Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text

Lucy Havens[†] **Melissa Terras** [‡] **Benjamin Bach**[†] **Beatrice Alex**[§†]

[†]School of Informatics

[‡]College of Arts, Humanities and Social Sciences

[§]Edinburgh Futures Institute; School of Literatures, Languages and Cultures

University of Edinburgh

lucy.havens@ed.ac.uk, m.terras@ed.ac.uk

bbach@inf.ed.ac.uk, balex@ed.ac.uk

Abstract

Mitigating harms from gender biased language in Natural Language Processing (NLP) systems remains a challenge, and the situated nature of language means bias is inescapable in NLP data. Though efforts to mitigate gender bias in NLP are numerous, they often vaguely define gender and bias, only consider two genders, and do not incorporate uncertainty into models. To address these limitations, in this paper we present a taxonomy of gender biased language and apply it to create annotated datasets. We created the taxonomy and annotated data with the aim of making gender bias in language transparent. If biases are communicated clearly, varieties of biased language can be better identified and measured. Our taxonomy contains eleven types of gender biases inclusive of people whose gender expressions do not fit into the binary conceptions of woman and man, and whose gender differs from that they were assigned at birth, while also allowing annotators to document unknown gender information. The taxonomy and annotated data will, in future work, underpin analysis and more equitable language model development.

1 Background and Introduction

The need to mitigate bias in data has become urgent as evidence of harms from such data grows (Birhane and Prabhu, 2021; O'Neill et al., 2021; Perez, 2019; Noble, 2018; Vainapel et al., 2015; Sweeney, 2013). Due to the complexities of bias often overlooked in Machine Learning (ML) bias research, including Natural Language Processing (NLP) (Devinney et al., 2022; Stańczak and Augenstein, 2021), Blodgett et al. (2020), Leavy (2018), and Crawford (2017) call for greater interdisciplinary engagement and stakeholder collaboration. The Gallery, Library, Archive, and Museum (GLAM) sector has made similar calls for interdisciplinary engagement, looking to applications of data science and ML to better understand and mitigate bias in GLAM collections (Padilla, 2017, 2019; Geraci, 2019). Supporting the NLP and GLAM communities' shared aim of mitigating the minoritization[1] of certain people that biased language causes, we provide a taxonomy of gender biased language and demonstrate its application in a case study with GLAM documentation.

We use *GLAM documentation* to refer to the descriptions of heritage items written in GLAM catalogs. Adapting our previously published definition, we use *gender biased language* to refer to "language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their [gender] identity; and privileging other people through words or phrases that favor their [gender] identity" (Havens et al., 2020, 108). We focus on gender bias due to the contextual nature of gender and bias (they vary across time, location, culture, and people), as well as the existing efforts of our partner institution, the Archives of the Centre for Research Collections at the University of Edinburgh, to mitigate gender bias in its documentation.

GLAM documentation provides a unique benefit compared to many text sources: it contains historical and contemporary language. GLAM continually acquire and describe heritage items to enable the items' discoverability. In archives, heritage items include photographs, handwritten documents, instruments, and tweets, among other materials. Heritage items and the language that describes them influence society's understanding of the past,

[1]This paper uses *minoritization* in the sense D'Ignazio and Klein (2020) use the term: as a descriptor to emphasize a group of people's experience of oppression, rather than using the noun *minority*, which defines people as oppressed.

the present, and the direction society is moving into the future (Benjamin, 2019; Welsh, 2016; Yale, 2015; Cook, 2011; Smith, 2006). Through research with GLAM documentation, variations in biased language could be better understood. Should diachronic patterns emerge, the NLP community could train models to identify newly-emerging, previously unseen types of bias.

This paper presents an annotation taxonomy (§5) to label gender biased language inclusive of trans and gender diverse identities,[2] as well as a dataset of historical and contemporary language from British English archival documentation annotated according to the taxonomy. Linguistics, gender studies, information sciences, and NLP literature inform the taxonomy's categorization of gender biased language. As a result, the taxonomy holds relevance beyond the GLAM sector in which we situate our work. The taxonomy may be applied when creating NLP datasets or models, or when measuring varieties of gender bias in language, because the taxonomy's definitions of types of gender biases are rooted in the language of text, rather than an abstracted representation of text. Uniquely, our taxonomy includes labels that record uncertainty about a person's gender.

As we situate our work in the GLAM sector, this paper provides a case study (§6) demonstrating how the annotation taxonomy was applied to create an annotated dataset of archival documentation. For future NLP work, the resulting dataset of historical and contemporary language annotated for gender biases provides a corpus to analyze gender biased language for diachronic patterns, to analyze correlations between types of gender biases, and to develop gender bias classification models. Specific to the GLAM sector, gender bias classification models could enhance reparative description practices. A model's ability to automatically identify descriptions of heritage items that contain gender biases would enable efficient prioritization of the additions and revisions needed on outdated, harmful descriptions in GLAM documentation.

2 Bias Statement

This paper adopts our previously published definition of biased language (Havens et al., 2020), narrowing the focus to gender bias as written in §1. Gender biased language may cause representational or allocative harms to a person of any gender (Blodgett et al., 2020; Crawford, 2017). The taxonomy created in this paper considers a person's gender to be self-described and changeable, rather than being limited to the binary and static conceptualization of gender as either a man or woman since birth (Keyes, 2018; Scheuerman et al., 2020). Recognizing that a person's gender may be impossible to determine from the information available about them, the taxonomy also allows annotators to record uncertainty (Shopland, 2020). Furthermore, the paper acknowledges that characteristics other than gender, such as racialized ethnicity and economic class, influence experiences of power and oppression (Crenshaw, 1991). Drawing on archival science and feminist theories, the paper considers knowledge derived from language as situated in a particular perspective and, as a result, incomplete (Tanselle, 2002; Harding, 1995; Haraway, 1988).

To communicate this paper's perspective, we as authors report our identification as three women and one man; and our nationalities, as American, German, and Scots. Annotators identify as women (one specifying queer woman and two, cis women); they are of American, British, Hungarian, and Scots nationalities. Though annotators do not represent great gender diversity,[3] the annotation process still contributes to the advancement of gender equity.

As women, the annotators identify as a minoritized gender. The evolution of British English demonstrates the historical dominance of the perspective of the heteronormative man, and the pejoration of terms for women (Spencer, 2000; Schulz, 2000; Lakoff, 1989).[4] Creating a women-produced dataset challenges the dominant gender perspective by explicitly labeling where minoritized genders' perspectives are missing (D'Ignazio and Klein, 2020; Smith, 2006; Fairclough, 2003).

3 Related Work

Evidence of bias in ML data and models abound regarding gender (Kurita et al., 2019; Zhao et al., 2019), disability (Hutchinson et al., 2020), racial-

[2]This paper uses *gender diverse* in the sense the Trans Metadata Collective (2022) uses the term: to include gender expressions that do not fit within binary conceptualizations of gender, that differ from one's gender assigned at birth, and that cannot be described with the culturally-specific term *trans*.

[3]The availability of people who responded to the annotator application and the annotation timeline limited the gender diversity that could be achieved among annotators.

[4]In the 16^{th} century, grammarians instructed writers to write "men" or "man" before "women" or "woman." In the 18^{th} century, "man" and "he" began to be employed as universal terms, rather than "human" and "they" (Spencer, 2000).

ized ethnicities (Sap et al., 2019), politics and economics (Elejalde et al., 2017), and, for an intersectional approach (Crenshaw, 1991), a combination of characteristics (Jiang and Fellbaum, 2020; Sweeney and Najafian, 2019; Tan and Celis, 2019). Harms from such biases are also well documented (Birhane and Prabhu, 2021; Costanza-Chock and Philip, 2018; Noble, 2018; Vainapel et al., 2015; Sweeney, 2013). Despite numerous bias mitigation approaches put forth (Cao and Daumé III, 2020; Dinan et al., 2020a; Hube and Fetahu, 2019; Webster et al., 2018; Zhao et al., 2018), many have limited efficacy, failing to address the complexity of biased language (Stańczak and Augenstein, 2021; Blodgett et al., 2021; Gonen and Goldberg, 2019).

Methods of removing bias tend to be mathematically focused, such as Basta et al. (2020) and Borkan et al. (2019). As McCradden et al. (2020) state, typical ML bias mitigation approaches assume biases' harms can be mathematically represented, though no evidence of the relevance of proposed bias metrics to the real world exists. On the contrary, Goldfarb-Tarrant et al. (2021) found no correlation between a commonly used intrinsic bias metric, Word Embedding Association Test, and extrinsic metrics in the downstream tasks of coreference resolution and hate speech detection. Due to the misalignment between abstract representations of bias and the presence and impact of bias, this paper presents a taxonomy to measure biased language at its foundation: words.

Limitations to bias mitigation efforts also result from overly simplistic conceptualizations of bias (Devinney et al., 2022; Stańczak and Augenstein, 2021; Blodgett et al., 2020). NLP gender bias work, for example, often uses a binary gender framework either in its conceptualization (such as Webster et al. (2018)) or application (such as Dinan et al. (2020b)), and tends to focus on one variety of gender bias, stereotypes (Stańczak and Augenstein, 2021; Doughman et al., 2021; Bolukbasi et al., 2016). NLP bias work more generally often asserts a single ground truth (Davani et al., 2022; Sang and Stanton, 2022; Basile et al., 2021). Despite evidence that bias varies across domains (Basta et al., 2020), approaches to mitigating bias have yet to address the contextual nature of biased language, such as how it varies across time, location, and culture (Bjorkman, 2017; Bucholtz, 1999; Corbett, 1990). This paper adopts a data feminist (D'Ignazio and Klein, 2020) and perspectivist approach (Basile, 2022) to situate identification and measurement of bias in a particular context.

Data feminism views data as situated and partial, drawing on feminist theories' view of knowledge as particular to a time, place, and people (Harding, 1995; Crenshaw, 1991; Haraway, 1988). Similarly, the Perspectivist Data Manifesto encourages disaggregated publication of annotated data, recognizing that conflicting annotations may all be valid (Basile, 2022). Indigenous epistemologies, such as the Lakota's concept of *waȟkàŋ*, further the notion of the impossibility of a universal truth. Translated as "that which cannot be understood," waȟkàŋ communicates that knowledge may come from a place beyond what we can imagine (Lewis et al., 2018). Our taxonomy thus permits annotations to overlap and record uncertainty, and our aggregated dataset incorporates all annotators' perspectives.

Encouraging greater transparency in dataset creation, Bender et al. (2021) and Jo and Gebru (2020) caution against creating datasets too large to be adequately interrogated. Hutchinson et al. (2021), Mitchell et al. (2019), and Bender and Friedman (2018) propose new documentation methods to facilitate critical interrogation of data and the models trained on them. Our appendices include a data statement documenting the creation of the annotated data presented in this paper (§B). To maximize the transparency of our data documentation, we will publish the data only after further interrogation of its gender bias annotations, including collaborative analysis with the Centre for Research Collections.

4 Methodology

To practically apply theories and approaches from NLP, data feminism, and indigenous epistemologies, we apply the case study method, common to social science and design research. Case studies use a combination of data and information gathering approaches to study particular phenomena in context (Martin and Hanington, 2012), suitable for annotating gender biased language because gender and bias vary across time, location, and culture. Furthermore, case studies report and reflect upon outliers discovered in the research process (ibid.), supporting our effort to create space for the perspectives of people minoritized due to their gender identity. After first developing the annotation taxonomy through an interdisciplinary literature review and participatory action research with archivists

(§5), we applied the taxonomy in a case study to create datasets annotated for gender bias (§6).

