Lost in Translation?
Approaches to Gender Representation in Multilingual Archives

Mrinalini Luthra, Brecht Nijman
GLOBALISE, Huygens Insitute,
Koninklijke Nederlandse Akademie van Wetenschappen (KNAW),
Oudezijds Achterburgwal 185, 1012 DK Amsterdam
{mrinalini.luthra,brecht.nijman}@huygens.knaw.nl

Abstract

The GLOBALISE project's digitalisation of the Dutch East India Company (VOC) archives raises questions about representing gender and marginalised identities. This paper outlines the challenges of accurately conveying gender information in the archives, highlighting issues such as the lack of self-identified gender descriptions, low representation of marginalised groups, colonial context, and multilingualism in the collection. Machine learning (ML) and machine translation (MT) used in the digitalisation process may amplify existing biases and under-representation. To address these issues, the paper proposes a gender policy for GLOBALISE, offering guidelines and methodologies for handling gender information and increasing the visibility of marginalised identities. The policy contributes to discussions about representing gender and diversity in digital historical research, ML, and MT.

Disclaimer. In this paper, words and phrases presented in *"quotation marks and italicised"* are taken from the VOC archives. The records and metadata within these archives contain language and descriptions that are offensive, biased, or distorted. They reflect the prevailing societal attitudes of the VOC, and do not represent our views or those of our institution. Please be aware that engaging with this material may cause distress. We advise approaching the content with care and consideration.

1 Introduction

Gender has been the subject of much debate and analysis across various disciplines. In historical archives, gender representation often reflects the biases and power dynamics of the societies that produced them, leading to the marginalisation or erasure of non-normative gender identities. Attempting to describe various genders within historical contexts and across different languages and cultures, while beneficial, often simplifies complexities. Such reductions can inadvertently perpetuate (post)colonial and state-driven narratives of visibility, thus homogenising differences in time, place, and circumstances (Fanon, 1967; Dutta, 2013; Hinchy, 2019).

In parallel, the use of gender as a variable in artificial intelligence (AI) and machine learning (ML) within systems like recommender systems, information retrieval models, and machine translation is growing. However, there is a significant gap in critical analysis on how gender, especially non-binary identities, should be represented, taking into account intersectionality and the racialised nature of gender constructs (Pinney et al., 2023). Addressing gender biases is crucial both in historical archives and AI systems to avoid perpetuating existing inequalities and introducing new biases (Hicks, 2017; Noble, 2018).

The digitalisation (Brennen and Kreiss, 2016) of historical archives, combined with the application of machine learning and machine translation, presents unique challenges and implications for gender representation. This paper uses the GLOBALISE project as a case study to explore these issues. GLOBALISE aims to innovate historical research practices by creating an infrastructure that enables researchers and the public to access

Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies, pages 42–55
June 27, 2024 ©2024 European Association for Machine Translation

and explore the Dutch East India Company (VOC) archives, offering insights into the history of colonial expansion and the societies that endured and resisted the VOC's dominance. The project employs both historic and semantic contextualisation, in the form of entity recognition (ER) and event detection (ED), to enrich the archives with additional layers of information (Petram and van Rossum, 2022; Verkijk and Vossen, 2023).

While the primary focus of this paper is on the challenges of accurately representing gender in the context of ML, ER, and ED tasks, the insights gained are particularly relevant for machine translation (MT) as well. The inherently colonial nature of the VOC archives, combined with their complex historical context, the multilingual nature of the documents, and the absence of self-identified gender descriptions, poses significant challenges for accurately translating and representing gender across languages and cultures. Misrepresenting or erasing gender diversity in the translation process can further perpetuate the marginalisation of non-normative identities and distort historical narratives.

This paper thus attempts to answer the question posed in the title "Lost in Translation?" by grappling with the challenges of accurately translating and representing gender diversity across languages and cultures in the multilingual context of the GLOBALISE project and the VOC archives. The question highlights the potential for gender identities and expressions to be misinterpreted, oversimplified, or erased when historical documents are digitised and subjected to machine learning and translation processes. We explore these challenges and propose approaches to mitigate the potential loss or misrepresentation of gender diversity in the digitalisation and translation process.

The remainder of the paper is structured as follows: Section 2 presents a bias statement; Section 3 discusses the use of gender as an analytical variable; Section 4 provides a detailed discussion on the GLOBALISE project, examining the multilingual nature of the archives, gender representation and its challenges within the VOC archives, supplemented by specific examples; Section 5 outlines the first steps toward a gender policy; Section 6 concludes with a discussion on future work and broader implications.

2 Bias Statement

VOC archives present a significant challenge for contemporary researchers seeking to uncover the histories of marginalised communities. The vast majority of the records were created by European men employed by the VOC, reflecting their biases, interests, and the prevailing societal attitudes of the time (Wamelen, 2014; Meersbergen, 2017). Searching these records often leads to disappointment due to the violent categorisations of the past, which turned enslaved and colonised people into "nonpersons" (Hartman, 2008; Patterson, 2018; Fuentes, 2016; Zijlstra, 2021). While the colonised population left hardly any self-produced traces, the archive is full of records about them. However, due to the current organisation and accessibility of these archives, the experiences and perspectives of marginalised groups are not only underrepresented but also frequently misrepresented and, in most cases, extremely difficult to access (Trouillot, 2015). Taking from Bowker and Star's (2000) argument in their landmark work "Sorting Things Out: Classifications and their Consequences", the research infrastructures we create have the power to shape and reinforce social categories and power dynamics. As researchers creating an infrastructure to access colonial archives, we must critically examine our own practices to uncover these marginalised histories (Ghosh, 2004; Kars, 2020). Our approach is informed by personal experiences and academic backgrounds, which highlight the limitations and dangers of singular, totalising knowledge systems. This awareness underscores the importance of adopting pluralistic approaches that acknowledge the coexistence of diverse perspectives.

