Multilingual Contextual Affective Analysis of LGBT People Portrayals in Wikipedia

Chan Young Park[*] **Xinru Yan**[*] **Anjalie Field**[*] **Yulia Tsvetkov**
Language Technologies Institute
Carnegie Mellon University
{chanyoun, xinruyan, anjalief, ytsvetko}@cs.cmu.edu

Abstract

Specific lexical choices in how people are portrayed both reflect the writer's attitudes towards people in the narrative and influence the audience's reactions. Prior work has examined descriptions of people in English using *contextual affective analysis*, a natural language processing (NLP) technique that seeks to analyze how people are portrayed along dimensions of *power*, *agency*, and *sentiment*. Our work presents an extension of this methodology to *multilingual* settings, which is enabled by a new corpus that we collect and a new multilingual model. We additionally show how word connotations differ across languages and cultures, which makes existing English datasets and methods difficult to generalize. We then demonstrate the usefulness of our method by analyzing Wikipedia biography pages of members of the LGBT[1] community across three languages: English, Russian, and Spanish. Our results show systematic differences in how the LGBT community is portrayed across languages, surfacing cultural differences in narratives and signs of social biases. Practically, this model can be used to surface Wikipedia articles for further manual analysis—articles that might contain content gaps or an imbalanced representation of particular social groups.

1 Introduction

In 1952, Alan Turing was prosecuted for being gay and subsequently underwent a hormonal injection; two years later he committed suicide. Figure 1 shows parallel sentences drawn from his English, Spanish, and Russian Wikipedia pages. Although all three sentences describe the same situation, their connotations subtly differ. The English edition uses the verb *accepted*, which suggests that Turing had little control over the situation (low agency). In contrast, the verbs *chose* in Spanish and *preferred* in Russian imply that he actively made the decision (high agency). The verb *preferred* in Russian can even imply positive sentiment towards the injections, while the English connotation is more negative. Thus, Russian, Spanish, and English readers who search for Alan Turing on Wikipedia may form different impressions about this part in his life.

English Wikipedia:
He *accepted* the option of injections of what was then called stilboestrol.

Spanish Wikipedia:
Finalmente escogió las inyecciones de estrógenos.
Finally he chose estrogen injections.

Russian Wikipedia:
Учёный предпочёл инъекции стильбэстрола
The scientist preferred stilbestrol injections.

Figure 1: Example from Alan Turing's biography page on Wikipedia in different languages. Verb choice in different languages can have subtly different connotations.

These subtle differences in phrasing can be indicative of social norms and perceptions about social roles (Eckert 2000; Tannen 1994). In general, analyzing narratives about people can shed light on stereotypes and power structures (Hall and Braunwald 1981; Fournier, Moskowitz, and Zuroff 2002) and examining how these concepts differ across cultures is an important component of social-oriented analysis (Almeida et al. 2009; Balsam et al. 2011; Harris et al. 2013). In the example in Figure 1, these discrepancies in phrasing could indicate that stereotypes and bias about LGBT people manifest differently in Russian, English, and Spanish-speaking cultures. On Wikipedia, which strives to present a "Neutral Point of View",[2] they can be signs of problematic bias.

In this work, we develop computational methods that facilitate large-scale analyses of people described in multilingual narrative text. This technology can aid readers or writers, such as Wikipedia editors or journalists, in identifying biases in sets of articles that can be further analyzed to understand social stereotypes or edited in order to reduce bias.

Recent advances in NLP have analyzed stereotypes and biases in narratives in English (Bamman, O'Connor, and Smith 2013; Wagner et al. 2015; Sap et al. 2017). Field, Bhat, and Tsvetkov (2019) establish a framework called Contextual Affective Analysis (CAA) that focuses on affective

[*] Equal contribution

[1] Our research focuses on the LGBT sub-community of the LGBTQIA+ community due to data scarcity of other groups.

