2021.findings-emnlp.211. URL https://aclanthology.org/2021.findings-emnlp.211.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of* *the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Molly Lewis and Gary Lupyan. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028, 2020. URL https://doi.org/10.1038/s41562-020-0918-6.

Fangfang Li. The development of gender-specific patterns in the production of voiceless sibilant fricatives in mandarin chinese. *Linguistics*, 55(5):1021–1044, 2017.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.194. URL https://aclanthology.org/2022.findings-acl.194.

Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. Unmet Needs and Opportunities for Mobile Translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376261. URL https://doi.org/10.1145/3313831.3376261.

Anna Lindqvist, Emma A. Renström, and Marie Gustafsson Sendén. Reducing a Male Bias in Language? Establishing the Efficiency of three Different Gender-fair Language Strategies. *Sex Roles*, 81(1-2):109–117, 2019. URL https://doi.org/10.1007/s11199-018-0974-9.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. Character-based Neural Machine Translation. *arXiv preprint arXiv:1511.04586*, 2015.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Zachary C. Lipton. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue*, 16(3):31–57, jun 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.

Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16), pages 923–929, Portorož, SI, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1147.

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 25–32, 2005.

Katherine A. Liu and Natalie A. Dipietro Mager. Women's Involvement in Clinical Trials: Historical Perspective and Future Implications. *Pharmacy Practice*, 14(1):708, 2016. doi: 10.18549/pharmpract.2016.01.708.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. End-to-End Speech Translation with Knowledge Distillation. In *Proceedings of Interspeech 2019*, pages 1128–1132, Graz, Austria, September 2019. doi: 10.21437/Interspeech.2019-2582.

Chi-kiu Lo. YiSi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, 2019.

Arle Richard Lommel, Alojscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: A flexible system for assessing translation quality. *Proceedings of ASLIB: Translating and the Computer*, 35:311–318, 2013.

Arle Richard Lommel, Maja Popovic, and Aljoscha Burchardt. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*, pages 31–37. Language Resources and Evaluation Conference Reykjavik, 2014.

Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49, 2008.

Ártemis López, Susana Rodríguez Barcia, and María del Carmen Cabeza Pereiro. Visibilizar o interpretar: respuesta al Informe de la Real Academia Española sobre el lenguaje inclusivo y cuestiones conexas. http://www.ngenespanol.com/el-mundo/la-rae-rechaza-nuevamente-el-lenguaje-inclusivo/, 2020. Accessed: 2021-02-25.

Brandon C. Loudermilk. Implicit Attitudes and the Perception of Sociolinguistic Variation. *Responses to language varieties: Variability, processes and outcomes*, pages 137–156, 2015.

Leo Loveday. Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, 24(1):71–89, 1981.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender Bias in Neural Natural Language Processing. In *Logic, Language, and Security*, volume 12300 of *Lecture Notes in Computer Science*, pages 189–202. Springer, 2019. URL https://doi.org/10.1007/978-3-030-62077-6_14.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015a. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://aclanthology.org/D15-1166.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July 2015b. Association for Computational Linguistics. doi: 10.3115/v1/P15-1002. URL https://aclanthology.org/P15-1002.

John Lyons. *Semantics*, volume 2. Cambridge University Press, Cambrdige, UK, 1977.

Ártemis López. Cuando el lenguaje excluye: Consideraciones sobre el lenguaje no binario indirecto. *Cuarenta Naipes*, 3:295–312, June 23 2020. URL https://fh.mdp.edu.ar/revistas/index.php/cuarentanaipes/article/view/4891.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.602. URL https://aclanthology.org/2020.emnlp-main.602.

Kate MacKrill, Connor Silvester, James W. Pennebaker, and Keith J. Petrie. What makes an idea worth spreading? language markers of popularity in ted talks by academics and other speakers. *Journal of the Association for Information Science and Technology*, 2021.

Nishtha Madaan, Sameep Mehta, Taneea S. Agrawaal, Vrinda Malhotra, Aditi Aggarwal, and Mayank Saxena. Analyzing gender stereotyping in bollywood movies, 2017.

Nishtha Madaan, Sameep Mehta, Shravika Mittal, and Ashima Suvarna. Judging a book by its description : Analyzing gender stereotypes in the man bookers prize winning fiction, 2018.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. Socially Aware Bias Measurements for Hindi Language Representations. *arXiv preprint arXiv:2110.07871*, 2021.

Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

Michael Maratsos. How to get from words to sentences. In *Psycholinguistic Research*, pages 285–353. Psychology Press, 2013.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. 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 7297–7306, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.566. URL https://aclanthology.org/2021.acl-long.566.

Marianna Martindale and Marine Carpuat. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, USA, March 2018. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/W18-1803.

Marianna Martindale, Kevin Duh, and Marine Carpuat. Machine Translation Believability. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 88–95, 2021.

Giuseppe Martucci, Mauro Cettolo, Matteo Negri, and Marco Turchi. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Interspeech*, pages 2282–2286, 2021.

Ranko Matasović. *Gender in Indo-European*. Universitaetsverlag Winter, Heidelberg, 2004.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. Gender Bias in Natural Language Processing Across Human Languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, 2021.

Peter Hugoe Matthews. *The concise Oxford dictionary of linguistics*. Oxford University Press, 1997.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6(3):175–183, 2004.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. Customizing Neural Machine Translation for Subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5209. URL https://www.aclweb.org/anthology/W19-5209.

Chandler May. Deconstructing Gender Prediction in NLP. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Vancouver, CA, 2019. URL https://slideslive.com/38923450/deconstructing-gender-prediction-in-nlp?locale=en.

Michael Mccloskey and Neil J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *The Psychology of Learning and Motivation*, 24:104–169, 1989.