We acknowledge the potential for representational harm (Blodgett et al., 2020) in our work with the VOC archives. Marginalised groups, such as women, non-European actors, and individuals with non-binary genders, and other genders, appear only in traces within these colonial archives, often described by the colonial agents rather than represented in their own voices. This poses significant challenges in correctly attributing gender. The archives' inherent biases and the underrepresentation of these groups raise concerns that the developed information extraction models may fail to recognise them (under-representation) or attribute incorrect genders (such as stereotyping).

GLOBALISE is considering translating the

archives into various languages, such as Indonesian languages (Bahasa Indonesia, Javanese, Sundanese), Malay, Sinhala, Tamil, and Mandarin, to facilitate increased access. However, this process also carries the risk of perpetuating harm through translation. Misgendering or erasing diverse gender identities in the translated archives can further marginalise these communities and distort historical narratives.

The ramifications of such harms can be far-reaching, particularly for researchers and individuals from communities affected by Dutch colonialism, who may be attempting to write on marginalised histories or seek traces of their ancestors within these archives. Misrepresentation or erasure of their identities and experiences would perpetuate the very harms and marginalisation that these communities have and continue to endure.

3 To Gender or Not to Gender?

This section explores the potential benefits and drawbacks of using gender as a variable in the context of the GLOBALISE project and its analysis of the VOC archives, through historical research, machine learning, and machine translation.

3.1 Potential Benefits

1. Revealing Power and Marginalised Histories Gender is a critical category for understanding power dynamics in historical and cultural contexts. Applying gender as an analytical lens can provide insights into the social, economic, and power dynamics in different societies and time periods (Scott, 1986). Examining gender in conjunction with other identity categories such as race, socio-economic class, and nationality can reveal the complex ways in which power and privilege were and are distributed and experienced in context (Crenshaw, 1991). Focusing on gender can also help uncover the experiences and perspectives of women and other marginalised groups who may have been overlooked in traditional (historical) narratives (Luthra et al., 2023).

2. Auditing Bias in Machine Learning and Improving Translation Gender and other demographic variables can be useful to audit biases in machine learning systems. For instance, the incorporation of gender as a variable in ML models can help to uncover and mitigate biases in various domains, such as facial recognition systems, job recommendations, credit scoring, and healthcare (Buolamwini and Gebru, 2018; Omiye et al., 2023; Chen, 2023). Even when gender or other demographic features are not explicitly included in the data, ML models can still discriminate by picking up on proxy factors, as seen in Amazon's hiring algorithm that discriminated against women (Dastin, 2022).

In the field of machine translation, incorporating gender information can help produce more accurate and contextually appropriate translations. For example, in languages with grammatical gender, knowing the gender of the referent can help select the correct pronouns, adjective forms, and other gender-specific linguistic features (Vanmassenhove et al., 2018; Elaraby et al., 2018). Additionally, explicitly modeling gender in machine translation can identify and mitigate gender biases present in training data and algorithms (Saunders et al., 2020; Prates et al., 2020).

3.2 Possible Drawbacks

1. Anachronistic Categories and Limited Sources Applying modern understandings of gender to historical contexts risks imposing anachronistic categories and obscuring the specific ways in which gender was constructed and experienced in the past (Hartman, 2012). Colonial archives, such as those of the VOC, may not provide sufficient or unbiased information about the (gendered) experiences of all individuals and groups, particularly those who were marginalised or oppressed (Spivak, 1985; Jeurgens and Karabinos, 2020; Hinchy, 2022). Researchers must approach colonial archives critically, recognising their limitations and biases, and seeking to read between the lines and "along and against the grain" to uncover histories of gender (Stoler, 2008).

2. Reinforcing Binaries and Obscuring Intersectional Identities Relying solely on gender as a primary analytical category may inadvertently reinforce binary and essentialist notions of gender, failing to capture the diversity and fluidity of gender identities and expressions (Scott, 2010). The Hijra community in South Asia serves as a poignant example of the complexities surrounding gender identity. While sometimes referred to as the "third gender," this term is not without debate, as it risks oversimplifying the multifaceted nature of the Hijra identity. Focusing too heavily on gender alone may obscure other crucial axes of iden-

tity that the Hijra community holds dear, such as kinship, religion, class, and embodiment (Reddy, 2005). Moreover, an overemphasis on gender as a variable may also conceal other significant power relations and social categories that shaped the historical context of the VOC, including race, religion, and colonialism (Stoler, 2010). To fully understand the intricacies of identity and power dynamics in the VOC archives, it is essential to adopt an intersectional approach that considers the interplay between gender and other social categories.

3. Limitations of Machine-Learning and Translation The use of machine learning techniques to analyse historical documents and archives related to gender in the VOC context poses additional challenges. ML algorithms can perpetuate and amplify biases present in their training data (Noble, 2018; Buolamwini and Gebru, 2018) and may struggle to capture the nuances, ambiguities, and contextual factors crucial for understanding the complexities of gender in historical settings (Jo and Gebru, 2020). Many current approaches to incorporating gender in machine translation rely on binary gender classifications, which may not adequately capture the diversity of gender identities and expressions across cultures and languages (Savoldi et al., 2021; Saunders et al., 2020; Alhafni et al., 2020).

In conclusion, the use of gender as an analytical category in the GLOBALISE project and its analysis of the VOC archives presents both opportunities and challenges. As the project moves forward, it will be crucial to approach the use of gender as a variable with critical reflexivity, acknowledging the limitations and potential drawbacks while also leveraging its potential to uncover new insights and perspectives on the history of the VOC, colonialism, and globalisation.

4 Case Study: GLOBALISE

GLOBALISE aims to improve access to the Dutch East India Company archive through the creation of research infrastructure, which will offer an annotated machine-readable version of the *Letters and Papers Received* (OBP) section of VOC archives.[1] The OBP consists of the documents that the Dutch offices of the company received from its offices in its region of operation which ranged from the South African Cape to Japan.