[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

dimensions of *power* (strength/weakness), *agency* (activeness/passiveness), and *sentiment* (goodness/badness). Analyzing portrayals of people along these dimensions (e.g., are men or women portrayed as more powerful?) has revealed stereotypes and bias in various domains, including movie scripts and newspaper articles (Rashkin, Singh, and Choi 2016; Sap et al. 2017; Field and Tsvetkov 2019; Field, Bhat, and Tsvetkov 2019; Antoniak, Mimno, and Levy 2019).

Measuring these affective dimensions relies on *connotation frames*—lexicons of verbs annotated to elicit implications (Rashkin, Singh, and Choi 2016; Sap et al. 2017). These connotations can be subtle, and the same verb often has different connotations in different contexts (Field, Bhat, and Tsvetkov 2019; Field and Tsvetkov 2019). Until now, these manually-annotated affective lexicons have existed only for English. While lexicons can be machine-translated (Rashkin et al. 2017; Mohammad 2018), no work has yet conducted in-language evaluations of connotation translatability nor attempted extensive analysis in other languages.

Our ultimate goal is to measure the power, agency, and sentiment of people described in multilingual text; we here focus on English, Spanish, and Russian. These measurements allow us to conduct both in-language analysis (in Russian text, are LGBT people portrayed as more powerful than non-LGBT people?) and cross-language analysis (is the power differential between LGBT people and non-LGBT people greater in English or in Russian?). To accomplish this, we first crowdsource annotations of connotation frames in English, Spanish, and Russian (§2). We then analyze how connotations vary across contexts and languages in these data in order to demonstrate why existing English data sets are insufficient (§3). With the new dataset, we develop multilingual CAA classifiers of power, agency, and sentiment (§4).

Finally, we demonstrate the usefulness of our methodology in a semi-automated analysis §6, by collecting a new corpus (LGBTBio) and analyzing how members of the LGBT community are portrayed on Wikipedia in different languages. Our results show that the biography pages of LGBT people typically contain more negative connotations than the pages of other people, and that these trends can differ across languages; for example, Russian-language pages about LGBT people of Russian nationality are less negative than English or Spanish pages about the same people.

The key contributions of our work are annotated datasets, machine learning models, and a general methodology that enables nuanced analyses of narratives about people across languages. We additionally present an analysis of LGBT people on Wikipedia, whereas extensive prior computational work on Wikipedia biography pages has focused primarily on male/female gender bias (Callahan and Herring 2011; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013; Wagner et al. 2015; Chandrasekharan et al. 2017).[3]

2 Crowdsourcing Contextualized Connotation Frames

We first collected a corpus of multilingual connotation frames in English, Spanish, and Russian. Connotation frame annotations ask annotators to answer questions about the power, sentiment, and agency of the agent (approximated as the grammatical subject) and theme (approximated as the object) of verbs. At a high level, we seek to answer: (1) Does the subject have more/less/equal **power** as the object? (2) Does the subject have low/moderate/high **agency**? (3) Does the writer feel positive/negative/neutral about the subject ($\text{Sent}_{\text{subj}}$)? (4) Does the writer feel positive/negative/neutral about the object (Sent_{obj})?

Consider the sentence: *The firefighter rescued the boy.* The verb *rescued* implies that the subject, *firefighter* has more power than the object, *boy*. The firefighter is also active and in control of his actions, which shows high agency. Because *rescuing* is a positive action (e.g. as opposed to *killing*), the writer likely feels positively about the subject. In the absence of information conveying positive or negative sentiment about *boy*, we can infer that the writer feels neutrally about the object.

Our connotation frames differ from existing lexicons in two primary ways: first, no prior work has collected connotation frames in languages other than English, and second, we collect all annotations in complete contexts drawn from newspaper articles, e.g., *the firefighter rescued the boy*, whereas prior work uses either simplified tuples or artificial placeholders, e.g., *X rescues Y* (Sap et al. 2017; Rashkin, Singh, and Choi 2016). Because these affective dimensions can be difficult to define, we took numerous steps to ensure annotations would be of high quality. We briefly summarize here and provide details in Appendix A to facilitate reproducibility.