Sally McConnell-Ginet. Gender and its Relation to Sex: The Myth of 'Natural' Gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton, Berlin, DE, 2013. URL https://doi.org/10.1515/9783110307337.3.

Thomas R. McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, IT, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://www.aclweb.org/anthology/P19-1334.

Thomas R. McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *arXiv e-prints*, pages arXiv–2111, 2021.

Elin McCready. *The semantics and pragmatics of honorification: Register and social meaning*, volume 11. Oxford University Press, USA, 2019.

Hayley McGlashan and Katie Fitzpatrick. 'I use any pronouns, and I'm questioning everything else': Transgender youth and the issue of gender pronouns. *Sex Education*, 18(3):239–252, 2018. doi: 10.1080/14681811.2017.1419949. URL https://doi.org/10.1080/14681811.2017.1419949.

Kevin A. McLemore. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74, 2015.

John McWhorter. The Power of Babel. *A natural history of language*, 2001.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*, page 169, 2021.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. *arXiv preprint arXiv:2112.10508*, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Viviana Patti, Tommaso Caselli, and Malvina Nissim. Frame Semantics for Social NLP in Italian: Analyzing Responsibility Framing in Femicide News Reports. In *Italian Conference on Computational Linguistics 2021: CLiC-it 2021*. CEUR Workshop Proceedings (CEUR-WS. org), 2021.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. Motivating Personality-Aware Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, PT, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1130. URL https://www.aclweb.org/anthology/D15-1130.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. Diversity and Inclusion Metrics in Subset Selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 117–123, New York, USA, 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3375627.3375832.

Melanie Mitchell. Why AI is Harder than We Think. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '21, page 3, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383509. doi: 10.1145/3449639.3465421. URL https://doi.org/10.1145/3449639.3465421.

Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2022.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2): 2053951716679679, 2016. doi: 10.1177/2053951716679679. URL https://doi.org/10.1177/2053951716679679.

Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn. Understanding how deep belief networks perform acoustic modelling. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4273–4276. IEEE, 2012.

Saif Mohammad. Ethics Sheets for AI Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379, 2022.

Britta Mondorf. Gender Differences in English Syntax. *Journal of English Linguistics*, 30: 158–180, 06 2002. doi: 10.1177/007242030002005.

John Money. Hermaphroditism, gender and precocity in hyperadrenocorticism: Psychologic findings. *Bulletin of the Johns Hopkins Hospital*, 96(6):253–264, 1955.

Johanna Monti. Questioni di Genere in Traduzione Automatica. *Al femminile. Scritti linguistici in onore di Cristina Vallini*, 139:411–431, 2017. URL http://hdl.handle.net/11574/178435.

Johanna Monti. Gender Issues in Machine Translation: An Unsolved Problem? In Luise von Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge, 2020. URL https://www.taylorfrancis.com/chapters/gender-issues-machine-translation-johanna-monti/e/10.4324/9781315158938-39.

Amit Moryossef, Roee Aharoni, and Yoav Goldberg. Filling Gender & Number Gaps in Neural Machine Translation with Black-Box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, IT, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3807. URL https://www.aclweb.org/anthology/W19-3807.

Heiko Motschenbacher. Grammatical gender as a challenge for language policy: The (im)possibility of non-heteronormative language use in German versus English. *Language policy*, 13(3):243–261, 2014. URL https://doi.org/10.1007/s10993-013-9300-0.

Janice Moulton, George M. Robinson, and Cherin Elias. Sex bias in language use:" Neutral" pronouns that aren't. *American psychologist*, 33(11):1032, 1978.

Anthony Mulac, James J. Bradac, and Pamela Gibbons. Empirical Support for the Gender-as-Culture Hypothesis. *Human Communication Research*, 27:121– 152, 01 2001. doi: 10.1111/j.1468-2958.2001.tb00778.x.

National Geographic El Mundo. La RAE rechaza nuevamente el lenguaje inclusivo. https://www.ngenespanol.com/el-mundo/la-rae-rechaza-nuevamente-el-lenguaje-inclusivo/, 2018. Accessed: 2021-02-25.

Benjamin Munson and Molly Babel. Loose lips and silver tongues, or, projecting sexual orientation through speech. *Language and Linguistics Compass*, 1(5):416–449, 2007.

David A. B. Murray. Who is Takatāpui? Māori Language, Sexuality and Identity in Aotearoa/New Zealand. *Anthropologica*, pages 233–244, 2003. URL https://doi.org/10.2307/25606143.

Iftekhar Naim, Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, 2015. doi: 10.1109/FG.2015.7163127.

Terttu Nevalainen and Helena Raumolin-Brunberg. Its Strength and the Beauty of it: The Standardization of the Third Person Neuter Possessive in Early Modern English. In Dieter Stein and Ingrid Tieken-Boon van Ostade, editors, *Towards a Standard English*, pages 171–216. De Gruyter, Berlin, DE, 1993. URL https://doi.org/10.1515/9783110864281.171.

Matthew L. Newman, Carla J Groom, Lori D. Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3): 211–236, 2008.

Dong Nguyen, Seza Doğruöz A., Carolyn P. Rosé, and Franciska de Jong. Computational Sociolinguistics: A Survey. *Computational linguistics*, 42(3):537–593, 2016. URL https://doi.org/10.1162/COLI_a_00258.

Dong Nguyen, Laura Rosseel, and Jack Grieve. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online, June 2021. Association for Computational Linguistics.

doi: 10.18653/v1/2021.naacl-main.50. URL https://aclanthology.org/2021.naacl-main.50.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proceedings of the 2020 International Conference on Acoustics, Speech, and Signal Processing – IEEE-ICASSP-2020*, Barcelona, Spain, May 2020.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT)*, Bruges, Belgium, October 2018.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, Hong Kong, November 2019.