Adopting our previously published bias-aware methodology (Havens et al., 2020), we employed participatory action research (Swantz, 2008; Reid and Frisby, 2008), collaborating with the institution that manages our data source: the Centre for Research Collections. Due to validity (Welty et al., 2019) and ethical concerns (Gleibs, 2017) with crowdsourcing, we hired annotators with expertise in archives (the domain area of the case study's data) and gender studies (the focus area of this paper's bias mitigation) to apply the taxonomy in a case study. Hiring a small number of annotators will enable us to publish disaggregated versions of the annotated data, implementing data perspectivism (Basile, 2022; Basile et al., 2021).

Following the approach of Smith (2006) to heritage, we consider heritage to be a process of engaging with the past, present, and future. Annotators in this paper's case study visited, interpreted, and negotiated with heritage (Smith, 2006) in the form of archival documentation. Annotating archival documentation with labels that mark specific text spans as gender biased transforms the documentation, challenging the "authorized heritage discourse" (ibid., 29) of the heteronormative man. We aim such explicit labeling to recontextualize the archival documentation, transforming its language by placing it in a new social context (Fairclough, 2003): the 21st century United Kingdom, with gender conceptualized as a self-defined, changeable identity characteristic. We aim this negotiation-through-annotation to guide the NLP models we will create with the data in the future towards more equitable representations of gender.

5 Annotation Taxonomy

Our annotation taxonomy organizes labels (lettered) into three categories (numbered). Category and label names are **bolded**. Each label's listing includes a definition and example. Examples are *italicized*; labeled text in each example is underlined. For every label, annotators could label a single word or multiple words. Examples come from the archival documentation summarized in §6 except for 1(a), *Non-binary*, and 3(d), *Empowering*, because annotators did not find text relevant to their definitions (the "Fonds ID," or collection identifier, indicates where in the documentation example descriptions may be found). §7 further explains the rationale for the taxonomy's labels, and how they facilitate analysis and measurement of gender biased language.

1. **Person Name:** the name of a person, including any pre-nominal titles (i.e., Professor, Mrs., Sir, Queen), when the person is the primary entity being described (rather than a location named after a person, for example)
   (a) **Non-binary:** the pronouns, titles, or roles of the named person are non-binary
   Example 1(a): Francis McDonald went to the University of Edinburgh where they studied law.
   (b) **Feminine:** the pronouns, titles, or roles of the named person are feminine
   Example 1(b): "Jewel took an active interest in her husband's work..." (Fonds ID: Coll-1036)
   (c) **Masculine:** the pronouns, titles, or roles of the named person are masculine
   Example 1(c): "Martin Luther, the man and his work." (Fonds ID: BAI)
   (d) **Unknown:** any pronouns, titles, or roles of the named person are gender neutral, or none are provided
   Example 1(d): "Testimonials and additional testimonials in favour of Niecks, candidacy for the Chair of Music, 1891." (Fonds ID: Coll-1086)

2. **Linguistic:** gender marked in the way a word or words reference a person or people, assigning them a specific gender that cannot be determined with certainty from the word(s)
   (a) **Generalization:** use of a gender-specific term (i.e., roles, titles) to refer to a group of people that could identify as more than the specified gender
   Example 2(a): "His classes included Anatomy, Practical Anatomy...Midwifery and Diseases of Women, Therapeutics, Neurology...Public Health, and Diseases of the Skin." (Fonds ID: Coll-1118)
   (b) **Gendered Role:** use of a word denoting a person's role that marks either a non-binary, feminine, or masculine gender
   Example 2(b): "New map of Scotland for Ladies Needlework, 1797" (Fonds ID: Coll-1111)

(c) **Gendered Pronoun:** marking a person or people's gender with gendered pronouns (i.e., she, he, ey, xe, or they as a singular pronoun)
Example 2(c): "He obtained surgical qualifications from Edinburgh University in 1873" (Fonds ID: Coll-1096)

3. **Contextual:** expectations about a gender or genders that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (i.e., sentence structure or word choice)

   (a) **Stereotype:** a word or words that communicate an expectation of a person or people's behaviors or preferences that does not reflect the extent of their possible behaviors or preferences; or that focus on a single aspect of a person that doesn't represent that person holistically
   Example 3(a): "The engraving depicts a walking figure (female) set against sunlight, and holding/releasing a bird." (Fonds ID: Coll-1116)

   (b) **Omission:** focusing on the presence, responsibility, or contribution of one gender in a situation where more than one gender has a presence, responsibility or contribution; or defining a person in terms of their relation to another person
   Example 3(b): "This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version in 1909 entitledApollinaire [sic] and his friends." (Fonds ID: Coll-1090).

   (c) **Occupation:** a word or words that refer to a person or people's job title for which the person or people received payment, excluding occupations in pre-nominal titles (for example, "Colonel Sir Thomas" should not have an occupation label)
   Example 3(c): "He became a surgeon with the Indian Medical Service." (Fonds ID: Coll-1096).

   (d) **Empowering:** reclaiming derogatory words as positive
   Example 3(d): a person describing themself as queer in a self-affirming manner

We chose to build on the gender bias taxonomy of Hitti et al. (2019) because the authors grounded their definitions of types of gender bias in gender studies and linguistics, and focused on identifying gender bias at the word level, aligning with our approach. Though Dinan et al. (2020b) also provide a framework for defining types of gender bias, their framework focuses on relationships between people in a conversation, identifying "bias when speaking ABOUT someone, bias when speaking TO someone, and bias from speaking AS someone" (316). The nature of our corpus makes these gender bias dimensions irrelevant to our work: GLAM documentation contains descriptions that only contain text written *about* a person or people (or other topics); it does not contain text that provides gender information about who is speaking or who is being spoken to. Additionally, despite writing of four gender values (unknown, neutral, feminine, and masculine), the dataset and classifiers of Dinan et al. (2020b) are limited to "*masculine* and *feminine* classes" (317). The authors also do not explain how they define "bias," limiting our ability to draw on their research.

Doughman et al. (2021) provide another gender bias taxonomy that builds on that of Hitti et al. (2019), resulting in overlaps between our taxonomies. However, Doughman et al. (2020) focus on gender stereotypes, while our taxonomy considers other types of gender biases. Though less explicit in the names of our taxonomy's labels, we also looked to the descriptions of gender and gender bias from Cao and Daumé III (2021), who point out the limited gender information available in language. The aim of our dataset creation differs from Cao and Daumé III (2021), though. They created data that represents trans and gender diverse identities in order to evaluate models' gender biases, specifically looking at where coreference resolution fails on trans and non-binary referents. By contrast, we aim to create a dataset that documents biased representations of gender, with the future aim of creating models that are able to identify types of gender bias in language.

6 Case Study

To demonstrate the application of the taxonomy, we present a case study situated in the United Kingdom in the 21st century, annotating archival documentation written in British English from the Centre for Research Collections at the University of Edinburgh (CRC Archives). This paper thus takes the first step in building a collection of case studies that situate NLP bias research in a specific context.

	Title	Biographical/Historical	Scope & Contents	Processing Information	Total
Count	4,834	576	6198	280	11,888
Words	51,904	75,032	269,892	3,129	399,957
Sentences	5,932	3,829	14,412	301	24,474

Table 1: Total counts, words and sentences for descriptive metadata fields in the aggregated dataset. Calculations were made using Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002).

A collection of case studies would enable the NLP community to determine which aspects of bias mitigation approaches generalize across time, location, culture, people, and identity characteristics.

The CRC's Archives' documentation served as a suitable data source because the documentation adheres to an international standard for organizing archival metadata (ISAD(G) (ICA, 2011)), the archivists at the institution had found gender bias in the documentation's language, and the archivists were already engaged in efforts to mitigate gender bias in the archival documentation. The documentation describes a variety of heritage collections and items, such as letters, journals, photographs, degree certificates, and drawings; on a variety of topics, such as religion, research, teaching, architecture, and town planning. Employees at the partner institution describe themselves as activists changing archival practices to more accurately represent the diverse groups of people that the archival collections are intended to serve.

The annotation corpus consists of 24,474 sentences and 399,957 words, selected from the first 20% of the entire corpus of archival documentation from the partner institution's catalog (see §B.9 for more on this corpus). Table 1 provides a breakdown of the size of the annotation corpus by metadata field. 90% of the annotation corpus (circa 22,027 sentences and 359,961 words) was doubly annotated with all labels, and 10% of the annotation corpus (circa 2,447 sentences and 39,996 words) was triply annotated with all labels. In total, the annotation process amounted to circa 400 hours of work and £5,333.76, funded by a variety of internal institutional funds. Each of the four hired annotators worked for 72 hours over eight weeks at £18.52 per hour (minimum wage is £9.50 per hour (Gov.uk, 2022)). The hired annotators were PhD students selected for their experience in gender studies or archives, with three of the annotators having experience in both. The lead annotator worked for 86 hours over 16 weeks as part of their PhD research.

The categories of labels in the annotation taxonomy were divided among annotators according to the textual relations the labels record. Hired annotators 1 and 2 (A1 and A2) labeled internal relations of the text with *Person Name* and *Linguistic* categories, hired annotators 3 and 4 (A3 and A4) labeled external relations of the text with the *Contextual* category, and the lead annotator (A0) labeled both relations with all categories. A1 and A3 labeled the same subset of archival documentation, and A2 and A4 labeled the same subset of archival documentation, ensuring every description had labels from all categories. The lead annotator labeled the same descriptions as A1 and A3, and a subset of the descriptions that A2 and A4 labeled (due to time constraints, A0 could not label all the same descriptions). Prior to beginning annotation, *Gendered Pronoun*, *Gendered Role*, and *Occupation* labels were automatically applied. The annotators corrected mistakes from this automated process during their manual annotation.

We produced three instances of the annotation corpus: one for A0, one for each pair of hired annotators (A1 and A3, and A2 and A4), and one aggregated dataset. The aggregated dataset combines annotations from all five annotators, totaling 76,543 annotations with duplicates and 55,260 annotations after deduplication. Manual reviews of each annotator's dataset informed the aggregation approach, which involved a combination of programmatic and manual steps. The data statement in §B details the aggregation approach. Figure 1 displays the number of annotations in the aggregated dataset by label (§A contains additional annotation figures). In line with perspectivist NLP (Basile, 2022), the individual annotator's datasets will be published alongside the aggregated dataset, enabling researchers to interrogate patterns of agreement and disagreement, and enabling future work to compare the performance of classifiers trained on disaggregated and aggregated datasets.