We first extracted contexts that are representative of each verb's most common usage, and then asked annotators to annotate verbs in these contexts. For each language, we extracted all (subject, object, verb) tuples from a *News Crawl* corpus.[4] We chose the 300 most frequent transitive verbs to annotate. For each verb, we took the three most common (subject, object, verb) tuples as the most representative context. We restricted tuples to have at least one human subject or object by using the list of words in *noun.person* category of WordNet (Fellbaum 2012).[5] We then pulled phrases containing the chosen tuples from the news corpus, which served as our data to be annotated.

We used the same interfaces as Rashkin, Singh, and Choi (2016) and Sap et al. (2017), with minor modifications based on feedback during pilot studies.[6] For non-English annotation tasks, a native speaker translated the task instructions into the target language. We restricted the pool of annotators to the United States for English, Russia for Russian, and to several South American countries for Spanish.

[3] Code and data will be made publicly available, to facilitate future work in this area.

[4] A large monolingual corpus of newspaper articles (Barrault et al. 2019). Throughout this work, parsing was done with SpaCy.

[5] Extensive list of English nouns denoting people (Miller 1995); we translated to other languages with Google Translate.

[6] We will provide full annotation interface in the released data.

	English	Russian	Spanish
Power	↓29.2%	↓24.5%	↓26.7%
Agency	↓31.1%	↓30.9%	↓35.2%
Sent(subj)	↓18.5%	↓18.4%	↓29.6%
Sent(obj)	↓20.2%	↓23.6%	↓29.8%

Table 1: Assessment of how much information is lost when using verb-level ("rescues") annotations instead of context-level ("the firefighter rescued the boy"). Accuracy decreases by nearly 20% for all languages.

	Russian	Spanish
Power	↓40.1%	↓48.3%
Agency	↓54.2%	↓48.3%
Sent(subj)	↓19.9%	↓35.6%
Sent(obj)	↓34.7%	↓38.4%

Table 2: Assessment of information lost when using translated annotations instead of in-language annotations, evaluated over 147 Russian and 149 Spanish verbs that overlapped with English annotations.

For each of the three target languages, we collected power, agency, and sentiment annotations for 300 verbs in three contexts each (900 instances). For each instance, we collected judgements from three annotators, leading to $32,400$ total annotations.

Despite the steps taken to ensure annotation quality, we suspect that some annotators paid more attention to the task instructions and generated higher-quality judgements than others. We correct for this by discarding annotations from 14.4% of workers who frequently disagreed with other annotators (Appendix A). Krippendorff's alpha, averaged across tasks, was 0.22 for English, 0.31 for Russian, and 0.26 for Spanish. These agreement scores are comparable to prior work (Rashkin, Singh, and Choi 2016; Sap et al. 2017) and reflect the subjective nature of connotation frames; very high agreement would suggest that we over-simplified the task, for example by choosing non-representative samples to annotate. Additionally, the most common case of annotator disagreement was when one annotator labeled an instance as neutral and another did not, meaning polar-opposite annotations were rare; if we only count polar-opposite annotations as disagreements, the average pairwise agreement is 92.5%. We describe additional restrictions used to ensure annotation quality and full agreement metrics in Appendix A.

To aggregate annotations, we mapped each judgement to a $(-1, 0, 1)$ value and averaged annotator scores. We then ternerized the aggregated scores by labeling connotations as positive $[1, 0.35]$, neutral $(0.35, -0.35)$, and negative $[-0.35, -1]$. With these boundaries, a connotation is only scored as positive or negative if at least two annotators labeled it with this polarity, and thus, samples where annotators disagreed were labeled as neutral–not clearly indicative of a positive or negative connotation.