Uwe Kjær Nissen. Aspects of Translating Gender. *Linguistik Online*, 11(2), 2002. doi: 10.13092/lo.11.914.

Malvina Nissim and Rob van der Goot. Fair is Better than Sensational: Man is to Doctor as Woman is to Doctor. *Computational Linguistics*, 46(2):487–497, 2020. doi: 10.1162/coli_a_00379. URL https://doi.org/10.1162/coli_a_00379.

Safiya Umoja Noble. Algorithms of Oppression. In *Algorithms of Oppression*. New York University Press, 2018.

Jennifer Oates and Georgia Dacakis. Transgender voice and communication: Research evidence underpinning voice intervention for male-to-female transsexual women. *Perspectives on Voice and Voice Disorders*, 25(2):48–58, 2015.

Parmy Olson. The Algorithm That Helped Google Translate Become Sexist. https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/?sh=d675b9c7daa2, 2018. URL https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-/translate-become-sexist/?sh=d675b9c7daa2. Accessed: 2021-02-25.

Cathy O'Neil. *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. Broadway books, 2016.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, Brisbane, Australia, April 2015.

Dimitrios Papadimoulis. *GENDER-NEUTRAL LANGUAGE in the European Parliament*. European Parliament 2018, 2018. doi: https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf.

Benjamin Papadopoulos. *Morphological Gender Innovations in Spanish of Gender queer Speakers*. Department of Spanish and Portuguese, University of California, UC Berkeley, 2019. URL https://escholarship.org/uc/item/6j73t666.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617, Graz, Austria, September 2019. doi: 10.21437/Interspeech.2019-2680. URL http://dx.doi.org/10.21437/Interspeech.2019-2680.

Eluned S. Parris and Michael J. Carey. Language independent gender identification. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 685–688. IEEE, 1996.

Amandalynne Paullada. *Considerations for the social impact of natural language processing*. PhD thesis, University of Washington, 2021.

Amandalynne Paullada, Inioluwa D. Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA)*, Vitual, December 2020. URL https://ml-retrospectives.github.io/neurips2020/camera_ready/19.pdf.

James Pennebaker and Lori Stone. Words of Wisdom: Language Use Over the Life Span. *Journal of personality and social psychology*, 85:291–301, 09 2003. doi: 10.1037/0022-3514.85.2.291.

Sarah Perez. Nearly 70% of US smart speaker owners use Amazon Echo devices. https://techcrunch.com/2020/02/10/nearly-70-of-u-s-smart-speaker-owners-use-amazon-echo-devices/, 2020. Accessed: 2022-08-16.

Juan Antonio Pérez-Ortiz, Mikel L. Forcada, and Felipe Sánchez-Martínez. How neural machine translation works. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:141, 2022.

Nicolai Pharao, Marie Maegaard, Janus Spindler Møller, and Tore Kristiansen. Indexical meanings of [s] among copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, 43(1):1–31, 2014. doi: 10.1017/S0047404513000857.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. Harnessing Indirect Training Data for End-to-End Automatic Speech Translation: Tricks of the Trade. In *Proceedings of the 16th International Conference on Spoken Language Translation*, 2019.

Jeff Pitman. Google Translate: One billion installs, one billion stories, 2018. URL https://blog.google/products/translate/one-billion-installs/. Accessed: 2022-07-28.

Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 13–20. Bochumer Linguistische Arbeitsberichte, 2016.

Robert J. Podesva, Janneke Van Hofwegen, Erez Levon, and Ronald Beline Mendes. S/exuality in small-town California: Gender normativity and the acoustic realization of/s. *Language, sexuality, and power: Studies in intersectional linguistics*, pages 16–88, 2016.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601, 2019.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Il Post. C'è confusione su come chiamare Giorgia Meloni. https://www.ilpost.it/2022/10/28/giorgia-meloni-signor-presidente-del-consiglio/, 2022. Accessed: 2022-2-11.

Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018b. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013.

Tomasz Potapczyk and Pawel Przybysz. SRPOL's System for the IWSLT 2020 End-to-End Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.iwslt-1.9.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 1–4, Big Island, Hawaii, 2011. IEEE Signal Processing Society.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. In *Proceedings of INTERSPEECH 2016*, pages 2751–2755, San Francisco, California, 2016.

Sam Prance. What is the Penny Challenge on TikTok? The dangerous trend explained, 2021. URL https://www.popbuzz.com/internet/viral/penny-challenge-tiktok-explained/. Accessed: 2022-08-13.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. Assessing Gender Bias in Machine Translation: a Case Study with Google Translate. *Neural Computing and Applications*, 32 (10):1–19, 2018. URL https://doi.org/10.1007/s00521-019-04144-6.

Dennis R. Preston. Are you really smart (or stupid, or cute, or ugly, or cool)? Or do you just talk that way. *Language attitudes, standardization and language change. Oslo: Novus forlag*, pages 105–129, 2009.

Ella Rabinovich, Raj N. Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 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, Valencia, ES, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1101.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, and Matthew O. Jackson. Machine Behaviour. *Nature*, 568(7753):477–486, 2019. URL https://doi.org/10.1038/s41586-019-1138-y.

Rajat Raina, Anand Madhavan, and Andrew Y. Ng. Large-scale Deep Unsupervised Learning using Graphics Processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880, 2009.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. Evaluating Gender Bias in Hindi-English Machine Translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.3. URL https://aclanthology.org/2021.gebnlp-1.3.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*, 2021.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of 4th International Conference on Learning Representations ICLR*, San Juan, Puerto Rico, May 2-4 2016. URL https://arxiv.org/abs/1511.06732.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1172–1183, 2021.