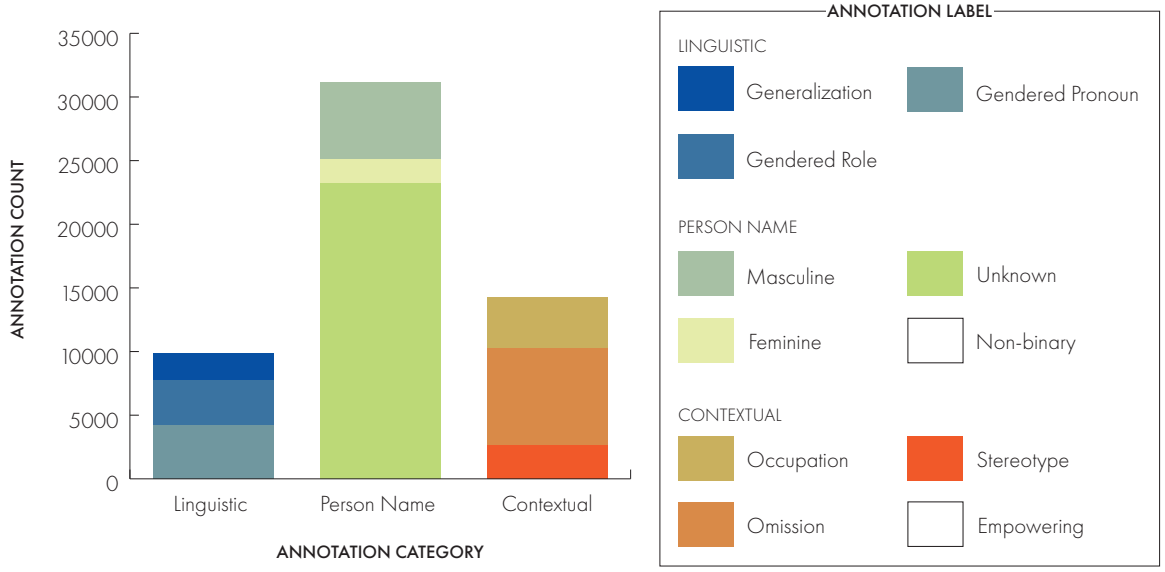

Figure 1: Total Annotations Per Label in the Aggregated Dataset. The stacked bar chart groups annotation labels into bars by category. Across all three categories, there are 55,260 annotations in the aggregated dataset. *Non-binary* (a *Person Name* label) and *Empowering* (a *Contextual* label) both have a count of zero.

6.1 Inter-Annotator Agreement

Due to our aim to create a training dataset for document classification models, identifying strictly matching text spans that annotators labeled was deemed less important than the presence of a label in a description. Consequently, inter-annotator agreement (IAA) calculations consider annotations with the same label to agree if their text spans match or overlap. Figures 2 and 3 display the F_1 scores for each label, with the aggregated dataset's labels as predicted and the annotators' labels as expected. Tables 2 and 3 in the appendices list true and false positives, false negatives, precision, and recall, in addition to F_1 scores, for IAA among the annotators and with the aggregated dataset.

IAA calculations reflect the subjectivity of gender bias in language. F_1 scores for the gendered language labels *Gendered Role* and *Gendered Pronoun* fall between 0.71 and 0.99. F_1 scores for annotating gender biased language are relatively low, with the greatest agreement on the *Generalization* label at only 0.56, on the *Omission* label at 0.48, and on the *Stereotype* label at 0.57. For *Person Name* labels, A0 and A2 agree more than A1: A0 and A2's F_1 scores for all *Person Name* labels are between 0.82 and 0.86, while A1's scores with either A0 or A2 are between 0.42 and 0.64. A1 has a particularly high false negative rate for the *Unknown* label compared to A0.

After creating the aggregated dataset, we calculated IAA between each annotator and the aggregated dataset. F_1 scores for all *Person Name* and *Linguistic* labels except *Generalization* are similarly high (0.74 to 0.98). *Generalization* proved particularly difficult to label. Annotators used *Generalization* and *Gendered Role* inconsistently. As a result, during the aggregation process, we revised the definition of *Generalization* to more clearly distinguish it from *Gendered Role*. Consequently the IAA between annotators and the aggregated dataset for this label is particularly low (0.1 to 0.4).

For *Contextual* labels, F_1 scores with the aggregated dataset as "expected" and an annotator as "predicted" increased more dramatically than the *Person Name* and *Linguistic* labels' F_1 scores. Besides *Omission* with A3, all F_1 scores are between 0.76 and 0.91. For *Stereotype*, A3 agreed more strongly with the aggregated dataset than A0 and A4. The reverse is true for *Omission* and *Occupation*, with A0 and A4 agreeing more strongly with the aggregated dataset than A3. A3's notes explain that she did not annotate an incomplete version of a person's name as an omission if the complete version was provided elsewhere in the collection's descriptions, whereas A0 and A4 annotated incomplete versions of people's names as omission unless the complete version appeared in the same description.

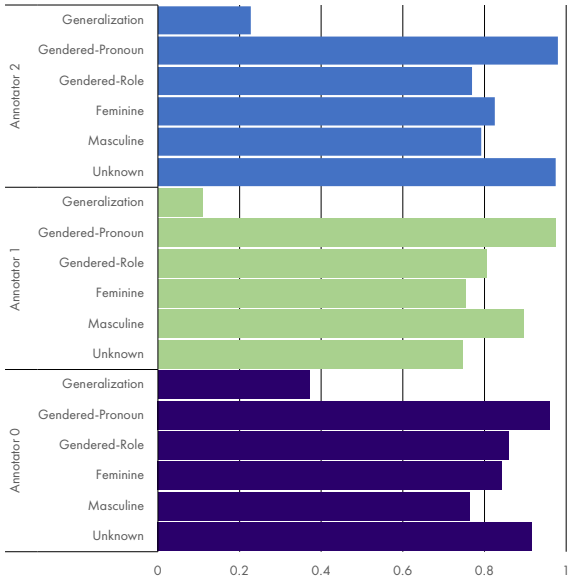

Figure 2: Linguistic and Person Name labels' F_1 scores (on the X axis) with the aggregated dataset's labels as the expected labels and each annotator's labels as predicted labels. Annotators (on the Y axis) did not use the *Non-binary* label (from the *Person Name* category) so it does not appear in the aggregated dataset.

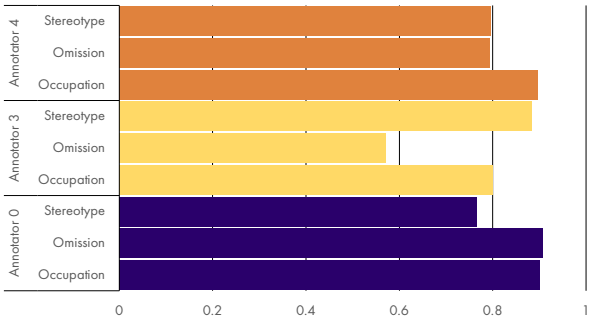

Figure 3: Contextual labels' F_1 scores (on the X axis), with the aggregated dataset's labels as the expected labels and each annotator's labels as predicted labels. Annotators (on the Y axis) did not use the *Empowering* label as defined in the annotation instructions, so it does not appear in the aggregated dataset.

Two labels were not applied according to the taxonomy's definitions: *Empowering* and *Non-binary*. *Empowering* was used by A3 according to a different definition than that of the taxonomy (see §B). As only 80 instances of the label exist in A3's dataset, though, there are likely to be insufficient examples for effectively training classifiers on this label in future work.

The annotators did not use the *Non-binary* label. That being said, this does not mean there were not people who would identify as non-binary represented in the text of the annotation corpus. Additional linguistic and historical research may identify people who were likely to identify as non-binary in the corpus of archival documentation, as well as more specific gender identities for people whose names were annotated as *Masculine* or *Feminine*. Metadata entries for people in the partner institution's catalog may also provide more information relevant to gender identities. Shopland (2020) finds that focusing on actions that people were described doing can help to locate people of minoritized genders (and sexualities) in historical texts. However, Shopland also cautions researchers against assuming too much: a full understanding of a person's gender often remains unattainable from the documentation that exists about them.

As Figure 1 displays, *Unknown* is the most prevalent label in the *Person Name* category, because each annotation of a person's name was informed by words within the description in which that name appears. Consequently, for people named in more than one description, there may be different person name labels applied to their name across those descriptions. The rationale for this approach comes from the aim to train document classification models on the annotated data where each description serves as a document. Should a person change their gender during their lifetime, and archival documentation exists that describes them as different genders, the person may wish a model to use the most recent description of a person to determine their gender, or not use any gender information about the person, in case obviating their change of gender leads to safety concerns (Dunsire, 2018). Furthermore, many GLAM content management systems do not have versioning control, so dates of descriptions may not exist to determine the most recent description of a person's gender. *Person Name* labels are thus based on the description in which a name appears to minimize the risk of misgendering (Scheuerman et al., 2020).

7 Discussion and Limitations

The paper's annotation taxonomy builds on biased language research from NLP, information sciences, gender studies, and linguistics literature. The gender bias taxonomy of Hitti et al. (2019), which categorizes gender biases based on whether the bias comes from the sentence structure or the context (i.e. people, relationships, time period, location) of the language, served as a foundation. We adopted

four labels from that taxonomy: *Gendered Pronoun*, *Gendered Role*, *Generalization*, and *Stereotype* (merging Hitti et al.'s Societal Stereotype and Behavioral Stereotype categories). Drawing on archival science and critical discourse analysis, and guided by participatory action research with archivists (e.g., interviews, workshops), we added to and restructured Hitti et al.'s taxonomy. The *Person Name* labels were added so that the representation of people of different genders in the archival documentation could be estimated. Annotators chose which label to apply to a person's name based on gendered pronouns or roles that refer to that person in the description in which their name appears. For example, "they" as singular for *Non-binary*, "his" for *Masculine*, and "she" for *Feminine*; or "Mx." for *Non-binary*, "Lady" for *Feminine*, or "son" for *Masculine*. The *Unknown*, *Feminine*, and *Masculine* labels distinguish our approach from previous NLP gender bias work that has not allowed for uncertainty.

Guessing a person's gender risks misgendering (Scheuerman et al., 2020), a representational harm (Blodgett et al., 2020; Crawford, 2017), and fails to acknowledge that sufficient information often is not available to determine a person's gender with certainty (Shopland, 2020). This led us to replace the initial labels of *Woman* and *Man* with *Feminine* and *Masculine*, recognizing that pronouns and roles are insufficient for determining how people define their gender. Each *Person Name* label encompasses multiple genders. For instance, a person who identifies as a transwoman, as genderfluid, or as a cis woman may use feminine pronouns, such as "she," or feminine roles, such as "wife." Though we aimed to create a taxonomy inclusive of all genders, we acknowledge this may not have been achieved, and welcome feedback on how to represent any genders inadvertently excluded.

We also added three labels to the *Contextual* category: *Occupation*, *Omission*, and *Empowering*. *Occupation* was added because, when combined with historical employment statistics, *Occupation*-labeled text spans could inform estimates of the representation of particular genders within the collaborating archive's collections. Furthermore, *Person Name* annotations combined with their occupations could guide researchers to material beyond the archive that may provide information about those people's gender identity. *Omission* was added because, during group interviews, representatives from the collaborating archive described finding gender bias through the lack of information provided about women relative to the detail provided about men. *Empowering* was added to account for how communities reclaim certain derogatory terms, such as "queer," in a positive, self-affirming manner (Bucholtz, 1999).

Figure 1 displays how prevalent *Omission* was in the annotated data: this label is the most commonly applied label from the *Contextual* category. Such prevalence demonstrates the value of interdisciplinary collaboration and stakeholder engagement, carried out in our participatory action research with domain experts. Had archivists at the partner institution not been consulted, we would not have known how relevant omitted information regarding gender identities would be to identifying and measuring gender bias in archival documentation.