3 Crowdsourced Connotation Analysis

As described in §2, our data differs from existing lexicons in two primary ways: (1) we collected annotations in context and (2) in various languages. We analyzed our data to assess how these differences impact lexicon quality.

How important is contextualization? We first address this question by examining how much accuracy would be lost if we use a single connotation score for each verb (e.g., one score for *deserves* in all contexts where it appears), rather than different scores when the verb appears in different contexts (e.g., different scores for *the boy deserves a reward* and *the boy deserves a punishment*) (Field and Tsvetkov 2019). To compute this, we "decontextualize" our lexicons by averaging annotations for each verb, regardless of the context it was annotated in, into a single verb-level score. We compare this decontextualized score with the contextualized score in our actual dataset, where we only average annotations for verbs annotated in the same context. Table 1 shows how often the verb-level score differs from the context-level score. If verbs had the same connotations in different contexts, all values in this table would be 0%. Instead, we see that ignoring contextualization can result in an $> 30\%$ drop in accuracy. While Field and Tsvetkov (2019) have similar findings for sentiment connotations in English, Table 1 extends these findings to power and agency connotations and to Spanish and Russian.

How important are in-language annotations? We cannot directly compare contextualized annotations across languages because we annotate different contexts for different languages; however, there is some overlap between the most frequent verbs in each language. For each non-English language, we first aggregate annotations into the same decontextualized verb-level scores as in the previous paragraph. We then use Google Translate to translate verbs into English and intersect them with our annotated English verbs.[7] Finally, we measure how often the decontextualized English annotations differ from the original in-language annotations.

Table 2 reports the results. Because we assess the decontextualized scores, if word-level translation were effective for obtaining multilingual connotations, all scores would be similar to the ones in Table 1. Instead, they are substantially higher, showing that information is lost because of translation. Both Tables 1 and 2 suggest that our annotations can facilitate higher-quality analyses than ones collected in prior work. Because word-level translations are often inaccurate, in §5, we also explore using a cross-lingual model and sentence-level translation to translate connotations.

4 Classification of Connotations

Our goal is to develop methodology for analyzing how people are portrayed in different languages. The multilingual annotations alone are insufficient, as 900 contexts represent a tiny subset of all verb usages in a corpus. Thus, we need a method to obtain connotation scores for unseen verbs and

[7] Google Translate has previously been used to obtain multilingual lexicons (Mohammad and Turney 2013)

Tgt	Src	$\text{Sent}_{\text{subj}}$	Sent_{obj}	Pow.	Agen.
EN	EN	**43.4**	43.0	**41.1**	**48.2**
	ES	38.1	43.4	29.5	43.4
	RU	41.1	**44.3**	40.1	41.4
ES	EN	38.9	36.6	24.5	31.3
	ES	**49.5**	**51.2**	**43.6**	**43.6**
	RU	39.0	42.2	34.0	38.9
RU	EN	43.6	49.2	36.4	44.5
	ES	37.2	49.3	38.2	42.7
	RU	**46.4**	**54.9**	**45.3**	**49.9**

Table 3: Macro F1 score of classifiers trained and evaluated with different target and source languages. Matching the language of the training and test data achieves better results than training on different languages.

contexts. We describe our method here and provide reproducibility details in Appendix B.

We follow prior work in developing a supervised classifier trained on our contextualized multilingual annotations that can predict a connotation frame label y_v for any unseen (in-context) verb v (Rashkin, Singh, and Choi 2016; Field, Bhat, and Tsvetkov 2019). Unlike prior models, ours is trained on contextualized annotations, and it leverages pretrained cross-lingual language models (CLMs). CLMs produce language-agnostic feature representations, allowing us to combine different languages in the training and test data.

We obtain multilingual embedding representations (c_v) of verbs in-context by extracting the last hidden layer of the pretrained model called XLM, which achieves state-of-the-art performance in a variety of cross-lingual tasks (Conneau and Lample 2019). We then use c_v as features in a classifier.