The GLOBALISE project team consists of two main groups: historians who collect and curate reference data related to the collection, and a team responsible for training language models for entity recognition (ER) and event detection (ED). The entities identified in the archives will be linked to the curated reference data as well as existing reference vocabularies. Furthermore, the annotations will be interconnected to provide additional context for future users of the GLOBALISE research infrastructure. The annotators creating the ground truth data for the ER and ED models are from within the project team, ensuring a close collaboration between the historical and computational aspects of the project. This structure allows for a multidisciplinary approach to the digitalisation and enrichment of the VOC archives, leveraging the expertise of historians and computational linguists to create a comprehensive and accessible research resource. By providing annotated and contextualised data, the GLOBALISE project aims to facilitate new insights into the history of the VOC and its impact on the regions under its influence.

4.1 The Corpus

The OBP consists of approximately 5 million scans of handwritten material, making up 1,042,989,589 tokens.[2] It consists of a wide variety of documents, including but not limited to internal and external correspondence, resolutions, court cases, censuses, and summarising reports tying these together called the General Letters. TThe majority of these documents were written in Dutch by European men employed by the company (Meersbergen, 2017). Nevertheless, the archive is seeped through with languages other than Dutch, including many local non-European languages. In addition to a small series of letters sent over in their original language and a more substantial series of such letters in translation (±5% of documents)[3], the Dutch of the archive is laden with

[1] https://globalise.huygens.knaw.nl

[2] The most recent version of transcriptions can be accessed here: https://transcriptions.globalise.huygens.knaw.nl/. The whole corpus can also be downloaded at: https://hdl.handle.net/10622/LVXSBW.

[3] This is five percent of documents, not five percent of the corpus, and consists of 8214 documents marked as "Translaat" (translation) within the "Indigenous correspondence" section of the "Towards a New Age of Partnership" (TANAP) index of the OBP. This is by no means an exhaustive list of translations in the archive. Translated documents do not always carry "translation" in their title. Additionally translation appears in other forms as well.

a vocabulary originating from the languages of the region (Pepping, 2024).

4.2 Gender in the Corpus

Gender is rarely self-identified in the VOC archives and is usually assigned by third parties, such as scribes and translators, often in reductive and incorrect ways. This poses challenges for accurately representing gender in the GLOBALISE project, as the information reflects the assumptions and biases of the record creators rather than the lived experiences and identities of the individuals described. The lack of historical and cultural context further complicates the interpretation of gender, as the concepts of sex and gender may not have been distinguished by the Europeans writing these records.[4]

Nevertheless, some traces of gender identities outside the historical Western binary categorisations are found in the archives. Given these traces and challenges of self-identification, gender in the GLOBALISE project is taken as a construct encompassing both sex and gender, given that in most cases they are conflated and indistinguishable, and are mostly not based on self-identification but based on the assumptions of the writers of the archives. Moreover, care must be taken not to flatten these identities when approaching them through modern Western concepts of gender. Many of these identities had intrinsic relations to sacredness, positions at court or spiritual roles, and connections to social status and enslavement (Arvas, 2019; Andaya, 2018; Bowie, 2023; Hinchy, 2022; Ismoyo, 2020; Peletz, 2009).

4.2.1 Explicit Gender Indicators

term (Dutch)	term (English)	count
bisoe — bisoes — bissoe — bissoes	*bissu*	34
sida sida	*sida sida*	0
hijra — hisra	hijra	1
besnedene	"castrated"	152
eunuch — eunich	eunuch	65

Table 1: Occurrence of selected terms describing individuals outside the gender binary.

The most explicit form of gendering in the archive occurs where people are explicitly described as belonging to a particular gender (See Tables 2 and 1). This is most often a form of "man" or "woman" (see example (1)). Although mentions of genders outside the binary are very rare, they are sometimes explicitly referenced, though this does not necessarily mean they are acknowledged as distinct gender identities. Example (2) shows a case where *bissu*, one of the Bugis genders, are explicitly mentioned (Ismoyo, 2020). While the *Bissu* in this passage are described as distinct from men or women, they are still misgendered as *"men"* within the passage and their identity is conflated with sexual practice (*"knowledge"* in this case refers to sexual intercourse). In other cases, these identities are even further obscured, either by being grouped together into catch-all categories describing gendered or sexual "otherness", such as *"Eunuch"*, *"hermafrodiet"* (*"hermaphrodite"*), or *"besnedene"* (*"castrate"*), or by simply being subsumed into binary categorisations (Andaya, 2018; Gannon, 2011; Hinchy, 2017). As Table 1 shows, these terms are relatively more common compared to in-community terms such as *bissu* or *hijra*. Both are very rare compared to terms referencing binary gender (Table 2). Additionally, these gender identities may be described only through another one of their identity axes, for instance solely as "priests".

(1) *6: **manspersoonen** ende een hollantsche **vrouw***[5]

6: **man** persons and a Hollandish **woman**

(2) *Buijten nog eenige sleep van ruijm 200. Coppen zoo **mans** als **vrouwen** die haar in de [z]aal en buijten op de stoep nedersette[,] ongerekent nog 20. **bissoes** of zoo genaamde mannen die de bekenning der vrouwen zouden hebben afgeswooren* [6]

Outside some more followers, roughly 200 heads, **men** as well as **women**, who waited in the hall as well as outside on the street, uncounted another 20 **bissu** or so-called men who have sworn off the knowledge of women.[7]

[4]This is further complicated by the fact that both sex and gender are socially constructed (Browne, 2010).

[5]National Archives (NA), the Hague, Archive of the Verenigde Oost-Indische Compagnie (VOC), 1.04.02, inventory number 1121, p. 833, https://transcriptions.globalise.huygens.knaw.nl/detail/urn:globalise:NL-HaNA_1.04.02_1121_0060

[6]NA, VOC, 8194, fo. 205. https://transcriptions.globalise.huygens.knaw.nl/detail/urn:globalise:NL-HaNA_1.04.02_8194_0213

[7]NA, VOC, 8194, fo. 205r. https://transcriptions.globalise.huygens.knaw.

term (English)	masculine term (Dutch)	count	feminine term (Dutch)	count
person	mansper*	1548	vrouwsper*	1642
"man slave" / "woman slave"	mansla*f — manslav*	3450	vrouwsla*f — vrouwslav*	69
"slave"	sla*f — slav*	167114	slavin*	25358
farmer	boer	3230	boerin	67
encik	intje	13589	–	–
njai	–	–	njaij — njeij	601
widow(er)	weduwna*r	20	weduwe	22099
son / daughter	zoon — soon	79659	dogter — dochter — doghter	19266
king / queen	koning — coning — coninck	304109	koningin — coningin	3407

Table 2: Occurrence of selected gendered terms.