The final annotation taxonomy includes labels for gendered language (specifically, *Gendered Role*, *Gendered Pronoun*, and all labels in the *Person Name* category), rather than only explicitly gender biased language (specifically, *Generalization*, *Stereotype*, and *Omission*), because measuring the use of gendered words across an entire archives' collection provides information about gender bias at the overall collections' level. For example, using a gendered pronoun such as "he" is not inherently biased, but if the use of this masculine gendered pronoun far outnumbers the use of other gendered pronouns in our dataset, we can observe that the masculine is over-represented, indicating a masculine bias in the archives' collections overall. Labeling gender-biased language focuses on the individual description level. For example, the stereotype of a wife playing a supporting role to her husband comes through in this description:
Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more.

Instructions for applying the taxonomy permitted labels to overlap as each annotator saw fit, and asked annotators to annotate from their contemporary perspective. Approaching the archival metadata descriptions as discourse (meaning language as representations of the material, mental, and social worlds (Fairclough, 2003)), the taxonomy of labels represents the "internal relations" and "external relations" of the descriptions (ibid., 37). The *Person Name* and *Linguistic* categories annotate in-

ternal relations, meaning the "vocabulary (or 'lexical') relations" (ibid., 37) of the descriptions. To apply their labels, annotators looked for the presence of particular words and phrases (i.e., gendered pronouns, gendered titles, familial roles).

The *Contextual* category annotates external relations: relations with "social events ... social practices and social structures" (Fairclough, 2003, 36). To apply *Contextual* labels, annotators reflected on the production and reception of the language in the archival documentation. For instance, to apply the *Stereotype* label, annotators considered the relationship between a description's language with social hierarchies in 21st century British society, determining whether the term or phase adequately represented the possible gender diversity of people being described.

8 Conclusion and Future Work

This paper has presented a taxonomy of gender biased language with a case study to support clarity and alignment in NLP gender bias research. Recognizing the value of clearly defined metrics for advancing bias mitigation, the taxonomy provides a structure for identifying types of gender biased language at the level they originate (words and phrases), rather than at a level of abstraction (i.e., vector spaces). Still, the case study presented in this paper demonstrates the difficulty of determining people's gender with certainty. While recognizing the value of NLP systems for mitigating harms from gender biased language at large scale, we contend that conceptualizations of gender must extend to trans and gender diverse gender expressions if NLP systems are to empower minoritized gender communities.

Future work will include the publication of the case study's datasets, analysis of the datasets, and document classification models trained on the datasets. The datasets will include each individual annotator's dataset and two aggregated datasets, one with duplicates across different annotators, and one deduplicated to exclude matching and overlapping annotations from different annotators. The analysis of the datasets and creation of models trained on them will be informed by participatory action research, incorporating perspectives from archivists, and from people of trans and gender diverse identities not represented in the research team. The dataset will be published in the same location as the code written to create the corpus of archival documentation and the annotated datasets.[5] The taxonomy and forthcoming datasets aim to guide NLP systems towards measurable and inclusive conceptualizations of gender.

Acknowledgements

Thank you to our collaborators, Rachel Hosker and her team at the Centre for Research Collections; our annotators, Suzanne Black, Ashlyn Cudney, Anna Kuslits, and Iona Walker; and Richard Tobin, who wrote the pre-annotation scripts for this paper's annotation process. We also extend our gratitude to the organizations who provided grants to support the research reported in this paper: the University of Edinburgh's Edinburgh Futures Institute, Centre for Data, Culture &Society, Institute for Language, Cognition and Computation, and School of Informatics; and the UK's Engineering and Physical Sciences Research Council. Additional thanks go to the organizers of the Fourth Workshop on Gender Bias in Natural Language Processing, for the opportunity to submit this paper, and to the reviewers who gave feedback on this paper.

References

Valerio Basile. 2022. The Perspectivist Data Manifesto. [Online; accessed March 21, 2022].

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *CoRR*, abs/2109.04270.

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2020. Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings. *Neural Computing & Applications*, 33(8):3371–3384.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, 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*, FAccT '21, page 610–623, New York, USA. Association for Computing Machinery.

Ruha Benjamin. 2019. *Race after technology : abolitionist tools for the new Jim code*. Polity, Cambridge, UK.

[5] github.com/thegoose20/annot

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Bronwyn M Bjorkman. 2017. Singular They and the Syntactic Representation of Gender in English. *Glossa (London)*, 2(1):1.

Su Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. 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. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, USA. Association for Computing Machinery.

Mary Bucholtz. 1999. Gender. *Journal of linguistic anthropology*, 9(1-2):80–83.

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

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Terry Cook. 2011. 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists*, 32(2):173–189.

M.Z. Corbett. 1990. Clearing the air: some thoughts on gender-neutral writing. *IEEE Transactions on Professional Communication*, 33(1):2–6.

Sasha Costanza-Chock and Nick Philip. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*.

Kate Crawford. 2017. The Trouble with Bias. In *Neural Information Processing Systems Conference Keynote*. [Online; accessed 10-July-2020].

CRC. 2018. Collection: Papers and artwork of Yolanda Sonnabend relating to her collaboration with C.H. Waddington. [Online; accessed 19 May 2022.

Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. *Computing Research Repository*.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. Strong ideas series. MIT Press, Cambridge, USA.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Jad Doughman, Fatima Abu Salem, and Shady Elbassuoni. 2020. Time-aware word embeddings for three Lebanese news archives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4717–4725, Marseille, France. European Language Resources Association.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. 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. Association for Computational Linguistics.

Gordon Dunsire. 2018. Ethical issues in catalogue content standards. In *Catalogue & Index*, volume 191, pages 11–15.

Erick Elejalde, Leo Ferres, and Eelco Herder. 2017. The Nature of Real and Perceived Bias in Chilean Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 95–104, New York, USA. Association for Computing Machinery.

Norman Fairclough. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Routledge, London, UK.

Noah Geraci. 2019. Programmatic approaches to bias in descriptive metadata. In *Code4Lib Conference 2019*. [Online; accessed 28-May-2020].

Ilka H. Gleibs. 2017. Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market addresss. *Behavior Research Methods*, 49(4):1333–1342.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. 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. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL 2019*, arXiv:1903.03862v2.

Gov.uk. 2022. National Minimum Wage and National Living Wage rates.

Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575.

Sandra Harding. 1995. "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3).

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2020. Situated data, situated systems: A methodology to engage with power relations in natural language processing research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 107–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. 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, IT. Association for Computational Linguistics.

Christoph Hube and Besnik Fetahu. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 195–203, Melbourne, AU. ACM.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, USA. Association for Computing Machinery.

ICA. 2011. ISAD(G): General International Standard Archival Description - Second edition.

May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Eun Seo Jo 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*, FAT* '20, page 306–316, New York, USA. Association for Computing Machinery.

Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW).

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. *CoRR*, abs/1906.07337.

Robin Lakoff. 1989. *Language and Woman's Place*. Harper & Row, New York, USA.

Susan Leavy. 2018. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, GE '18, page 14–16, New York, USA. Association for Computing Machinery.

Jason Edward Lewis, Nick Philip, Noelani Arista, Archer Pechawis, and Suzanne Kite. 2018. Making Kin with the Machines. *Journal of Design and Science*.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, USA. Association for Computational Linguistics.

Bella Martin and Bruce Hanington. 2012. 11 Case studies. In *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*, Beverly, USA. Rockport Publishers.

Melissa McCradden, Mjaye Mazwi, Shalmali Joshi, and James A. Anderson. 2020. *When Your Only Tool Is A Hammer: Ethical Limitations of Algorithmic Fairness Solutions in Healthcare Machine Learning*, page 109. Association for Computing Machinery, New York, USA.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT∗'19*.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, USA.

Helen O'Neill, Anne Welsh, David A Smith, Glenn Roe, and Melissa Terras. 2021. Text mining Mill: Computationally detecting influence in the writings of John Stuart Mill from library records. *Digital Scholarship in the Humanities*, 36(4):1013–1029.

Thomas Padilla. 2017. On a Collections as Data Imperative. *UC Santa Barbara Previously Published Works*.

Thomas Padilla. 2019. Responsible Operations: Data Science, Machine Learning, and AI in Libraries. *OCLC Research*.

Caroline Criado Perez. 2019. *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, London, UK.

Colleen Reid and Wendy Frisby. 2008. 6 Continuing the Journey: Articulating Dimensions of Feminist Participatory Action Research (FPAR). In *The SAGE Handbook of Action Research*, pages 93–105. SAGE Publications Ltd.

Yisi Sang and Jeffrey Stanton. 2022. The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, Lecture Notes in Computer Science, pages 425–444. Springer International Publishing, Cham.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. 2020. HCI Guidelines for Gender Equity and Inclusion: Misgendering.

Muriel R. Schulz. 2000. The Semantic Derogation of Women. In Lucy Burke, Tony Crowley, and Alan Girvin, editors, *The Routledge language and cultural theory reader*. Routledge, London, UK.

Norena Shopland. 2020. *A Practical Guide to Searching LGBTQIA Historical Records*. Taylor & Francis Group, Milton.

Laurajane Smith. 2006. *Uses of Heritage*. Routledge, London, UK.

Dale Spencer. 2000. Language and reality: Who made the world? (1980). In Lucy Burke, Tony Crowley, and Alan Girvin, editors, *The Routledge language and cultural theory reader*. Routledge, London, UK.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Sophia Ananiadou Tomoko Ohta, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*. Association for Computational Linguistics.

Marja Liisa Swantz. 2008. 2 Participatory Action Research as Practice. In *The SAGE Handbook of Action Research*, pages 31–48. SAGE Publications Ltd.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. *CoRR*, abs/1911.01485.

G. Thomas Tanselle. 2002. The World as Archive. *Common Knowledge*, 8(2):402–406.

Trans Metadata Collective. 2022. A Mandate for Trans and Gender Diverse Metadata (draft; working title).

Sigal Vainapel, Opher Y. Shamir, Yulie Tenenbaum, and Gadi Gilam. 2015. The dark side of gendered language: The masculine-generic form as a cause for self-report bias. *Psychological Assessment*, 27(4):1513–1519.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Computing Research Repository*, arXiv:1810.05201.

Anne Welsh. 2016. The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly*, 54(5–6):317–337.

Chris Welty, Praveen Paritosh, and Lora Aroyo. 2019. Metrology for AI: From Benchmarks to Instruments. *CoRR*, abs/1911.01875.

Elizabeth Yale. 2015. The History of Archives: The State of the Discipline. *Book History*, 18(1):332–359.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 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. Association for Computational Linguistics.