Our primary classifier is a logistic regression model with sample weighting. The classifier is trained to predict a connotation frame label y_v from the input representation c_v. Sample weights are tuned over a dev set. While the classifier architecture is the same as in (Rashkin, Singh, and Choi 2016; Field, Bhat, and Tsvetkov 2019), inputs to our model differ. Prior connotation frame lexicons only contain verb-level annotations, and classifiers were trained using non-contextual embeddings (e.g., word2vec (Rashkin, Singh, and Choi 2016)) or decontextualized embeddings (e.g., ELMo (Field, Bhat, and Tsvetkov 2019)). With new annotations over verbs in-context, we directly train the classifier on contextual embeddings. Finally, we use the trained classifier to predict connotation frame labels for verbs provided with their context in a target language corpus.

5 Connotation Classification Evaluation

Tgt	$\text{Sent}_{\text{subj}}$	Sent_{obj}	Power	Agency
ES	21.1	20.1	31.7	27.3
RU	18.9	30.8	34.4	24.2

Table 4: Macro F1 for Machine Translation approach is strictly worse than for cross-lingual model (Table 3).

Tgt	Src	S_{subj}	S_{obj}	Pow.	Agen.
EN	EN	43.4	43.0	41.1	48.2
	+ES	44.8	**45.2**	40.5	49.7
	+RU	**46.5**	43.2	**41.8**	49.9
	+ES+RU	45.0	44.3	41.7	**50.0**
ES	ES	49.5	51.2	43.6	43.6
	+EN	50.4	51.6	36.4	45.5
	+RU	51.0	**55.0**	**42.1**	**45.6**
	+EN+RU	**51.8**	54.8	40.8	44.9
RU	RU	46.4	54.9	45.3	49.9
	+EN	45.6	55.7	44.1	50.9
	+ES	46.0	**59.2**	42.1	49.8
	+EN+ES	**47.7**	53.7	**46.9**	**51.7**

Table 5: Macro F1 score of classifiers, where in-language training data is augmented with training data from other languages (e.g., Tgt=EN, Src=+ES indicates the model was trained on English and Spanish data). Augmentation improves performance in all cases.

We evaluate our model on the contextualized multilingual annotation data described in §2 using 5-fold cross-validation and splitting data into train, development, and test sets with the ratio of 6:2:2. Table 3 reports macro F1 scores[8] for single-language evaluations, where we train on the training set of one language ("Src") and evaluate on the test set of a second language ("Tgt"). Unsurprisingly, training and testing on the same language achieves better performance than training on one language and testing on another. While the cross-lingual model works well in certain cases, in most cases transferring languages yields a substantial decrease in performance. For power in Spanish, F1 decreases by 19 points (44%) when the training language is English instead of Spanish. These results offer further evidence of the importance of in-language connotations.

Machine Translated Classification In Table 4, we consider an alternative approach to in-language annotations. We translate the Spanish and Russian test sentences into English through Google Translate, and then we use the model trained on English annotations to predict connotations in these translated sentences. This method simulates translating a corpus into English, and using a model trained on English for analysis. Machine translation performs strictly worse than the cross-lingual model, suggesting it cannot replace in-language data.

Augmented Cross-Lingual Connotation Classification While Table 3 suggests that in-language training data results in better performance than out-of-language training data, we also explore using in-language and out-of-language data in combination. This experimentation is based on the hypothe-

[8] We provide evidence that our model performs comparably with prior work in Appendix C

sis that while connotations differ in different languages, there may be enough overlap for the cross-lingual model to learn useful signals from out-of-language data. Table 5 shows results. For all connotation dimensions, the best performing augmented models outperform the non-augmented models. In §6, we use the best performing models from Table 5 to analyze biography pages of LGBT people.