Eric S. Raymond. *The New Hacker's Dictionary*. MIT Press, 1996.

Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature human behaviour*, 3(11):1171–1179, 2019. URL https://doi.org/10.1038/s41562-019-0686-3.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, 2020.

NEC Press Release. World's First Speech-to-Speech Translation Capable Robot. http://www.nec.co.jp/press/en/0401/0601.html, 2004. Accessed: 2022-02-09.

Adithya Renduchintala and Adina Williams. Investigating Failures of Automatic Translation in the Case of Unambiguous Gender. *arXiv preprint arXiv:2104.07838*, 2021.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. Gender bias amplification during Speed-Quality optimization in Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.15. URL https://aclanthology.org/2021.acl-short.15.

La Repubblica. Sanremo, Beatrice Venezi:“Direttore, non direttrice". E i social si spaccano sulla scelta. https://www.repubblica.it/dossier/spettacoli/sanremo-2021/2021/03/06/news/sanremo_beatrice_venezi_direttore_non_direttrice_e_i_social_si_spaccano_sulla_scelta-290592565/#:~:text=Sanremo%202021%2C%20Beatrice%20Venezi%20si,video%20in%20fase%20di%20caricamento., 2021. Accessed: 2022-2-11.

Argentina A. Rescigno, Johanna Monti, Andy Way, and Eva Vanmassenhove. A Case Study of Natural Gender Phenomena in Translation: A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In *Proceedings of the Workshop on the Impact of Machine Translation (iMpacT 2020)*, pages 62–90, Online, October 2020. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/2020.amta-impact.4.

Natália Resende, Benjamin Cowan, and Andy Way. MT syntactic priming effects on L2 English speakers. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 245–253, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.26.

David Reynolds, Alan Garnham, and Jane Oakhill. Evidence of immediate activation of gender information from a social role name. *Quarterly Journal of Experimental Psychology*, 59(5): 886–903, 2006. doi: 10.1080/02724980543000088. URL https://doi.org/10.1080/02724980543000088. PMID: 16608753.

Alexander S. Rich and Todd M. Gureckis. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, 1(4):174–180, 2019. URL https://doi.org/10.1038/s42256-019-0038-z.

Christina Richards, Walter P. Bouman, Leighton Seal, Meg John Barker, Timo O. Nieder, and Guy T'Sjoen. Non-binary or Genderqueer Genders. *International Review of Psychiatry*, 28 (1):95–102, 2016. URL https://doi.org/10.3109/09540261.2015.1106446.

John R. Rickford. *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press, 2016.

John R. Rickford and Sharese King. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, pages 948–988, 2016.

Barbara J. Risman. Gender as a Social Structure. In Barbara J. Risman, Carissa Froyum, and William J. Scarborough, editors, *Handbook of the Sociology of Gender*, pages 19–43. Springer, 2018. doi: 10.1007/978-3-319-76333-0.

Elizabeth Ritter. Where's gender? *Linguistic Inquiry*, 24(4):795–803, 1993. ISSN 00243892, 15309150. URL http://www.jstor.org/stable/4178843.

Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C. Lipton. Decoding and Diversity in Machine Translation. In *Proceedings of the Resistance AI Workshop at 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, February 2020. URL https://drive.google.com/file/d/1crAS9oknszKV6Gssr9W8zyAupZ-i8sg9/view.

James A. Rodger and Parag C. Pendharkar. A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*, 60(5-6):529–544, 2004.

Anna Rogers. Changing the World by Changing the Data. 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 2182–2194, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.170. URL https://aclanthology.org/2021.acl-long.170.

Suzanne Romaine. Gender, grammar, and the space in between. *Communicating Gender in Context*, 42:51, 1997.

Suzanne Romaine. *Communicating Gender*. Lawrence Erlbaum, Mahwah, USA, 1999.

Suzanne Romaine. A Corpus-Based View of Gender in British and American English. *Gender across languages*, 1:153–175, 2001. URL https://doi.org/10.1075/impact.9.12rom.

Jonathan Rosa and Nelson Flores. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, 46(5):621–647, 2017.

Khushi Roy, Subhra Debdas, Sayantan Kundu, Shalini Chouhan, Shivangi Mohanty, and Biswarup Biswas. Application of Natural Language Processing in Healthcare. *Computational Intelligence and Healthcare Informatics*, pages 393–407, 2021.

Donald L. Rubin and Kathryn L.. Greene. Effects of biological and psychological gender, age cohort, and interviewer gender on attitudes toward gender-inclusive/exclusive language. *Sex Roles*, 24(7):391–412, 1991.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://www.aclweb.org/anthology/N18-2002.

Ram Samar. Machines Are Indifferent, We Are Not: Yann LeCun's Tweet Sparks ML Bias Debate. https://analyticsindiamag.com/yann-lecun-machine-learning-bias-debate/, 2020. Accessed: 2021-02-25.

Steven Samuel, Geoff Cole, and Madeline J. Eacott. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic bulletin & review*, 26(6):1767–1786, 2019.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proceedings of Visually Grounded Interaction and Language (ViGIL)*, Montréal, Canada, December 2018. Neural Information Processing Society (NeurIPS). URL https://hal.archives-ouvertes.fr/hal-02431947.

Anthoy J. Sanford. *Cognition and cognitive psychology*. Weidenfeld and Nicolson, London, UK, 1985.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486.

Naomi Saphra and Adam Lopez. Language Models Learn POS First. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 328–330, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5438. URL https://aclanthology.org/W18-5438.

Naomi Saphra and Adam Lopez. Understanding Learning Dynamics Of Language Models with SVCCA. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1329. URL https://aclanthology.org/N19-1329.

Danielle Saunders and Bill Byrne. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.690. URL https://www.aclweb.org/anthology/2020.acl-main.690.