A Additional Tables and Figures

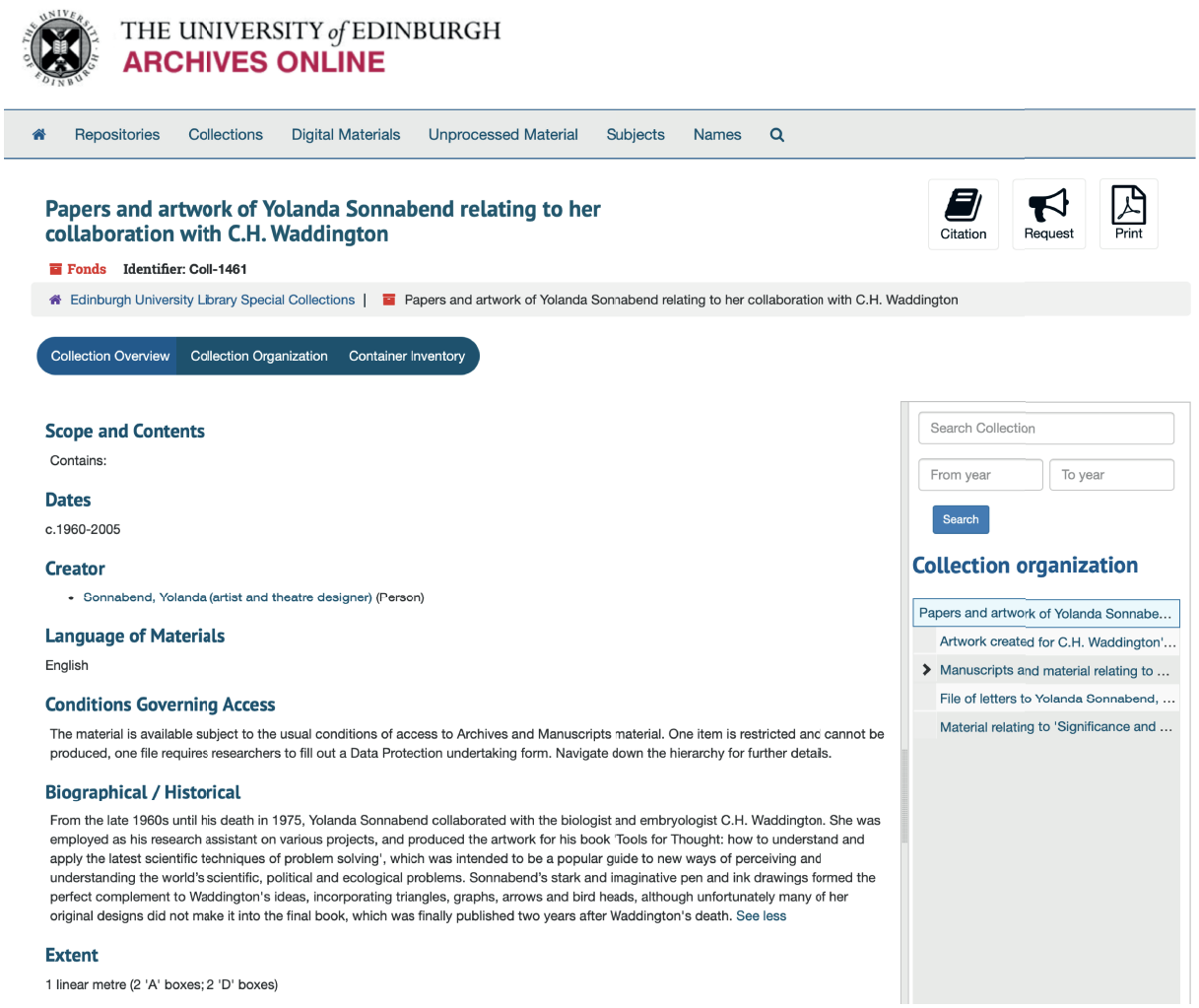

Figure 4: An example of GLAM documentation from the archival catalog of the Centre for Research Collections at the University of Edinburgh (2018). Metadata field names bolded in blue and their descriptions, regular, black text. The 'Title' field, however, is bolded in blue at the top of the page ("Papers and artwork of...").

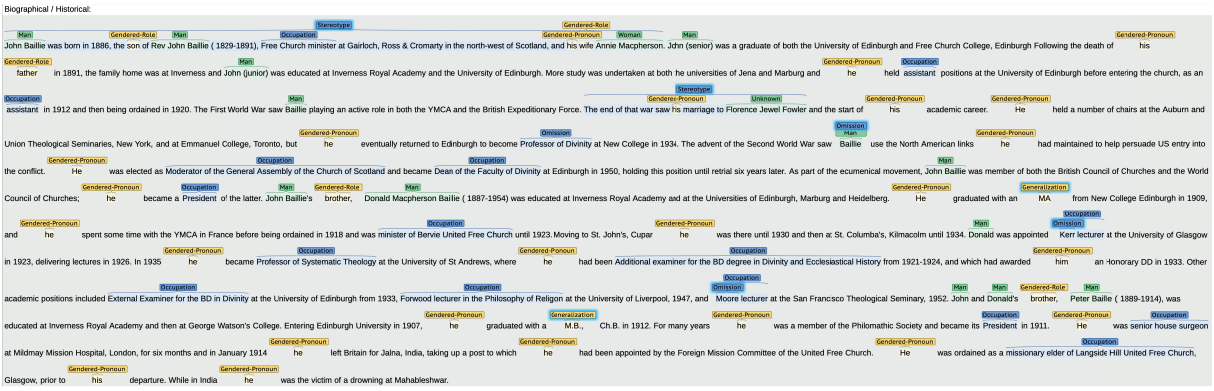

Figure 5: An example of a "Biographical / Historical" metadata field's description annotated with all labels from the taxonomy in the online annotation platform brat (Stenetorp et al., 2012).

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
0	1	Unknown	5031	1524	4268	0.76751	0.54103	0.63467	584
0	2	Unknown	2776	537	432	0.83791	0.86534	0.85140	170
1	2	Unknown	1048	1421	315	0.42446	0.76889	0.54697	72
0	1	Masculine	2367	2372	1079	0.49947	0.68688	0.57838	584
0	2	Masculine	728	111	146	0.86770	0.83295	0.84997	170
1	2	Masculine	380	169	411	0.69217	0.48040	0.56716	72
0	1	Feminine	627	427	642	0.59488	0.49409	0.53982	584
0	2	Feminine	724	128	178	0.84977	0.80266	0.82554	170
1	2	Feminine	287	496	279	0.36654	0.50707	0.42550	72
0	1	Non-binary	0	0	0	-	-	-	584
0	2	Non-binary	0	0	0	-	-	-	170
1	2	Non-binary	0	0	0	-	-	-	72
0	1	Gendered Role	1802	306	882	0.85484	0.67139	0.75209	584
0	2	Gendered Role	1404	162	257	0.89655	0.84527	0.87016	170
1	2	Gendered Role	438	292	52	0.60000	0.89388	0.71803	72
0	1	Gendered Pronoun	3398	101	190	0.97113	0.94705	0.95894	584
0	2	Gendered Pronoun	869	70	60	0.92545	0.93541	0.93041	170
1	2	Gendered Pronoun	518	7	11	0.98667	0.97921	0.98292	72
0	1	Generalization	37	35	262	0.51389	0.12375	0.19946	584
0	2	Generalization	74	51	63	0.59200	0.54015	0.56489	170
1	2	Generalization	2	50	7	0.03846	0.22222	0.06557	72

Table 2: Inter-annotator agreement measures for annotators who used the *Person Name* and *Linguistic* categories of labels to annotate archival documentation. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation "pos" is for "positive;" "neg," for "negative." The last column lists the number of files with annotations by both annotators for that row. No annotators applied the "Non-binary" label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
0	3	Occupation	1988	613	724	0.76432	0.73303	0.74835	485
0	4	Occupation	738	396	240	0.65079	0.75460	0.69886	149
3	4	Occupation	422	327	134	0.56341	0.75899	0.64674	57
0	3	Omission	1376	914	3259	0.60087	0.29687	0.39740	485
0	4	Omission	416	317	875	0.56753	0.32223	0.41106	149
3	4	Omission	215	315	155	0.40566	0.58108	0.47777	57
0	3	Stereotype	505	539	227	0.48371	0.68989	0.56869	485
0	4	Stereotype	507	525	600	0.49127	0.45799	0.47405	149
3	4	Stereotype	34	60	161	0.36170	0.17435	0.23529	57
0	3	Empowering	0	80	0	-	-	-	485
0	4	Empowering	0	0	0	-	-	-	149
3	4	Empowering	0	0	80	-	-	-	57

Table 3: Inter-annotator agreement measures for annotators who used the *Contextual* category of labels to annotate archival metadata descriptions. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation "pos" is for "positive;" "neg," for "negative." The last column lists the number of files with annotations by both annotators for that row. Only annotator 3 applied the "Empowering" label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
Agg	0	Unknown	10561	36	1900	0.99660	0.84752	0.91604	714
Agg	1	Unknown	6608	0	4511	1.00000	0.59430	0.74553	597
Agg	2	Unknown	15140	117	679	0.99233	0.95708	0.97439	444
Agg	0	Masculine	3963	18	2446	0.99548	0.61835	0.76285	714
Agg	1	Masculine	4749	1	1099	0.99979	0.81207	0.89621	597
Agg	2	Masculine	1007	5	525	0.99506	0.65731	0.79167	444
Agg	0	Feminine	1454	19	523	0.98710	0.73546	0.84290	714
Agg	1	Feminine	1076	0	707	1.00000	0.60348	0.75271	597
Agg	2	Feminine	994	12	410	0.98807	0.70798	0.82490	444
Agg	0	Nonbinary	0	0	0	-	-	-	714
Agg	1	Nonbinary	0	0	0	-	-	-	597
Agg	2	Nonbinary	0	0	0	-	-	-	444
Agg	0	Gendered-Role	3108	697	330	0.81682	0.90401	0.85821	714
Agg	1	Gendered-Role	1924	218	716	0.89823	0.72879	0.80468	597
Agg	2	Gendered-Role	1471	652	230	0.69289	0.86479	0.76935	444
Agg	0	Gendered-Pronoun	3933	160	165	0.96091	0.95974	0.96032	714
Agg	1	Gendered-Pronoun	3498	3	190	0.99914	0.94848	0.97315	597
Agg	2	Gendered-Pronoun	1016	1	41	0.99902	0.96121	0.97975	444
Agg	0	Generalization	405	1	1370	0.99754	0.22817	0.37139	714
Agg	1	Generalization	69	4	1123	0.94521	0.05789	0.10909	597
Agg	2	Generalization	127	0	862	1.00000	0.12841	0.22760	444

Table 4: Inter-annotator agreement between the aggregated dataset and annotators for the *Person Name* and *Linguistic* categories of labels to annotate archival documentation. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation "pos" is for "positive;" "neg," for "negative." The last column lists the number of files with annotations by both annotators for that row. No annotators applied the "Non-binary" label.

exp	pred	label	true pos	false pos	false neg	precision	recall	F_1	files
Agg	0	Occupation	2725	23	571	0.99163	0.82676	0.90172	631
Agg	3	Occupation	2320	290	873	0.88889	0.72659	0.79959	508
Agg	4	Occupation	1746	147	253	0.92235	0.87344	0.89723	450
Agg	0	Omission	5916	12	1187	0.99798	0.83289	0.90799	631
Agg	3	Omission	2310	13	3475	0.99440	0.39931	0.56981	508
Agg	4	Omission	1876	5	967	0.99734	0.65987	0.79424	450
Agg	0	Stereotype	1748	11	1058	0.99375	0.62295	0.76583	631
Agg	3	Stereotype	1089	9	279	0.99180	0.79605	0.88321	508
Agg	4	Stereotype	1400	2	715	0.99857	0.66194	0.79613	450
Agg	0	Empowering	0	0	0	-	-	-	631
Agg	3	Empowering	0	80	0	0.0	-	0.0	508
Agg	4	Empowering	0	0	0	-	-	-	450

Table 5: Inter-annotator agreement between the aggregated dataset and annotators for the *Contextual* category of labels to annotate archival metadata descriptions. The first two columns note the annotator whose labels were considered expected or predicted, respectively. The abbreviation "pos" is for "positive;" "neg," for "negative." The last column lists the number of files with annotations by both annotators for that row. Only annotator 3 applied the "Empowering" label.