4.2.2 Personal Nouns as Gender Indicators

As Dutch is a grammatical gender language, the archive is in no shortage of explicit gender-markers. Personal nouns in particular carry potentially valuable information on (perceived) social gender. A term's grammatical gender does not always coincide with the social gender associated with it. For instance, *"wijf"*, (wife or woman) is grammatically neuter, but socially feminine. However, in the majority of cases grammatical and social gender of personal nouns in Dutch align. See for instance example (3)).

(3) *"12* ***boeren*** *en* ***boerinnen****"*.[8]
12 **farmers**$_M$ and **farmers**$_F$

This approach also carries a number of pitfalls. First, following on from the previous point, these forms make it nearly impossible to recognise any gender that falls outside the man–woman binary. Second, commonly occurring issues regarding gender in language, such as masculine generics (see example (4)) and marked femaleness (see example (5)), also complicate this strategy. As Stahlberg et al. (2007) point out, it cannot be generally assumed that by using the masculine, particularly the masculine plural, the author considered an individual to be a man (let alone how that individual identified). At the same time, annotating only non-masculine terms reinforces the "othering" of women and genders beyond the binary. Finally, as mentioned, a small but non-negligible number of documents are translations of documents received in other languages (this does not include translations of spoken accounts or summaries of in-person interactions in other languages). Many of these are translations from genderless languages such as Malay or Javanese, and gender may have been introduced in the process of translation.

(4) *"de principaalste* ***actrice*** *van deese* ***gaauwdieven*** *troep"*.[9]
the principal **actress** of this gang of **thieves**$_M$.

(5) *"en Conting groot 15. Coij[ang]s bem[an]t met 8. chineesen en 12. javanen waeronder een* ***vrouwspersoon****"*.[10]
a kunting, large 15 koj[ang]s, manned by 8 chinese and 12 javanese including a **woman** person.

4.2.3 Loan Words as Gender Indicators

Loanwords are words adopted from one language into another without translation, often as a result of cultural contact or influence (Durkin, 2014). In colonial archives, loanwords originate from interactions between colonisers and indigenous populations, serving to facilitate communi-

nl/detail/urn:globalise:NL-HaNA_1.04.02_8194_0213

[8]NA, VOC, 4074, fo. 16r, https://transcriptions.globalise.huygens.knaw.nl/detail/urn:globalise:NL-HaNA_1.04.02_4074_0022.

[9]NA, VOC, 10936, https://transcriptions.globalise.huygens.knaw.nl/detail/urn:globalise:NL-HaNA_1.04.02_10936_0243.

[10]NA, VOC, 1945, 75, https://transcriptions.globalise.huygens.knaw.nl/detail/urn:globalise:NL-HaNA_1.04.02_1945_0086.

cation and reflect the integration of indigenous systems into colonial structures. They offer insights into cultural exchange, administrative integration, and power dynamics within colonial societies (Naregal, 1999; Cohn, 1996). For instance, "Baboe", from Javanese and Malay (also spelled as "babu") originally referred to a female servant or domestic worker in Southeast Asian societies. In Dutch colonial households, "baboes" were often employed to perform domestic chores and childcare duties for Dutch families. The use of this term in colonial archives highlights the hierarchical relationship between Dutch employers and indigenous domestic workers, reflecting the social stratification based on race and class in colonial society.[11] Moreover, many of the gendered identities in the VOC archives such as "bissu" are also loanwords.

Loanwords, particularly those used as terms of address, titles, and professions, constitute the final group of gender indicators to consider in the archive. Not considering loanwords while considering gendered words in Dutch would result in a disparity between the gendering of indigenous individuals and Europeans in the corpus. Furthermore, one could argue that these terms are more likely to reflect personal identity, though they may still be assigned from within the same language group. A particular pitfall with these terms is that they tend not to be recognised, even by human annotators, often being identified as parts of names. Identifying gendered loanwords can be particularly difficult due to several factors. Firstly, they originate from hundreds of languages throughout the region. Secondly, they are often poorly transliterated using Early Modern Dutch spelling, at times rendering them unrecognisable even to (native) speakers. For instance "Encik" (Mr. in Malay) is commonly written as "Intje" in the corpus. Lastly, successful identification of gendered loanwords is limited further where languages which have become endangered or even extinct, oftentimes in direct result to violence enacted by the VOC (Pepping, 2024). Additionally, care should be taken not to introduce gender to titles which do not explicitly carry it. Many forms of address might say more about class, caste, race or closeness than they do about gender (Yusra et al., 2023). Historically, gender neutral titles have been glossed in explicitly gendered ways.

Note that names have not been listed as a gender indicator in this section. Names are dubious carriers of gender in any context, and only more so in a multilingual one (Das and Paik, 2021; Saunders and Olsen, 2023). The same name may have very different associations across languages. Additionally, the Early Modern Dutch transliterations of names may render them indistinguishable or unrecognisable.

5 Developing a Gender Policy for GLOBALISE

GLOBALISE aims to develop a gender policy that respectfully represents the diversity of gender identities and experiences within the VOC archives. This policy will serve as a framework for addressing the challenges and limitations of working with historical sources, where gender information may be incomplete, biased, or absent. The gender policy will guide the project's approaches to data collection, annotation, analysis, and interpretation, ensuring that the resulting research infrastructure is sensitive to the complexities of gender across different historical and cultural contexts.