Danielle Saunders, Rosie Sallis, and Bill Byrne. Neural Machine Translation Doesn’t Translate Gender Coreference Right Unless You Make It. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.4.

Danielle Saunders, Rosie Sallis, and Bill Byrne. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland, May 2022. Association for Computational

Linguistics. doi: 10.18653/v1/2022.findings-acl.301. URL https://aclanthology.org/2022.findings-acl.301.

Beatrice Savoldi and Luisa Bentivogli. Gender bias and machine translation: On first looking into parallel corpora. In *Book of Abstracts*, page 143, 2021. URL https://events.unibo.it/uccts2021/book-of-abstracts.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021. doi: 10.1162/tacl_a_00401. URL https://aclanthology.org/2021.tacl-1.51.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. On the Dynamics of Gender Learning in Speech Translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111, Seattle, Washington, July 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.gebnlp-1.12.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. *ArXiv e-prints arXiv:2203.09866. Accepted at ACL 2022*, 2022b. doi: https://doi.org/10.48550/arXiv.2203.09866.

Paul Schachter and Timothy Shopen. Parts-of-speech systems. *Language Typology and Syntactic Description. Vol. 1: Clause Structure*, pages 1–60, 2007. URL http://services.cambridge.org/us/academic/subjects/languages-linguistics/grammar-and-syntax/language-typology-and-syntactic-description-volume-1-2nd-edition?format=HB.

Elmar Schafroth. *Gender in French: Structural properties, incongruences and asymmetries*, pages 50–72. John Benjamins Publishing, Amsterdam, The Netherlands, 2003.

Londa Schiebinger. Scientific Research Must Take Gender into Account. *Nature*, 507(9), 2014. doi: 10.1038/507009a.

Tyler Schnoebelen. The carrots and sticks of ethical NLP. https://medium.com/@TSchnoebelen/ethics-and-nlp-some-further-thoughts-53bd7cc3ff69, 2017. Accessed: 2022-08-06.

Herbert Schriefers and Jörg D. Jescheniak. Representation and processing of grammatical gender in language production: A review. *Journal of psycholinguistic research*, 28(6):575–600, 1999.

Muriel R. Schulz. The Semantic Derogation of Woman. In Barrie Thorne and Nancy Henley, editors, *Sex and language. Difference and dominance*, pages 64–75. Newbury House, Rowley, USA, 1975.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, CN, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1341. URL https://www.aclweb.org/anthology/D19-1341.

Stefan R. Schweinberger, Hideki Kawahara, Adrian P. Simpson, Verena G. Skuk, and Romi Zäske. Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1):15–25, 2014.

William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Pubulic Opinion Quarterly*, 19:321–325, 1955.

Sabine Sczesny, Christa Nater, and Alice H. Eagly. Agency and communion: Their implications for gender stereotypes and gender identities. In *Agency and Communion in Social Psychology*, pages 103–116. Taylor and Francis, August 2018. ISBN 9781138570269. doi: 10.4324/9780203703663.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, USA, 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

Rico Sennrich. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2060.

Rico Sennrich and Biao Zhang. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Maria D. Sera, Chryle Elieff, James Forbes, Melissa Clark Burch, Wanda Rodríguez, and Diane Poulin Dubois. When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General*, 131 (3):377, 2002.

Deven S. Shah, Hansen A. Schwartz, and Dirk Hovy. 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*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://www.aclweb.org/anthology/2020.acl-main.468.

Saqib Shah and Julian Chokkattu. Microsoft's disastrous Tay experiment shows the hidden dangers of AI, 2016. URL https://www.digitaltrends.com/social-media/microsoft-tay-chatbot/. Accessed: 2022-08-13.

Sam Shead. Amazon's Alexa assistant told a child to do a potentially lethal challenge, 2022. URL https://www.cnbc.com/2021/12/29/amazons-alexa-told-a-child-to-do-a-potentially-lethal-challenge.html. Accessed: 2022-08-13.

Emily Sheng and David Uthus. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.9.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.189. URL https://aclanthology.org/2021.naacl-main.189.

Jeanette Silveira. Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3):165–178, 1980. URL https://www.sciencedirect.com/science/article/pii/S0148068580921132.

Adrian P. Simpson. Phonetic differences between male and female speech. *Language and linguistics compass*, 3(2):621–640, 2009.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-2006. URL https://www.aclweb.org/anthology/E14-2006.

Janet Smith. Gendered Structures in Japanese. *Gender across languages: The linguistic representation of women and men*, 3:201–227, 2003. URL https://doi.org/10.1075/impact.11.12shi.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231, Cambridge, Massachusetts, August 2006a. The Association for Machine Translation in the Americas. URL http://mt-archive.info/AMTA-2006-Snover.pdf.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,*, pages 223–231, Cambridge, August 2006b. Association for Machine Translation in the Americas.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021.

Robyn Speer. ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors. https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-wordvectors/, 2017. Accessed: 2021-02-25.

Linda E. Spencer. Speech Characteristics of Male-to-Female Transsexuals: A Perceptual and Acoustic Study. *Folia Phoniatrica et Logopaedica*, 40(1):31–42, 1988.

Matthias Sperber and Matthias Paulik. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, 2020.

Matthias Sperber, Jan Niehues, and Alex Waibel. Toward Robust Neural Machine Translation for Noisy Input Sequences. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2017.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. Self-Attentional Acoustic Models. In *Proceedings of Interspeech 2018*, pages 3723–3727, 2018. doi: 10.21437/Interspeech.2018-1910. URL http://dx.doi.org/10.21437/Interspeech.2018-1910.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. Consistent Transcription and Translation of Speech. *Transactions of the Association for Computational Linguistics (TACL)*, 8:695–709, 2020.