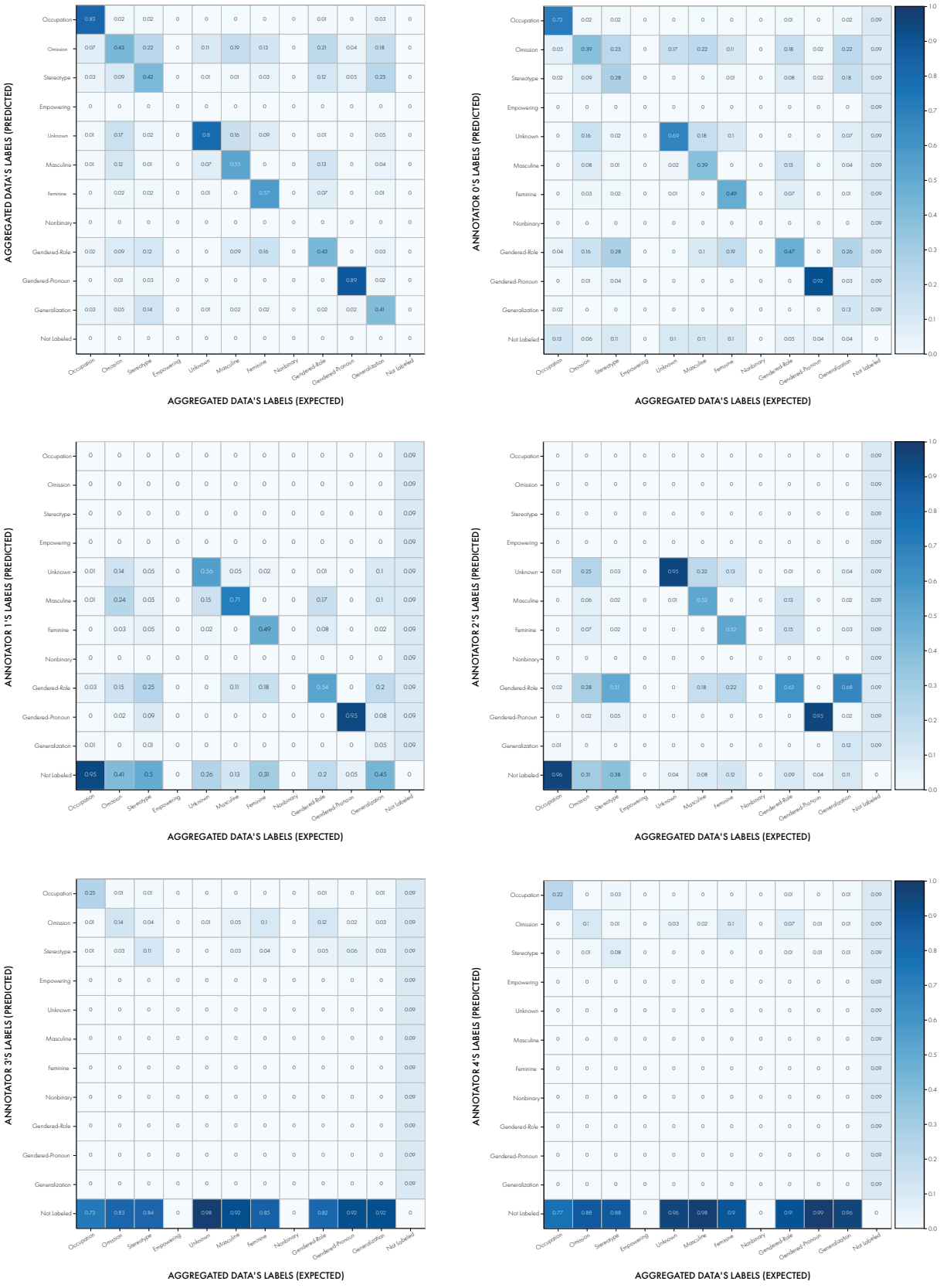

Figure 6: Confusion matrices normalized with a weighted average on the aggregated data's labels, so that class imbalances are taken into account. The top left confusion matrix displays intersections between the aggregated datasets labels, illustrating where the same text spans have more than one label. The remaining confusion matrices to display the agreement between an annotator's labels (Y axis) and the aggregated data's labels (X axis). The Y axis scale is the same for all matrices, ranging from zero to one.

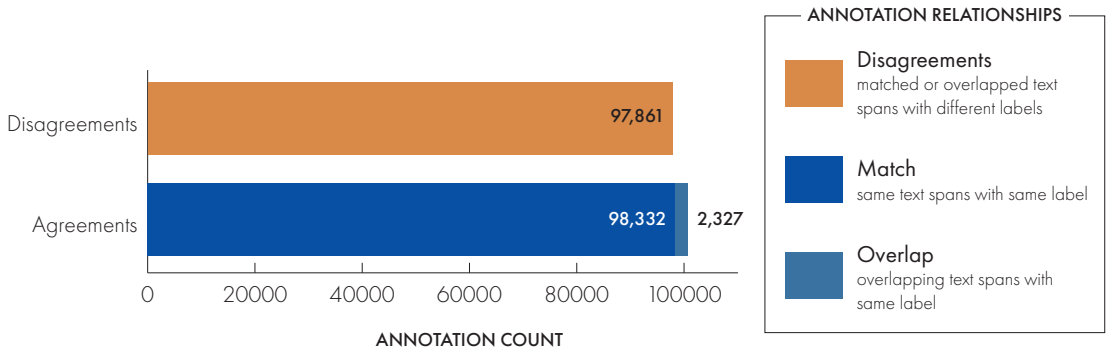

Figure 7: Disagreeing and Agreeing Label Counts Across All Annotators' Datasets. The bar chart displays counts of the occurrence of disagreements and agreements across annotators' labels. Annotations by two annotators with the same or overlapping text span but different labels are considered to be in disagreement. Annotations by two annotators with the same or overlapping text span and the same labels are considered to be in agreement. Agreements with the same text span are considered to be exact matches. Agreements with different but overlapping text spans are considered to be overlaps. Combined, the annotated datasets contain 198,520 annotations.

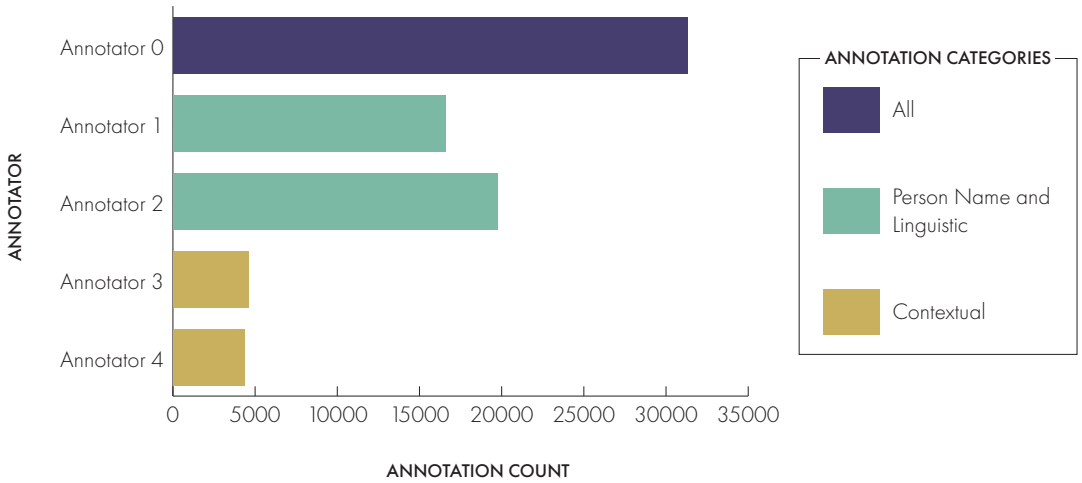

Figure 8: Total Annotations Per Annotator in the Aggregated Dataset. The bar chart displays the total annotations from each annotator included in the aggregated dataset, with colors indicating the category of labels each annotator used. For annotations that matched or overlapped, only one was added to the aggregated dataset, so the total number of annotations in the aggregated dataset (55,260) is 21,283 less than the sum of the annotators' annotations in this chart (76,543).

B Data Statement: Annotated Datasets of Archival Documentation

B.1 Curation Rationale

These datasets were created from a corpus of 1,460 files of archival metadata descriptions totaling circa 15,419 sentences and 255,943 words. That corpus is the first 20% of text from the corpus described in the Provenance Appendix (§B.9), annotated for gender bias according the the taxonomy in Other (§B.8). 73 of files (10% of the text) were triply annotated; the remaining 1,387 files (90% of the text) were doubly annotated. There are six instances of the annotated corpus: one for each of the five annotators and one that aggregates all annotators' labels. Participatory action research with archivists led the project to choose four metadata fields were chosen in the archival catalog to extract for annotation: Title, Scope and Contents, Biographical / Historical, and Processing Information.

The five annotated datasets were merged into a single aggregated dataset for classifier training and evaluation, so comparisons could be made on classifiers' performances after training on an individual annotator's dataset versus on the aggregated dataset. The merging process began with a one-hour manual review of each annotator's labels to identify patterns and common mistakes in their labeling, which informed the subsequent steps for merging the five annotated datasets.

The second step of the merging process was to manually review disagreeing labels for the same text span and add the correct label to the aggregated dataset. Disagreeing labels for the same text span were reviewed for all *Person Name*, *Linguistic*, and *Contextual* categories of labels. For *Person Name* and *Linguistic* labels, where three annotators labeled the same span of text, majority voting determined the correct label: if two out of the three annotators used one label and the other annotator used a different label, the label used by the two annotators was deemed correct and added to the aggregated dataset. For *Contextual* labels, unless an obvious mistake was made, the union of all three annotators' labels was included in the aggregated dataset.

Thirdly, the "Occupation" and "Gendered Pronoun" labels were reviewed. A unique list of the text spans with these labels was generated and incorrect text spans were removed from this list. The "Occupation" and "Gendered Pronoun" labels in the annotated datasets with text spans in the unique lists of valid text spans were added to the aggregated dataset. Fourthly, the remaining *Linguistic* labels ("Gendered Pronoun," "Gendered Role," and "Generalization") not deemed incorrect in the annotated datasets were added to the aggregated dataset. Due to common mistakes in annotating *Person Name* labels with one annotator, only data from the other two annotators who annotated with *Person Name* labels was added to the aggregated dataset. Fifthly, for annotations with overlapping text spans and the same label, the annotation with the longer text span was added to the aggregated dataset. The sixth and final step to constructing the aggregated dataset was to take the union of the remaining *Contextual* labels ("Stereotype," "Omission," "Occupation," and "Empowering") not deemed incorrect in the three annotated datasets with these labels and add them to the aggregated dataset.