6 Case Study: Multilingual Affective Analysis of LGBT People

Finally, we demonstrate how the new corpus of annotations (§2) and the cross-lingual model (§4) facilitate multilingual analysis by examining portrayals of LGBT people on Wikipedia. We focus on LGBT people because discrimination against the LGBT community is an increasingly important global issue. Although pride marches are held ≥ 158 cities world-wide (Lisitza 2017), social oppression is prevalent in many countries (Balsam et al. 2011; Doi and Stewart 2019). Nevertheless, almost no prior computational work has studied narratives about the community, likely due to data scarcity. To facilitate analysis, we collect a new corpus, titled LGBTBio, which contains Wikipedia biography pages of LGBT people and pages of non-LGBT people (controls). We explain the motivation and methods behind our corpus collection in §6.1 and discuss results in §6.2.[9]

6.1 LGBTBio Corpus

We collected a multilingual corpus of $1,044$ (L: 174; G: 570, B: 254 T: 43) Wikipedia biography pages for people in the LGBT community using lists of LGBT people from Wikipedia (details in Appendix D). This corpus allows us to analyze how the same person is portrayed in different languages. However, we cannot draw conclusions from this corpus alone because we need to control for overall language differences. For example, if we find that LGBT people have higher power in English pages than in Russian pages, we cannot determine if this difference occurs because LGBT people are actually described differently or because English verbs tend to have more positive power connotations than Russian verbs.

To isolate the effects of our variable of interest (i.e., sexual orientation), we built a control corpus by matching each LGBT person with a non-LGBT person who has similar characteristics using a matching method from Anonymous (2020). Our goal is not to construct perfectly-matched pairs, and we do not analyze any single pair. Rather, we seek to build a control corpus that has a similar distribution of characteristics as the main corpus *on aggregate*, excepting of sexual orientation, which is a common practice in causal inference studies (Stuart 2010).

To identify matches, for each person Y in the LGBT corpus, we first scraped their associated Wikipedia categories (e.g., *American fashion businesspeople*). We then identified all other people listed in these categories and constructed weighted TF-IDF vectors from each person's associated categories. We selected the control person as the one who has the most similar category vector as Y. We refer to Appendix D and Anonymous (2020) for details about the corpus and matching algorithm. The 1,044 pages about LGBT people and their matched controls together constitute the LGBTBio corpus.

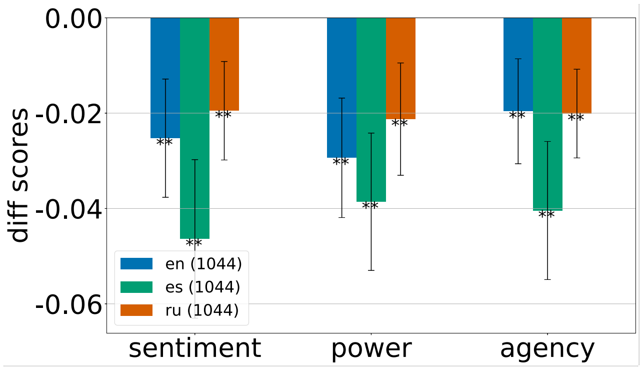

Figure 2: Average differences in affective scores in narratives about LGBT people vs. matched control people across languages. In all languages, LGBT people are consistently portrayed more negatively, with lower power, and with lower agency. For all figures, asterisks indicate scores are statistically different from zero (independent t-test, $*$:p<0.05 and $**$:p<0.01) and brackets denote 0.95 confidence intervals. Numbers in the legend or in the title indicate the number of biographies in each group.

6.2 Contextual Affective Analysis of Narratives Describing LGBT People

We first use our model to compute high-level trends that identify directions for further exploration. We then leverage our model to extract samples that exemplify these trends, which are manually analyzed and inform the computation of corroborating statistics. Given the difficulty of the task, the subjective nature of connotations, and the focus of this work on a sensitive social dimension, we do not use our model off-the-shelf to draw black-box inferences (and we do not recommend that others use it in this way), but rather use it to facilitate manual analyses that would otherwise be prohibitively expensive given the volume of Wikipedia and other Web-scale data.