Richard Sproat. *Language, Technology, & Society*. Oxford University Press, 2010.

Shikaripur N. Sridhar. *Kannada: Descriptive Grammar*. Croom Helm Descriptive Grammars. Routledge, London, 1990.

Patrick Stadler, Vivien Macketanz, and Eleftherios Avramidis. Observing the Learning Curve of NMT Systems With Regard to Linguistic Phenomena. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 186–196, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-srw.20. URL https://aclanthology.org/2021.acl-srw.20.

Artūrs Stafanovičs, Mārcis Pinnis, and Toms Bergmanis. Mitigating Gender Bias in Machine Translation with Target Gender Annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.73.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. Representation of the Sexes in Language. *Social communication*, pages 163–187, 2007. URL https://psycnet.apa.org/record/2007-01308-006.

Karolina Stanczak and Isabelle Augenstein. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint arXiv:2112.14168*, 2021.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://www.aclweb.org/anthology/P19-1164.

Daniel Stein. Machine translation: Past, present and future. *Language technologies for a multilingual Europe*, 4(5), 2018.

Katy Steinmetz. Everything You Need to Know About the Word 'Suffragette', 2015. URL https://time.com/4079176/suffragette-word-history-film/. Accessed: 2021-02-25.

Fred WM Stentiford and Martin G. Steer. Machine translation of speech. *British Telecom technology journal*, 6(2):116–123, 1988. URL https://www.researchgate.net/profile/Fred-Stentiford/publication/292097271_MACHINE_TRANSLATION_OF_SPEECH/links/594ba6c8aca2720930f45309/MACHINE-TRANSLATION-OF-SPEECH.pdf.

Elizabeth A. Strand. *Gender stereotype effects in speech processing*. PhD thesis, The Ohio State University, 2000.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.14. URL https://aclanthology.org/2020.nlpcovid19-2.14.

Lucy A. Suchman. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge university press, 2007.

Jiao Sun and Nanyun Peng. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, 2021.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 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, IT, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://www.aclweb.org/anthology/P19-1159.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 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, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://www.aclweb.org/anthology/P19-1159.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv preprint arXiv:2102.06788*, 2021. URL https://arxiv.org/abs/2102.06788.

Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019. URL https://arxiv.org/abs/1901.10002.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States, June 2016.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.gebnlp-1.5.

Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. A Japanese-to-English speech translation system: ATR-MATRIX. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, November–December 1998.

Tina Tallon. A Century of "Shrill": How Bias in Technology Has Hurt Women's Voices. *The New Yorker*, March 2019. URL https://www.newyorker.com/culture/cultural-comment/a-century-of-shrill-how-bias-in-technology-has-hurt-womens-voices.

Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://www.aclweb.org/anthology/W17-1606.

Mildred C. Templin. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, 1957.

Jörg Tiedemann. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384, 2016. ISSN 2255-8942. Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT).

Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk. *Meaning and Lexicography*, volume 28. John Benjamins Publishing, 1990.

Antonio Toral and Víctor M. Sánchez-Cartagena. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1100.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, 2018.

Jonas-Dario Troles and Ute Schmid. Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives. *arXiv preprint arXiv:2107.11584*, pages 531–541, November 2021. URL https://aclanthology.org/2021.wmt-1.61.

Peter Trudgill. Language contact and the function of linguistic gender. *Poznan studies in contemporary linguistics*, 35:133–152, 1999.

Peter Trudgill. *Sociolinguistics: An Introduction to Language and Society*. Penguin Books, London, UK, 2000.

Peter Trudgill. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press, 2011.

Junichi Tsujii. Natural language processing and computational linguistics. *Computational Linguistics*, 47(4):707–727, 2021.

G. Richard Tucker, Wallace E. Lambert, and André Albert Rigault. The French speaker's skill with grammatical gender. In *The French Speaker's Skill with Grammatical Gender*. De Gruyter Mouton, 1977.

Anne M. Turner, Megumu K. Brownstein, Kate Cole, Hilary Karasz, and Katrin Kirchhoff. Modeling Workflow to Design Machine Translation Applications for Public Health practice. *Journal of Biomedical Informatics*, 53:136–146, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2014.10.005. URL https://www.sciencedirect.com/science/article/pii/S1532046414002251.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973. URL https://www.sciencedirect.com/science/article/pii/0010028573900339.

Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974. URL https://doi.org/10.1126/science.185.4157.1124.

Mark Twain. The Awful German Language. A Trump Abroad, 1880. URL https://en.wikisource.org/wiki/A_Tramp_Abroad/Appendix_D. Accessed: 2022-02-10.

Fiona J. Tweedie and Harald R. Baayen. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

Nneka Umera-Okeke. Linguistic sexism: an overview of the English language in everyday discourse. *AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies*, 1(1):1–17, 2012.

Rhoda K. Unger and Mary Crawford. Sex and gender—The troubled relationship between terms and concepts. *Psychological science*, 4(2):122–124, 1993.

Jannis Vamvas and Rico Sennrich. Contrastive Conditioning for Assessing Disambiguation in MT: A Case Study of Distilled Bias. In *2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL Anthology, November 2021. URL https://doi.org/10.5167/uzh-206179.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. The Birth of Bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.8. URL https://aclanthology.org/2022.gebnlp-1.8.

Emiel van Miltenburg. Stereotyping and bias in the Flickr30k dataset. In *Proceedings of multimodal corpora: computer vision and language processing (mmc 2016)*, 2016. 11th workshop on multimodal corpora: computer vision and language processing ; Conference date: 01-01-2016 Through 01-01-2016.