B.2 Language Variety

The metadata descriptions extracted from the Archive's catalog are written primarily in British English, with the occasional word in another language such as French or Latin.

B.3 Producer Demographic

The producing research team are of American, German, and Scots nationalities, and are three women and one man. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us is audited an online course on feminist and social justice studies.

B.4 Annotator Demographic

The five annotators are of American and European nationalities and identify as women. Four annotators were hired by the lead annotator for their experience in gender studies and archives. The four annotators worked 72 hours each over eight weeks in 2022, receiving £1,333.44 each (£18.52 per hour). The lead annotator completed the work for her PhD project, which totaled to 86 hours of work over 16 weeks.

B.5 Speech or Publication Situation

The archival metadata descriptions describe material about a range of topics, such as teaching, research, town planning, music, and religion. The materials described also vary, from letters and journals

to photographs and audio recordings. The descriptions in this project's dataset with a known date (which describe 38.5% of the archives' records) were written from 1896 through 2020.

The annotated dataset will be published with a forthcoming paper detailing the methodology and theoretical framework that guided the development of the annotation taxonomy and the annotation process, accompanied by analysis of patterns and outliers in the annotated dataset.

B.6 Data Characteristics

The datasets were organized for annotation in a web-based annotation paltform, the brat rapid annotation tool (Stenetorp et al., 2012). Consequently, the data formats conform to the brat formats: plain text files that end in '.txt' contain the original text and plain text files that end in '.ann' contain the annotations. The annotation files include the starting and ending text span of a label, the actual text contained in that span, the label name, and any notes annotators recorded about the rationale for applying the label they did. The names of all the files consist of the name of the fonds (the archival term for a collection) and a number indicating the starting line number of the descriptions. Descriptions from a single fonds were split across files so that no file contained more than 100 lines, because brat could not handle the extensive length of certian fonds' descriptions.

B.7 Data Quality

A subset of annotations were applied automatically with a grep script and then corrected during the manual annotation process. All three categories of the annotation taxonomy were manually applied by the annotators. The lead annotator then manually checked the labels for accuracy. That being said, due to time constraints, mistakes are likely to remain in the application of labels (for example, the starting letter may be missing from a labeled text span or a punctuation mark may have accidentally been included in a labeled text span).

B.8 Other: Annotation Schema

The detailed schema that guided the annotation process is listed below with examples for each label. In each example, the labeled text is underlined. All examples are taken from the dataset except for labels 1.1, "Non-binary," and 3.4, "Empowering," as the annotators did not find any text to which the provided label definitions applied. The annotation instructions permitted labels to overlap as each annotator saw fit, and asked annotators to read and annotate from their contemporary perspective. The categories of labels from the annotation taxonomy were divided among annotators: two hired annotators labeled with categories 1 and 2, two hired annotators labeled with category 3, and the lead annotator labeled with all categories.

The annotation taxonomy includes labels for *gendered* language, rather than only explicitly gender-biased language, because measuring the use of gendered words across an entire archives' collection provides information about gender bias at the overall collections' level. For example, using a gendered pronoun such as "he" is not inherently biased, but if the use of this masculine gendered pronoun far outnumbers the use of other gendered pronouns in our dataset, we can observe that the masculine is over-represented, indicating a masculine bias in the archives' collections overall. Labeling gender-biased language focuses on the individual description level. For example, the stereotype of a wife playing only or primarily a supporting role to her husband comes through in the following description:

> *Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more. She also wrote a preface to his Baptism and Conversion and a foreward [sic] to his A Reasoned Faith.* (Fonds Identifier: Coll-1036)

1. **Person Name:** the name of a person, including any pre-nominal titles (i.e., Professor, Mrs., Sir, Queen), when the person is the primary entity being described (rather than a location named after a person, for example)

    1.1 **Non-binary:*** the pronouns or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are non-binary

    Example 1.1: Francis McDonald went to the University of Edinburgh where they studied law.
    Note: the annotation process did not find suitable text on which to apply this label in the dataset.

1.2 **Feminine:** the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are feminine

Example 1.2: "Jewel took an active interest in her husband's work..." (Fonds Identifier: Coll-1036)

1.3 **Masculine:** the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are masculine

Example 1.3: "Martin Luther, the man and his work." (Fonds Identifier: BAI)

1.4 **Unknown:** any pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are gender neutral, or no such pronouns or roles are provided within the descriptive field

Example 1.4: "Testimonials and additional testimonials in favour of Niecks, candidacy for the Chair of Music, 1891" (Fonds Identifier: Coll-1086)

2. **Linguistic:** gender marked in the way a word, phrase or sentence references a person or people, assigning them a specific gender that does not account for all genders possible for that person or people

2.1 **Generalization:** use of a gender-specific term (i.e. roles, titles) to refer to a group of people that could identify as more than the specified gender

Example 2.1: "His classes included Anatomy, Practical Anatomy, ... Midwifery and Diseases of Women, Therapeutics, Neurology, ... Public Health, and Diseases of the Skin." (Fonds Identifier: Coll-1118)

2.2 **Gendered Role:** use of a title or word denoting a person's role that marks either a non-binary, feminine, or masculine gender

Example 2.2: "New map of Scotland for Ladies Needlework, 1797" (Fonds Identifier: Coll-1111)

2.3 **Gendered Pronoun:** explicitly marking the gender of a person or people through the use of pronouns (e.g., he, him, himself, his, her, herself, and she)

Example 2.3: "He obtained surgical qualifications from Edinburgh University in 1873 ([M.B.])." (Fonds Identifier: Coll-1096)

3. **Contextual:** expectations about a gender or genders that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (i.e., sentence structure or word choice)

3.1 **Stereotype:** a word, phrase, or sentence that communicates an expectation of a person or group of people's behaviors or preferences that does not reflect the reality of all their possible behaviors or preferences; or a word, phrase, or sentence that focuses on a particular aspect of a person that doesn't represent that person holistically

Example 3.1: "The engraving depicts a walking figure (female) set against sunlight, and holding/releasing a bird." (Fonds Identifier: Coll-1116)

3.2 **Omission:** focusing on the presence, responsibility, or contribution of a single gender in a situation in which more than one gender has a presence, responsibility or contribution; or defining one person's identity in terms of their relation to another person

Example 3.2: "This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version in 1909 entitledApollinaire [sic] and his friends." (Fonds Identifier: Coll-1090).

3.3 **Occupation:** a word or phrase that refers to a person or people's job title (singular or plural) for which the person or people received payment; do not annotate occupations used as a pre-nominal title (for example, "Colonel Sir Thomas Francis Fremantle" should not have an occupation label)

Example 3.3: "He became a surgeon with the Indian Medical Service." (Fonds Identifier: Coll-1096).

3.4 **Empowering:** reclaiming derogatory words or phrases to empower a minoritized person or people

Example 3.4: a person describing themself as queer in a self-affirming, positive manner
*Note: the annotation process did not find enough text on which to apply this label in the dataset to include it when training a classifier. One annotator used the label according to a different definition.***

*The "Non-binary" label was not used by the annotators. That being said, this does not mean there were not people who would identify as non-binary represented in the text of the annotation corpus. When relying only on descriptions written by people other than those represented in the descriptions, knowledge about people's gender identity remains incomplete (Shopland, 2020). Additional linguistic research informed by a knowledge of terminology for the relevant time period may identify people who were likely to identify as non-binary in the corpus of archival metadata descriptions. For example, Shopland (2020) finds that focusing on actions that people were described doing can help to locate people of minoritized genders (and sexualities) in historical texts, but also cautions researchers against assuming too much. A full understanding of a person's gender often remains unattainable from the documentation that exists about them.

**One annotator used the "Empowering" label in the following instances:

- When a person referenced with feminine terms was described as the active party in marriage
- Honor or achievement held by a woman (as indicated in the text)

  Note: Honors and achievements held by men were labeled as stereotypes, as there was a consistent focus on this type of detail about people, which involved spheres of life historically dominated by men in the UK. Spheres of life historically dominated by women in the UK were described with greater vagueness, eliminating the possibility of honors or achievements in these spheres to be identified.
- The fate of a wife is mentioned in an entry predominantly about the life of a husband
- Family members referenced with feminine terms are prioritized (i.e., they are listed first, more detail is given about them than those referenced with masculine terms)
- A gender-neutral term is used instead of gendered term

All annotators were encouraged to use the annotation tool's notes field to record their rationale for particular label choices, especially for text labeled with "Generalization," "Stereotype," or "Omission." The work intends these notes to lend transparency to the annotation process, providing anyone who wishes to use the data with insight onto the annotator's mindset when labeling the archival documentation.

B.9 Provenance Appendix

Data Statement: Corpus of Archival Documentation

B.9.1 Curation Rationale

We (the research team) will use the extracted metadata descriptions to create a gold standard dataset annotated for contextual gender bias. We adopt Hitti et al.'s definition of contextual gender bias in text: written language that connotes or implies an inclination or prejudice against a gender through the use of gender-marked keywords and their context (2019).

A member of our research team has extracted text from four descriptive metadata fields for all collections, subcollections, and items in the Archive's online catalog. The first field is a title field. The second field provides information about the people, time period, and places associated with the collection, subcollection, or item to which the field belongs. The third field summarizes the contents of the collection, subcollection, or item to which the field belongs. The last field records the person who wrote the text for the collection, subcollection, or item's descriptive metadata fields, and the date the person wrote the text (although not all of this information is available in each description; some are empty). Using the dataset of extracted text, we will experiment with training a discriminative classification algorithm to identify types of contextual gender bias. Additionally, the dataset will serve as a source of annotated, historical text to complement datasets composed of contemporary texts (i.e. from social media, Wikipedia, news articles).

We chose to use archival metadata descriptions as a data source because:

1. Metadata descriptions in the Archive's catalog (and most GLAM catalogs) are freely, publicly available online

2. GLAM metadata descriptions have yet to be analyzed at large scale using natural language processing (NLP) methods and, as records of cultural heritage, the descriptions have the potential to provide historical insights on changes in language and society (Welsh, 2016)

3. GLAM metadata standards are freely, publicly available, often online, meaning we can use historical changes in metadata standards used in the Archive to guide large-scale text analysis of changes in the language of the metadata descriptions over time

4. The Archive's policy acknowledges its responsibility to address legacy descriptions in its catalogs that use language considered biased or otherwise inappropriate today[6]

B.9.2 Language Variety

The metadata descriptions extracted from the Archive's catalog are written in British English.

B.9.3 Producer Demographic

We (the research team) are of American, German, and Scots nationalities, and are three females and one male. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us has been auditing a feminism and social justice course, and reading literature on feminist theories, queer theory, and indigenous epistemologies.

B.9.4 Annotator Demographic

Not applicable

B.9.5 Speech or Publication Situation

The metadata descriptions extracted from the Archive's online catalog using Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH). For OAI-PMH, an institution (in this case, the Archive) provides a URL to its catalog that displays its catalog metadata in XML format. A member of our research team wrote scripts in Python to extract three descriptive metadata fields for every collection, subcollection, and item in the Archive's online catalog (the metadata is organized hierarchically). Using Python and its Natural Language Toolkit library (Loper and Bird, 2002), the researcher removed duplicate sentences and calculated that the extracted metadata descriptions consist of a total of 966,763 words and 68,448 sentences across 1,231 collections. The minimum number of words in a collection is 7 and the maximum, 156,747, with an average of 1,306 words per collection and standard deviation of 7,784 words. The archival items described in resulting corpus consist of a variety of material, from photographs and manuscripts (letters, lecture notes, and other handwritten documents) to instruments and tweets.