Principles of the GLOBALISE gender policy include:

1. Recognise the historical and cultural specificity of gender categories and expressions
2. Acknowledge the limitations and biases inherent in historical sources, particularly those created within colonial contexts
3. Strive to represent gender diversity in a manner that is respectful, accurate, and inclusive
4. Engage with relevant communities, scholars, and stakeholders to inform the development and implementation of the policy
5. Ensure transparency and accountability in the project's handling of gender-related information

The remaining section outlines guidelines and strategies for handling gender-related information in the GLOBALISE project.

[11] Note: It's worth acknowledging that some of these loanwords have become fully integrated into the colonial language itself, reflecting the enduring influence of colonial interactions on linguistic evolution. For instance, words like "loot" (derived from the Hindi word "lut") meaning "plunder" and "jungle" (originating from the Hindi word meaning "dense forest") are now commonplace in English vocabulary, serving as reminders of the historical connections between colonial past and contemporary language usage.

5.1 Polyvocal Gender Vocabulary

To address the diverse gender identities in the VOC archives, as discussed in section 4, GLOBALISE adopts a "polyvocal gender vocabulary" (Peletz, 2009). This approach allows for the inclusion of multiple gender classifications within a single knowledge organisation system (Tudhope and Lykke Nielsen, 2006; Hjørland and Gnoli, 2016), enabling the representation of historical and cultural specificities of gender. By employing this method, GLOBALISE aims to avoid imposing anachronistic or Western-centric categories onto the historical records while making gender diversity visible and searchable. The project seeks to represent gender categories from various cultures on their own terms, rather than "being through others" (Fanon, 1967).

The polyvocal gender vocabulary draws on the concept of "polyvocality" or "polyphony" (Bakhtin, 1984), which is a narrative feature that emphasises the simultaneous inclusion of multiple voices and perspectives. This approach aligns with the concept of "practical ontology" developed in anthropology and science and technology studies (STS), which recognises the coexistence of multiple, culturally-specific ways of understanding and categorising the world (Gad et al., 2015; Barth, 1993; Geertz, 1973). By adopting a polyvocal gender vocabulary, GLOBALISE aims to represent the diverse and "situated" (Haraway, 2016) understandings of gender present in the VOC archives, while acknowledging the challenges and limitations of working across multiple cultural and historical contexts.

5.1.1 Detecting Gendered Terms and Constructing a Polyvocal Gender Vocabulary

To address the challenges of detecting gendered terms in the VOC archives, GLOBALISE will employ an iterative process involving vocabulary development, manual annotation, and machine learning, as illustrated in Figure 1.

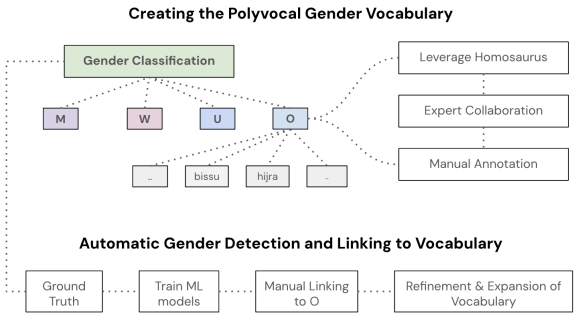

Figure 1: Creating the Polyvocal Gender Vocabulary and Automatic Gender Detection

Gender Classification The project will adopt a four-level hierarchy with the following top-level categories: M (man), W (woman), U (undefined, cannot infer gender information from the text), and O (other gender categories). The "O" category, an intermediate step, will be further subdivided manually into more specific gender identities based on the expertise of historians and cultural experts, allowing for granular representation while ensuring the machine learning models will not ignore and misrepresent and misclassify these other gender categories. This decision is based on our experience that marginalised groups are mentioned in the VOC archives but only in low frequencies and often with wrong and insufficient contextual information.[12]

Creating the Gender Vocabulary The gender vocabulary will be developed through an iterative process involving the use of existing linked data vocabularies, collaboration with experts, and manual annotation.

1. **Leverage the Homosaurus**: GLOBALISE will leverage existing vocabularies such as the Homosaurus (Homosaurus Editorial Board, 2019), a linked data vocabulary of LGBTQ+ terms developed for cultural heritage institutions, as a starting point. While the Homosaurus focuses on modern terminology, its principles of providing a standardised yet inclusive ontology for gender diversity will inform the development of GLOBALISE's gender vocabulary. However, the project will adapt these principles to address the specific challenges posed by the VOC archives, such as the lack of self-identified gender information and the presence of historical terms and categories that may not map neatly onto contemporary understandings of gender.[13]

2. **Collaborate with Experts**: GLOBALISE

[12] So far, we have only found about 5 instances of the "bissu" in the archives and their descriptions in the archive are reductive and incorrect.

[13] Terms like "Hijra" and "bissu" are present in the Homosaurus and can be used to check their presence in the GLOBALISE corpus.

will work with gender scholars, communities and individuals from different gender groups, and historians specialising in regions covered by the VOC to create, refine, and expand the gender vocabulary. These experts will help identify culturally-specific gender terms and categories relevant to the historical context of the VOC archives and the early modern "Indian Ocean world".

3. **Manual annotation and Iterative Refinement**: Using the initial gender vocabulary, annotators will manually label a subset of the VOC archives with gender information, employing the hierarchical gender classification scheme defined earlier. During the annotation process, new gendered terms are expected to be identified and added to the vocabulary.[14] This requires that annotators have the relevant linguistic background to recognise these terms.[15]

Automatic Gender and Linking to Gender Vocabulary With the manually annotated dataset serving as the ground truth, GLOBALISE will employ automatic gender detection using machine learning models to identify gendered terms at scale in the VOC archives, as illustrated in Figure 1.

1. **Training machine learning models**: The manually annotated dataset will be used to train machine learning models, such as entity recognition systems (Ehrmann et al., 2023), to automatically detect gendered terms in the larger corpus. The machine learning models will assign one of the four top-level gender categories (M, W, U, or O) to each detected gendered term.

2. **Applying the models to the full corpus**: The trained machine learning models will be applied to the entire VOC archive to automatically detect and classify gendered terms at scale.