Martijn Van Otterlo. A Machine Learning View on Profiling. *Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology. Abingdon: Routledge*, pages 41–64, 2013.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL https://www.aclweb.org/anthology/D18-1334.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://www.aclweb.org/anthology/W19-6622.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.704.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 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, 2021b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Proceedings of NIPS 2017*, pages 5998–6008, Long Beach, California, December 2017. NIPS.

Rivarol Vergin, Azarshid Farhat, and Douglas O'Shaughnessy. Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1081–1084. IEEE, 1996.

Lucas Nunes Vieira, Carol O'Sullivan, Xiaochun Zhang, and Minako O'Hagan. Machine translation in society: Insights from UK users. *Language Resources and Evaluation*, pages 1–22, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Elena Voita, Rico Sennrich, and Ivan Titov. Context-Aware Monolingual Repair for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1081. URL https://aclanthology.org/D19-1081.

Elena Voita, Rico Sennrich, and Ivan Titov. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://aclanthology.org/P19-1116.

Elena Voita, Rico Sennrich, and Ivan Titov. Analyzing the Source and Target Contributions to Predictions in Neural Machine Translation. 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 1126–1140, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.91. URL https://aclanthology.org/2021.acl-long.91.

Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. From Gender Biases to Gender-Inclusive Design: An Empirical Investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300283. URL https://doi.org/10.1145/3290605.3300283.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's Wikipedia?

Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. Janus: A speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP 1991*, pages 793–796, Toronto, Canada, May 14-17 1991.

Judy Wajcman. From women and technology to gendered technoscience. *Information, Community and Society*, 10(3):287–298, 2007.

Mario Wandruszka. *Sprachen: Vergleichbar und Vnvergleichlich.* R. Piper & Co. Verlag, Munich, DE, 1969. URL https://doi.org/10.1017/S0022226700002978.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, 2016.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. OpenReview Preprint, 2020. URL https://openreview.net/pdf?id=fkAxTMzy3fs.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, IT, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3801. URL https://www.aclweb.org/anthology/W19-3801.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Mark Weiser. The computer for the 21st century, SIGMOBILE Mob. *Computing Communication Review*, 3(3):3–11, 1999.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proceedings of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden, August 2017.

Candace West and Don H. Zimmerman. Doing gender. *Gender & society*, 1(2):125–151, 1987.

Benj Ide Wheeler. The origin of grammatical gender. *The Journal of Germanic Philology*, 2 (4):528–545, 1899.

Frederick Williams, Jack L. Whitehead, and Leslie M. Miller. Ethnic Stereotyping and Judgments of Children's Speech. *Communications monographs*, 38(3):166–170, 1971.

Langdon Winner. Do Artifacts Have Politics? *Daedalus*, pages 121–136, 1980.

Ilka B. Wolter and Bettina Hannover. Gender role self-concept at school start and its impact on academic self-concept and performance in mathematics and reading. *European Journal of Developmental Psychology*, 13(6):681–703, 2016. URL https://doi.org/10.1080/17405629.2016.1175343.

Ikuko Patricia Yuasa. *Culture and gender of voice pitch: A sociophonetic comparison of the Japanese and Americans*. Equinox Publishing, 2008.

Mohd Ali Yusnita, AM Hafiz, M Nor Fadzilah, Aida Zulia Zulhanip, and Mohaiyedin Idris. Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal. In *2017 7th IEEE international conference on control system, computing and engineering (ICCSCE)*, pages 372–377. IEEE, 2017.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems. In *Proc. Interspeech 2022*, pages 1278–1282, 2022. doi: 10.21437/Interspeech.2022-353.

Rabih Zbib, Gretchen Markiewicz, Spyros Matsoukas, Richard Schwartz, and John Makhoul. Systematic comparison of professional and crowdsourced reference translations for machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 612–616, 2013.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, DK, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, USA, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, BE, October-November 2018c. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://www.aclweb.org/anthology/D18-1521.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.260. URL https://www.aclweb.org/anthology/2020.acl-main.260.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, CN, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1531. URL https://www.aclweb.org/anthology/D19-1531.

Lal Zimman. Representing trans: Linguistic, legal and everyday perspectives. In Evan Hazenberg and Miriam Meyerhoff, editors, *Trans People's Linguistic Self-determination and the Dialogic Nature of Identity*, page 226–248. Victoria University Press, 2017.

Lal Zimman. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, 12(8):e12284, 2018.

Lal Zimman. Transgender language, transgender moment: Toward a trans linguistics. In Kira Hall and Rusty Barrett, editors, *The Oxford Handbook of Language and Sexuality*. Oxford University Press, 2020. doi: 10.1093/oxfordhb/9780190212926.013.45.

Lal Zimman. Gender diversity and the voice. In *The Routledge handbook of language, gender, and sexuality*, pages 69–90. Routledge, 2021.

Lal Zimman, Evan Hazenberg, and Miriam Meyerhoff. Trans people's linguistic self-determination and the dialogic nature of identity. *Representing trans: Linguistic, legal and everyday perspectives*, pages 226–248, 2017. URL http://www.lalzimman.com/papers/Zimman2017RepresentingTrans.pdf.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, IT, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://www.aclweb.org/anthology/P19-1161.

A

List of Publications and Activities

Here, I present all the activities and collaborations that have had an impact on the final shape of this thesis and its scientific growth, starting from my publications.

Publications. List of published manuscripts included in the thesis. The symbol * denotes equal contributions, with first authors listed in alphabetical order.[1]

- 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.

- 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. **Outstanding**

[1] For all works, I am first author, meaning that I was responsible for the research ideas, their implementation, and the paper writing, with the support and feedback of the other authors. Works marked with * are those in which technical solutions – contributed by my co-authors – are among the predominant aspects and/or contributions.

Paper.