For each language, we used the best performing model from Table 5 to predict contextualized verb annotations for all sentences that contain a target person's name or pronoun as the subject of a verb.[10] We then mapped these verb scores to entities (LGBT/matched control people) following Field, Bhat, and Tsvetkov (2019). We report *diff score* as the difference between sentiment/power/agency scores averaged across sentences about LGBT people and scores averaged

[9] All of our analysis is conducted over publicly available data. We do not expect our results to have any negative effects on the individuals analyzed or the broader LGBT community.

[10] We omitted Sent(Obj) and focused on Sent(Subj) since sentences where the individual was an object are rare. All subgroups analyzed contained at least 280 verbs.

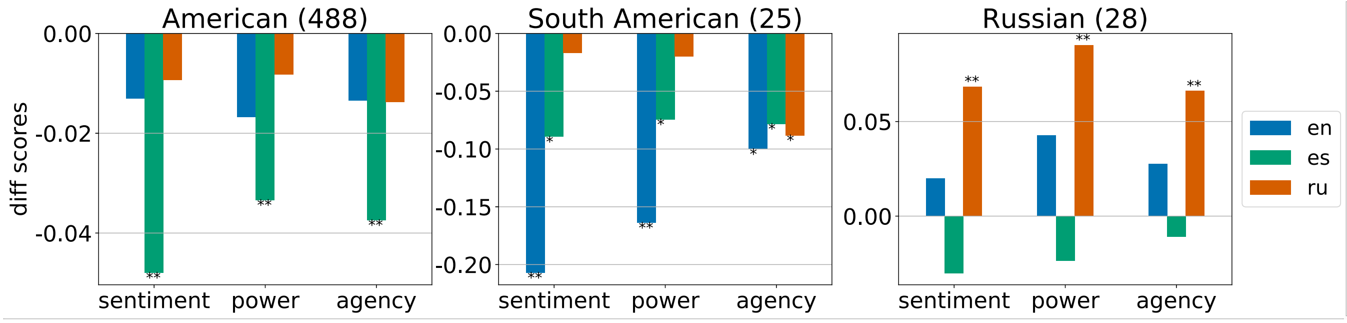

Figure 3: Average differences in affective scores for narratives about nationality subgroups in LGBTBio.

across sentences about matched controls (e.g. "average treatment effect"); a positive score means the LGBT people had a higher affect connotation in aggregate.

Figure 2 shows diff scores across the entire corpus. In all languages, connotations are negative for all dimensions, suggesting the Wikipedia biography pages portray LGBT people as having lower power, agency, and sentiment than non-LGBT people. While the magnitudes of the difference are similar for Russian and English, the largest difference occurs for Spanish articles, particularly for sentiment and agency.

Differences Between Languages Suggest Bias in Wikipedia Portrayals In order to further examine possible cultural differences and Wikipedia biases, we divided people in the LGBT corpus by nationality, restricting the corpus to biographies of LGBT people from the United States, South America, and Russia and their matched controls. We report diff scores for each subgroup in Figure 3.[11]

Each subgraph in Figure 3 displays connotation scores for the exact same set of people. We see similar local biases as prior work, where biography pages tend to be longer and more positive for people whose nationality match the language the page is written in (Callahan and Herring 2011; Eom et al. 2015). In this case, we see less evidence of bias against LGBT people for people of the same nationality as the language of the article.

In the left-most graph in Figure 3, which displays scores for U.S. people, sentiment is near-neutral in English articles. Power and agency are significantly negative, but they are less negative than Spanish versions of the same articles. In contrast, for South American people, the English articles are more negative for all connotations than Spanish versions, even though the Spanish articles are the most negative over the entire corpus (Figure 2).