3. **Manual linking to specific categories**: After the automatic gender detection process, the gendered terms classified as "O" will be examined and manually linked to more specific gender categories in the gender vocabulary by GLOBALISE's researchers, experts, and community members. This manual linking process allows for a more granular representation of non-binary and culturally-specific gender identities while ensuring that the machine learning models do not overlook these categories. Moreover, this step allows us to audit the outputs of the machine learning models to avoid misgendering (Kotek et al., 2023; Bender et al., 2021; Hamidi et al., 2018), especially given that descriptions that follow or precede gendered terms in the VOC archives, are often incorrect or reductive as explained in the example of "bissu" in subsection 4.2.

4. **Iterative refinement**: As new gendered terms are discovered during the automatic gender detection and manual linking processes, they will be incorporated into the gender vocabulary and used to refine the machine learning models. This iterative refinement ensures that the polyvocal gender vocabulary remains accurate, comprehensive, and culturally sensitive.

5.2 Gendered "Loanwords", and their Translations

As discussed in Section 4.2.3 on loanwords, GLOBALISE will pay close attention to the presence of loanwords in the corpus. Annotators will be trained in detecting loanwords, and the project will benefit from annotators of diverse cultural and historical backgrounds as well as insights from local communities and experts. Additionally, the project will investigate the use of multilingual models such as BERT (2018) and translation technologies to automatically detect loanwords in the VOC archives (Nath et al., 2022).

Recent research has emphasised the importance of considering gender diversity in machine translation, as neural machine translation systems often perpetuate gender biases and fail to accurately translate gender-neutral or non-binary language (Vanmassenhove et al., 2018; Savoldi et al., 2021). By developing a "polyvocal gender vocabulary," GLOBALISE can contribute to more gender- and culturally-sensitive machine translation by introducing contextual aspects of gender.

However, when translating the VOC archives, the project must also consider the presence of gen-

[14] Also based on our experience developing reference data on commodities in the GLOBALISE project (Pepping et al., 2023).

[15] Annotators without such knowledge are more likely to mistake gendered loanwords as names or (merely) professions.

dered loanwords, which reflect the complex linguistic and cultural dynamics of the colonial encounter. These loanwords often carry the "hierarchies of power" (Naregal, 1999) that characterized colonial bilingualism, and simply translating them into other languages risks overwriting or erasing important historical and cultural nuances. To address this issue, the GLOBALISE project's initial strategy will be to preserve gendered loanwords in their original form when translating the VOC archives into other languages. By retaining these loanwords, the project aims to maintain the visibility of the complex cultural and linguistic exchanges that took place in early modern history while providing context and explanations to help readers understand their meanings and connotations.

5.3 Gender Identification

As discussed in subsection 4.2, GLOBALISE treats gender as a complex construct encompassing both sex and gender, which are often conflated and indistinguishable in third-person historical records without self-identification.

Avoiding Gender Assignment Based on Names The project avoids assigning gender based solely on names as this can introduce biases (Luthra et al., 2023) and unwarranted assumptions about individuals' identities (Savoldi et al., 2021), especially for non European cultures, beyond that was already done in the creation of the records (Das and Paik, 2021). Instead, the project will rely on explicit references to gender or contextual "trigger words" (Ehrmann et al., 2023), such as titles, roles, and gendered nouns, to infer gender when possible. These include terms such as koningin" (queen), sultana," mevrouw" (madam), meneer" (sir), radja" (king), rani" (queen), coopvrouw" (merchant woman), priesteressen" (priestesses), "weduwe" (widow), "slavinne" (female slave), "capados", eunuch", and "bissu."

However, the project acknowledges that these references often reflect the assumptions and biases of the record creators rather than the self-identified gender of the individuals described. A not very straightforward example of this is the case of Matthias Panholsser, with a stereotypical Dutch male name, one of 52 persons in the VOC Opvarenden[16] [VOC Sailors], who was dismissed on the grounds of being a "vrouw" (woman). Here, we do not want to conclude that Matthias was a trans-man at the risk of "trans-ing" history (Hinchy, 2022). Perhaps Matthias only disguised as a man to serve on the VOC ships, looking for better economic opportunities. But of this we cannot know, due to the insufficient information, and thus adding our own interpretation. We return to the case of Matthias breifly in 5.4.

Handling Grammatical Gender in Dutch As Dutch is a grammatical gender language, personal nouns can offer valuable perceived gender information. However, the project will be mindful of the pitfalls associated with grammatical gender, such as masculine generics potentially obscuring women and non-binary individuals, and the "othering" reinforced by only annotating non-masculine terms (Stahlberg et al., 2007), as discussed in examples (4) and (5).

5.4 Modeling Gender Fluidity

Gender reassignments and gender fluidity have existed throughout history, as exemplified by the Hijras in South Asia (Hinchy, 2022; Reddy, 2005). However, current systems for classifying gender often fail to capture this reality, employing static categorisations that inadequately represent how individuals' genders can shift over their lifetimes (Andrews et al., 2024). Recognising this limitation, the GLOBALISE project aims to develop methods for modeling gender fluidity within the VOC archives to better represent changes in gender identity over time.

One approach is to use event-based modeling, where gender is treated as a temporal attribute that can change at specific points in an individual's life. This allows for the representation of gender transitions and the evolution of an individual's gender identity (Andrews et al., 2024). By employing event-based modeling, the project can capture pivotal moments when an individual's expressed or documented gender may have shifted, potentially due to societal pressures, personal needs, or changing circumstances. While the archival records may only provide a limited, biased glimpse into an individual's gender experience, modeling gender fluidity as a series of events acknowledges the possibility of more complex gender journeys than what is immediately apparent. This way, the project can

[16] https://www.nationaalarchief.nl/onderzoeken/index/nt00444/cc19411e-c864-11e6-9d8b-00505693001d

shed light on the diverse gender expressions, nonconformities, and transgressions (for instance the case of Matthias Panholsser who was dismissed for "being a woman") that have existed throughout history (Nationaal Archief, nd).