- Beatrice Savoldi and Luisa Bentivogli. 2021. Gender Bias and Machine Translation: On first looking into parallel corpora. In Castagnoli, S., S. Bernardini, A. Ferraresi, M. Milicevic Petrovic (eds) *Using Corpora in Contrastive and Translation Studies Conference* (6th Edition). Bertinoro, Italy.

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

- Marco Gaido*, Beatrice Savoldi*, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to Split: the Effect of Word Segmentation on Gender Bias in Speech Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online. Association for Computational Linguistics.

- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the Morphosyntactic Lens: A Multifaceted Evaluation of Gender Bias in Speech Translation. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. On the Dynamics of Gender Learning in Speech Translation. *In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111, Seattle, Washington. Association for Computational Linguistics.

Resource documentation. The creation of new dedicated linguistic resources represents a fundamental contribution of the thesis. Toward the enablement of future research by fostering open research and data, all resources have been publicly released.[2] Beside their public release, however, the scientifically and ethically responsible use of available resources requires accessing their detailed and transparent documentation. For this reason, in the wake of the landscape of best practices and documentation toolkits emerging in the field of technologies, the resources presented in the thesis are made available with their corresponding *Data Statement*; namely "a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" (Bender and Friedman, 2018). The data statements were created following my (online) partícitipation at the

[2] They are all available at `https://ict.fbk.eu/units/hlt-mt/resources/`.

2020 LREC workshop "Data Statements for NLP: Towards Best Practice" organized by Emily M. Bender, Batya Friedman, and Angelina McMillan-Major.

- Data Statement for MuST-SHE. 2021. http://techpolicylab.uw.edu
- Data Statement for MuST-Speakers. 2021. https://ict.fbk.eu

Dissemination, teaching, invited talks. In light of the interdisciplinary nature of my PhD and the societal relevance of the topic it addresses, it is of great value to establish exchanges across disciplines and make scientific findings accessible. Below, I list the teaching and dissemination activities that have allowed establishing such exchanges across different communities. Although they are not directly embedded within the thesis, they have nonetheless inspired and shaped its form, and should thus be acknowledged.

For two years (Autumn-Winter 2021 and 2022), I designed and taught a course on *Machine Translation and CAT Tools* at the University of Trento. The course was taught to MA students in Euramerican Literature, Translation and Literary Critique, and intended to engage and introduce them to the use of (otherwise foreign) translation technologies. Despite the limited opportunities for on-site collaborations due to the COVID pandemic, in March-July 2022, I was able to carry out a visiting period at the University of Groningen, where I had the opportunity to work with Prof. Malvina Nissim to co-design and -teach a cutting edge course on *Ethical Aspects in Natural Language Processing* for BSc students of Information Science.

On several occasions, I have been invited to take part in either research-led or dissemination activities reaching out to different audiences. Below, I list the events where I discussed my work:

- *Good but not always Fair: Gender Bias in Automatic Translation*. Invited talk, Language in the Human-Machine Era (LITHME) COST Action. Online, February 2023.
- *Situating Knowledge and Technologies*. Invited talk and panelist, Picture a Scientist: Un dialogo su stereotipi e gender gap nelle discipline STEM *Panel*. Pisa, February 2023.
- *Gender Bias in Machine Translation: Taking stock of where we are (going)*. Invited Talk, Gender Equality and Artificial Intelligence Workshop, L'Aquila University. November 2022.
- *Ethical Aspects in NLP: A Teaching and Learning Experience*. Invited talk, Milano-NLP Lab: Coding Aperitivo. Bocconi University, September 2022.
- *Panel on Gender as a Variable in Natural Language Processing*. Invited panelist, Queer in AI Workshop. Seattle, July 2022.

- *Gender in Danger: Bias and Automatic Translation*. Invited Talk, Translating Europe Workshop (TEW): Translation and the Use of Inclusive Language. Online, December 2021.

- *Understanding Gender Bias in Translation Technologies*. Invited Talk, European Night of Researchers: Mind the Gap. University of Parma. Online, September 2021.

- *Gender Bias and Automatic Translation*. Invited talk, University of Groningen Computational Linguistics Seminar series. Online, May 2021.

Other works. Lastly, I have been involved in other collaborations related to the topics of this thesis. They have led to manuscripts that are to be peer-reviewed or are currently under peer-review. For this reason, they have not been included as an integral part of this work. Nonetheless, chunks of them are integrated within the thesis, which partly draws inspiration on them. Hence, I acknowledge them below.

The following paper is the result of collaborative effort with other PhD students and researches within the MT unit at Fondazione Bruno Kessler:

- Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Matteo Negri, Luisa Bentivogli. 2023. From Inclusive Language to Gender-Neutral Machine Translation. *arXiv:2301.10075*. https://arxiv.org/abs/2301.10075.

Since 2022, I have become a member of the COST action *Language In The Human-Machine Era (LITHME)*, which has set out the goals of *i)* preparing the field of linguistics and its subdisciplines for the changes in communicative practices brought on by emerging technologies, and *ii)* facilitate longer term dialogue between linguists and technology developers. Specifically, my fruitful collaboration within the LITHME working group on Language Ideologies has led to two papers to be included in a forthcoming edited volume:

- Beatrice Savoldi. 2022. Concepts of harms and bias in NLP and how they relate to language ideologies. In Britta Schneider and Bettina Migge (eds.) *Changing Language Ideological Concepts in the Human-Machine Era. Questions, Themes and Topics*, pages 13-18. https://lithme.eu/working-groups/wg6/.

- Beatrice Savoldi. 2022. Gender bias in language technologies. In Britta Schneider and Bettina Migge (eds.) *Changing Language Ideological Concepts in the Human-Machine Era. Questions, Themes and Topics*, pages 23-27. https://lithme.eu/working-groups/wg6/.