B.9.6 Data Characteristics

Upon extracting the metadata descriptions using OAI-PMH, the XML tags were removed so that the total words and sentences of the metadata descriptions could be calculated to ensure the text source provided a sufficiently large dataset. A member of our research team has grouped all the extracted metadata descriptions by their collection (the "fonds" level in the XML data), preserving the context in which the metadata descriptions were written and will be read by visitors to the Archive's online catalog.

B.9.7 Data Quality

As a member of our research team extracts and filters metadata descriptions from the Archive's online catalog, they write assertions and tests to ensure as best as possible that metadata is not lost or unintentionally changed.

B.9.8 Other

The data can be freely accessed at: `datashare.ed.ac.uk/handle/10283/3794`. The data preparation code has been published at: `github.com/thegoose20/annot-prep`.

B.9.9 Provenance Appendix

The data described above was harvested from the University of Edinburgh's Centre for Research Collections' Archives catalog in 2020 (archives.collections.ed.ac.uk).

[6]The Archive is not alone; across the GLAM sector, institutions acknowledge and are exploring ways to address legacy language in their catalogs' descriptions. The "Note" in We Are What We Steal provides one example: `dxlab.sl.nsw.gov.au/we-are-what-we-steal/notes/`.

C Annotation Instructions

The annotation instructions were written to guide annotators in applying the taxonomy of to the annotation corpus of archival metadata descriptions. Prior to beginning the annotation process, an annotation pilot was undertaken with three participants to test the clarity of the annotation taxonomy. The pilot led to revisions of the instructions: more examples were added and annotators were explicitly instructed to read and interpret the descriptions from their contemporary perspective.

The annotation instructions below contain a slightly different annotation taxonomy than the final annotation taxonomy included above in the main body of the paper. This is due to the fact that during and after the annotation process, the taxonomy was revised based on the data that was being annotated. The definitions of Gendered Role and Generalization proved to be difficult to distinguish in practice, so the definitions were revised during the dataset aggregation process. Additionally, we realized during the annotation process that "Woman" and "Man" were inaccurate labels based on what we could learn about gender from text, so we changed these labels to "Feminine" and "Masculine," respectively, for the final annotation taxonomy.

C.1 Instructions

Step 1: As you read and label the archival metadata descriptions displayed on the screen, including text that quotes from source material, meaning text surrounded in quotation marks that reproduces something written in a letter, manuscript, or other text-based record from an archival collection.
NOTE: If you are unsure about an annotation, please make a note the file name and your question so that we can discuss it and decide on the way to annotate that sort of language moving forward!
Step 2: Please note that Gendered-Pronouns, Gendered-Roles, and Occupations have been pre-annotated. If any of these three categories of language have been annotated incorrectly, please correct them by clicking on the annotation label, deleting it, and making the correct annotation. If any of these three categories of language have been missed in the pre-annotation process, please annotate them yourself.
Step 3: Read the archival metadata descriptions displayed and while reading:

- Use your mouse to highlight a selection of text or click on a word that uses gendered language according to the schema in the table on the next page.
- Using the keyboard shortcuts (see the table) or your mouse, select the type of gendered language you've identified. Please select the most specific label possible (listed as i, ii, iii, or iv)! Please only select Person-Name, Linguistic or Contextual if you do not feel their subcategories are suitable to the gendered language you would like to annotate.
- If you select a subcategory of Contextual gendered language, please write a brief note explaining what you've annotated as gendered in the "Notes" section of the "New/Edit Annotation" pop-up window.
- If you used your mouse to open the pop-up window, press the Enter/Return key or the "OK" button to make the annotation.
- You may make overlapping annotations, meaning a single word or phrase may have multiple gendered language annotations.
- Please annotate all instances of a particular type of gendered language used for a specific person or people in the text.
- Please note that the labels to annotate with as defined below are intended to guide your interpretation of the text through a contemporary lens (not a historical lens).

The examples provided in the schema below are highlighted according to the words, phrases or sentences that should be highlighted or clicked in brat. If in doubt about how much to annotate, please annotate more words rather than less!

1. **Person-Name**: the name of a person including any pre-nominal titles they have (i.e., Professor, Mrs., Sir)

   NOTE 1: Please annotate every instance of a name in brat only (do not use a spreadsheet anymore). This means that each person may have multiple person-name labels annotating the same form of their name or different forms of their name.

   NOTE 2: Use the pronouns and roles that occur within the descriptive field in which the name appears (either "Title," "Scope

and Contents," "Biographical / Historical," or "Processing Information") to determine whether the annotation label should be Woman, Man, Nonbinary, or Unknown. Please do not use the occupation, name, or other information that implies a gender to determine the annotation label; only use explicit terms such as gender-marking pronouns (him, her, he, she, himself, herself, etc.) and gender-marking roles (mother, father, daughter, wife, husband, son, Mrs, Ms, Mr, etc.).

(a) **Woman**: the pronouns (i.e., she, her) or roles (i.e., mother, wife, daughter, grandmother, Mrs., Ms., Queen, Lady, Baroness) or use of term *nee [Last Name]* indicating a maiden name within the descriptive field in which the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") of the named person suggest they are a woman
Example: Mrs. Jane Bennet went to Huntsford.

(b) **Men**: the pronouns, roles, or titles of the named person suggest they are a man
Example: Conrad Hal Waddington lived in Edinburgh and he published scientific papers.

(c) **Non-binary**: the pronouns or roles of the named person within the descriptive field in which this instance of the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") suggest they are non-binary
NOTE: a preliminary search of the text returned no results for exclusively non-binary pronouns such as Mx, so most likely any non-binary person would be indicated with "they"); if the gender of a person is named and it's not a woman or man, please note this gender in the "Notes" section of the annotation pop-up window
Example: Francis McDonald went to the University of Edinburgh where they studied law.

(d) **Unknown**: there are no pronouns or roles for the named person within the descriptive field in which this instance of the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") that suggest their gender identity
Example: Jo McMahon visited Edinburgh in 1900.

2. **Linguistic**: gender marked in the way a sentence references a person or people, assigning them a specific gender that does not account for all genders possible for that person or group of people (Keyboard shortcut: L)

(a) **Generalization**: use of a gender-specific term to refer to a group of people (including the job title of a person) that could identify as more than the specified gender (Keyboard shortcut: G)
Example 1: The chairman of the university was born in 1980. Explanation: Chair would be the gender-neutral form of chairman
Example 2: Readers, scholars, and workmen Explanation: readers and scholars are gender-neutral, while workpeople or workers would be the gender-neutral form of workmen
Example 3: Housewife

(b) **Gendered Pronoun**: explicitly marking the gender of a person or people through the use of the pronouns he, him, his, her, and she (Keyboard shortcut: P)
Example 1: She studied at the University of Edinburgh. In 2000, she graduated with a degree in History.
Example 2: This manuscript belonged to Sir John Hope of Craighill. Sir John Hope was a judge. He lived in Scotland.

(c) **Gendered Role**: use of a title or word denoting a person's role that marks either a masculine or feminine gender (Keyboard shortcut: R)
Example 1: Sir Robert McDonald, son of Sir James McDonald
Example 2: Mrs. Jane Do
Example 3: Sam is the sister of Charles
Example 4: Sir Robert McDonald, son of Sir James McDonald

3. **Contextual**: gender bias that comes from knowledge about the time and place in which language is used, rather than from linguistic

patterns alone (i.e., sentence structure, word choice) (Keyboard shortcut: C)

(a) **Occupation**: occupations, whether or not they explicitly communicate a gender, should be annotated, as statistics from external data sources can be used to estimate the number of people of different genders who held such occupations; please label words as occupations if they'd be a person's job title and are how the person would make money, but not if the words are used as a title (Keyboard shortcut: J)

Example 1: minister

Example 2: Sergeant-Major-General

(b) **Stereotype**: language that communicates an expectation of a person or group of people's behaviors or preferences that does not reflect the reality of all possible behaviors/preferences that person or group of people may have, or language that focuses on a particular aspect of a person that doesn't represent that person holistically; for example, women described in relation to their family and home, and men in relation to their careers and workplace; men more associated with science and women more associated with liberal arts (Keyboard shortcut: S)

NOTE: Please label whichever words, phrases, or sentences you feel communicate the stereotype. Three different examples are shown below for how this may look. Include names being turned into ways of thought (e.g., Bouldingism, Keynsian).

Example 1: The event was sports-themed for all the fathers in attendance. *Explanation: The assumption here is that all fathers and only fathers would enjoy a sports-themed event. A neutral alternative sentence could read: The event was sports-themed for all the former athletes in attendance*

Example 2: A programmer works from his computer most of the day. *Explanation: The assumption here is that any programmer must be a man, since the indefinite article "A" is used with the pronoun "his"*

Example 3: A man with no doctorate degree being known as Dr. Jazz *Explanation: Women often receive negative attention for using titles such as Dr (see the WSJ op-ed on Dr Jill Biden for a recent example) while men typically do not*

(c) **Omission**: focusing on the presence, responsibility, or contribution of a single gender in a situation in which more than one gender has a presence, responsibility or contribution; or defining a person's identity in terms of their relation to another person (Keyboard shortcut: O)

NOTE: If initials are provided, consider that enough of a name that it doesn't need to be labeled as an omission!

Example 1: Mrs. John Williams lived in Edinburgh. *Explanation: Mrs. John Williams is, presumably, referred to by her husband's first and last name rather than her given name*

Example 2: Mr. Arthur Cane and Mrs. Cane were married in 1850. *Explanation: Mrs. Cane is not referred to by her given name*

Example 3: Mrs. Elizabeth Smith and her husband went to Scotland. *Explanation: The husband is not named, being referred to only by his relationship to Mrs. Elizabeth Smith*

Example 4: His name was Edward Kerry, son of Sir James Kerry. *Explanation: paternal relations only, no maternal relations*

Example 5: The novelist, Mrs. Oliphant, wrote a letter. *Explanation: Mrs. Oliphant is referred to by the last name she shares with her husband without including her given name*

(d) **Empowering**: use of gendered language to challenge stereotypes or norms that reclaims derogatory terms, empowering a minoritized person or people; for example, using the term queer in an empowering rather than a derogatory manner (Keyboard shortcut: E)

Example: "Queer" being used in a self-affirming, positive manner to describe oneself

Step 4: If you would like to change an annotation you have made, double click the annotation label.

If you would like to remove the annotation, click the "Delete" button in the pop-up window. If you would like to change the annotation, click the label you would like to change to and then click the "OK" button.

Step 5: Click the right arrow at the top left of the screen to navigate to the next archival metadata description (if you would like to return to a previous description, click the left arrow).

Step 6: If the screen does not advance when you click the right arrow, you've reached the end of the folder you're currently in. To move onto the next file, please hover over the blue bar at the top of the screen and click the "Collection" button. Click the first list item in the pop-up window "../" to exit your current folder and then double click the next folder in the list. Double click the first file in this next folder to begin annotating its text.

Step 7: Repeat from step 1.