5.5 Evaluation and Intersectionality

To ensure the accuracy and fairness of the gender vocabulary and detection methods, GLOBALISE will conduct regular evaluations and audits. One important aspect of this evaluation is studying bias along intersectional axes, such as race, ethnicity, and gender (Haim et al., 2024). By examining the interactions between these different dimensions of identity, the project can identify and mitigate potential biases in the gender classification and detection systems. The evaluation process will involve collaboration with experts in gender studies, history, and cultural heritage, as well as members of affected communities.

6 Conclusion, Discussions, Future Work

This paper encapsulates the central challenges we face in the GLOBALISE project in attempting to accurately representing and translating the diverse gender identities and expressions found within the multilingual VOC archives. Through this work, we aim to contribute to the broader discussion on the challenges of detecting gender at scale in multilingual and multicultural corpora. This paper has outlined the problems in our endeavor and proposed strategies to address them, grappling with the complex issues of navigating linguistic and cultural boundaries while striving for respectful and useful representations.

These challenges are not unique to GLOBALISE; they apply to those working in digital humanities and those working with socially constructed data in fields like machine learning and machine translation. Representing gender diversity across historical, linguistic, and cultural contexts in a culturally sensitive and computationally feasible manner is a broader issue that requires ongoing exploration and dialogue. We hope that this paper can contribute to informing and advancing these broader discussions and practices around the representation of gender diversity. As we continue to develop and refine our methods for representing gender diversity in the GLOBALISE project, we welcome input and collaboration from researchers and communities working on similar challenges in other domains.

In terms of future work, the GLOBALISE project will start with the creation of the polyvocal gender vocabulary and initial attempts at gender detection in the VOC archives. We plan to collaborate with area studies specialists to think about gender from the various regions once under the VOC empire, ensuring that our approaches are culturally informed and sensitive to the specific historical and linguistic contexts of the archives.

Acknowledgements: This publication is part of the project GLOBALISE (with project number 175.2019.003) of the Research Infrastructure research program financed by the Dutch Research Council (NWO). We extend our gratitude to the entire GLOBALISE team, with special thanks to the historical contextualisation team for their insightful discussions on gender classifications and their consequences. We would also like to express our appreciation to Sofia Coppola who inspired the title of this article with her most excellent film. Finally, we extend our appreciation to Kevin, Asawari, Julia, and Lolo for their invaluable support throughout the process.

References

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In Costa-jussà, Marta R., Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Andaya, Leonard Y. 2018. The *Bissu*: Study of a third gender in indonesia. In Zamfira, Andreea, Christian de Montlibert, and Daniela Radu, editors, *Gender in Focus: Identities, Codes, Stereotypes and Politics*, pages 62–87. Barbara Budrich, Opladen, German.

Andrews, Tara L, Marius Deierl, and Carla Ebel. 2024. Gender Assignment as an Event—a Contemporary Approach for the Adequate Depiction of Historical Gender Categories. *Digital Scholarship in the Humanities*, 39(1):5–12.

Arvas, Abdulhamit. 2019. Early modern eunuchs and the transing of gender and race. *Journal for Early Modern Cultural Studies*, 19(4):116–136.

Bakhtin, Mikhail. 1984. *Problems of Dostoevsky's poetics*. University of Minnesota Press, Minneapolis, Minnesota, United States.

Barth, Fredrik. 1993. *Balinese worlds*. University of Chicago Press, Chicago, Illnois, United States.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

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

Bowie, Katherine A. 2023. Eunuchs in burmese history: An overview. *Journal of Southeast Asian Studies*, 54(4):621–644.

Bowker, Geoffrey C and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

Brennen, J Scott and Daniel Kreiss. 2016. Digitalization. *The international encyclopedia of communication theory and philosophy*, pages 1–11.

Browne, Simone. 2010. Digital epidermalization: Race, identity and biometrics. *Critical Sociology*, 36(1):131–150.

Buolamwini, Joy and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Chen, Zhisheng. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1):567, September.

Cohn, Bernard S. 1996. *Colonialism and Its Forms of Knowledge*. Princeton University Press, Princeton, NJ.

Crenshaw, Kimberle. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Das, Sudeshna and Jiaul H Paik. 2021. Context-sensitive gender inference of named entities in text. *Information Processing & Management*, 58(1):102423.

Dastin, Jeffrey. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. In Martin, Kirsten, editor, *Ethics of Data and Analytics*. Auerbach Publications, Boca Raton, Florida, United States. Num Pages: 4.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Durkin, Philip. 2014. *Borrowed Words: A History of Loanwords in English*. Oxford University Press, January.

Dutta, Aniruddha, 2013. *An Epistemology of Collusion: Hijras, Kothis and the Historical (Dis) Continuity of Gender/Sexual Identities in Eastern India*, chapter 14, pages 305–329. John Wiley & Sons, Ltd.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Elaraby, Mostafa, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.

Fanon, Frantz. 1967. *Black skin, white masks*. Grove Press, New York.

Fuentes, Marisa J. 2016. *Dispossessed lives: Enslaved women, violence, and the archive*. University of Pennsylvania Press.

Gad, Christopher, Casper Bruun Jensen, and Brit Ross Winthereik. 2015. Practical Ontology: Worlds in STS and anthropology. *NatureCulture*, (3):67–86.

Gannon, Shane. 2011. Exclusion as language and the language of exclusion: Tracing regimes of gender through linguistic representations of the "eunuch". *Journal of the History of Sexuality*, 20(1):1–27.

Geertz, Clifford. 1973. *The interpretation of cultures*. Basic books, New York City, New York, United States.

Ghosh, Durba. 2004. Decoding the nameless: gender, subjectivity, and historical methodologies in reading the archives of colonial india. *A New Imperial History: Culture, Identity, and Modernity in Britain and the Empire*, pages 1660–1840.

Haim, Amit, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias.

Hamidi, Foad, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. 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, page 1–13, New York, NY, USA. Association for Computing Machinery.

Haraway, Donna. 2016. 'situated knowledges: The science question in feminism and the privilege of partial perspective'. In *Space, gender, knowledge: Feminist readings*, pages 53–72. Routledge, London, United Kingdom.

Hartman, Saidiya. 2008. Venus in two acts. *Small Axe: A Caribbean Journal of Criticism*, 12(2):1–14.

Hartman, Saidiya. 2012. The time of slavery. In *Enchantments of Modernity*, pages 447–468. Routledge India.

Hicks, Mar. 2017. *Programmed inequality: How Britain discarded women technologists and lost its edge in computing*. MIT press, Cambridge, Massachusetts, United States.

Hinchy, Jessica. 2017. The eunuch archive: Colonial records of non-normative gender and sexuality in india. *Culture, Theory and Critique*, 58(2):127–146.

Hinchy, Jessica. 2019. *Governing Gender and Sexuality in Colonial India: The Hijra, c.1850–1900*. Cambridge University Press.

Hinchy, Jessica Bridgette. 2022. Hijras and south asian historiography. *History Compass*, 20(1):e12706.

Hjørland, Birger and Claudio Gnoli. 2016. Isko encyclopedia of knowledge organization.

Homosaurus Editorial Board. 2019. Homosaurus: An international lgbtq linked data vocabulary. `http://homosaurus.org/`.

Ismoyo, Petsy Jessy. 2020. Decolonising gender identities in Indonesia: A study of bissue 'the transreligious leader' in bugis people. *Paradigma: Jurnal Kajian Budaya*, 10(3):277–288.

Jeurgens, Charles and Michael Karabinos. 2020. Paradoxes of curating colonial memory. *Archival Science*, 20(3):199–220, September.

Jo, Eun Seo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Kars, Marjoleine. 2020. *Blood on the River: A Chronicle of Mutiny and Freedom on the Wild Coast*. The New Press, New York City, New York, United Staes.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Luthra, Mrinalini, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*.

Meersbergen, Guido van. 2017. Writing east india company history after the cultural turn: Interdisciplinary perspectives on the seventeenth-century east india company and verenigde oostindische compagnie. *Journal for Early Modern Cultural Studies*, 17(3):10–36.

Naregal, Veena. 1999. Colonial bilingualism and hierarchies of language and power: Making of a vernacular sphere in western india. *Economic and Political Weekly*, pages 3446–3456.

Nath, Abhijnan, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In Calzolari, Nicoletta, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Nationaal Archief. n.d. VOC: Opvarenden, 1699 - 1794.

Noble, Safiya Umoja. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York University Press.

Omiye, Jesutofunmi A., Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1):195, October.

Patterson, Orlando. 2018. *Slavery and social death: A comparative study, with a new preface*. Harvard University Press.

Peletz, Michael G. 2009. *Gender Pluralism: Southeast Asia Since Early Modern Times*. Routledge.

Pepping, K., H. Vellinga, M. Kuruppath, L. Van Wissen, and M. Van Rossum. 2023. GLOBALISE Thesaurus - Commodities.

Pepping, K. W. 2024. Reflections on language tagging: working with the multilingual dimension of the Dutch East India Company archives. *Journal of Open Humanities Data*, 10(29):1–10.

Petram, Lodewijk and Matthias van Rossum. 2022. Transforming historical research practices – a digital infrastructure for the voc archives (globalise). *International Journal of Maritime History*, 34(3):494–502.

Pinney, Christine, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much ado about

gender: Current practices and future recommendations for appropriate gender-aware information access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, page 269–279, New York, NY, USA. Association for Computing Machinery.

Prates, Marcelo O. R., Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Comput. Appl.*, 32(10):6363–6381, may.

Reddy, Gayatri. 2005. *With Respect to Sex: Negotiating Hijra Identity in South India*. University of Chicago Press, Chicago, IL, July.

Saunders, Danielle and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In Vanmassenhove, Eva, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93. European Association for Machine Translation.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In Costa-jussà, Marta R., Christian Hardmeier, Will Radford, and Kellie Webster, editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Scott, Joan W. 1986. Gender: A useful category of historical analysis. *The American Historical Review*, 91(5):1053–1075.

Scott, Joan Wallach. 2010. Gender: Still a useful category of analysis? *Diogenes*, 57(1):7–14.

Spivak, Gayatri Chakravorty. 1985. The rani of sirmur: An essay in reading the archives. *History and theory*, 24(3):247–272.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes. In Fiedler, Klaus, editor, *Social Communication*, pages 163–187. Psychology Press.

Stoler, Ann Laura. 2008. *Along the archival grain: Epistemic anxieties and colonial common sense*. Princeton University Press.

Stoler, Ann Laura. 2010. *Carnal Knowledge and Imperial Power: Race and the Intimate in Colonial Rule, With a New Preface*. University of California Press, Berkeley, Los Angeles, United States; London, United Kingdom.

Trouillot, Michel-Rolph. 2015. *Silencing the past: Power and the production of history*. Beacon Press.

Tudhope, Douglas and Marianne Lykke Nielsen. 2006. Introduction to knowledge organization systems and services. *New Review of Hypermedia and multimedia*, 12(1):3–9.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Verkijk, Stella and Piek Vossen. 2023. Sunken ships shan't sail: Ontology design for reconstructing events in the dutch east india company archives. In *CEUR Workshop Proceedings*, pages 320–332. CEUR Workshop Proceedings.

Wamelen, Carla van. 2014. *Family life onder de VOC: Een handelscompagnie in huwelijks- en gezinszaken*. Uitgeverij Verloren.

Yusra, Kamaludin, Yuni Budi Lestari, and Jane Simpson. 2023. Borrowing of address forms for dimensions of social relation in a contact-induced multilingual community. *Journal of Politeness Research: Language, Behavior, Culture*, 19(1):217–248.

Zijlstra, Suzanne. 2021. *De voormoeders: een verborgen Nederlandse-Indische familiegeschiedenis*. Ambo Anthos, Amsterdam, The Netherlands.