Surprisingly, in Russian, the articles about LGBT people from Russia are more positive than controls for all three dimensions. We examined this trend by using our model to identify the biography pages which were most positive in Russian but most negative in English. In these pages, the English articles focused more on the sexual orientation of the person than the Russian articles. For example, the English article about Nikolai Zverev states that he "never married, and may have been homosexual", which is scored as implying a negative sentiment connotation, while the Russian article does not mention this at all. Similarly, for Anton Krasovsky, the English article focuses only on how he was fired (or left his job, depending on the source), after stating he was gay on the radio. The Russian article describes this incident as well, but also discusses other aspects of his life, including a lengthy description of his total career path. Thus, our results suggest that English articles over-emphasize sexual orientation, describing it even in articles that are not detailed, whereas Russian articles comparatively under-emphasize sexual orientation, only describing it in passing in detailed articles and omitting it in less-detailed articles. In order to substantiate this finding, we compiled a list of terms referring to sexual orientation such as "gay" and "transgender" for each language. We then counted what percent of sentences in each article use at least one of these terms, averaged across articles. In the set of biographies for LGBT people with Russian nationality, in English and Spanish 5% of sentences used one of these terms, compared to only 3% in Russian. Thus, while Figure 2 suggests that articles in English, Russian, and Spanish describe LGBT people with more negative connotations than non-LGBT people, Figure 3 suggests that the difference in portrayal is stronger for people whose nationality does not match the article language.

Similarities Between Languages Reflect Societal or Mixed Biases While differences across languages suggest biases in Wikipedia articles, similarities could be a sign of broader societal perceptions. In Figure 4, we subdivide the LGBT corpus by birth-year, grouping the biographies of people born in 20-year intervals (e.g., 1920–1940). All languages contain more negative portrayals for older age groups. The reduction in negativity over time aligns with the history of LGBT rights. For example, the Stonewall riots in 1969 led to the creation of pride marches, and homosexuality was removed as an "illness" in the American Psychiatric Association's diagnostic manual in 1973 (Morris 2019). These societal changes likely facilitated LGBT people born after 1960 achieving more accomplishments than LGBT people

[11] In theory, the matched control set may contain people of other nationalities, but in practice, our matching algorithm was effective in correcting this, e.g. 26/28 of the controls for Russian LGBT people were also Russian.

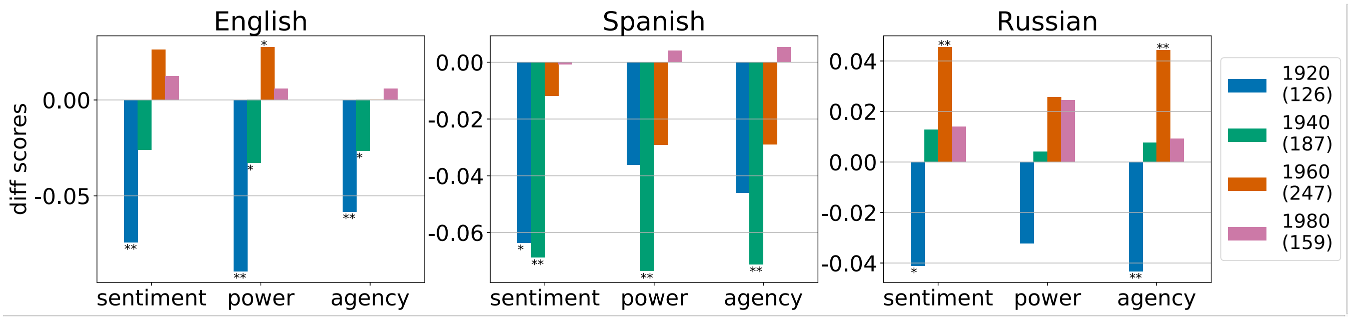

Figure 4: Differences in sentiment/power/agency scores for narratives about age subgroups in LGBTBio.

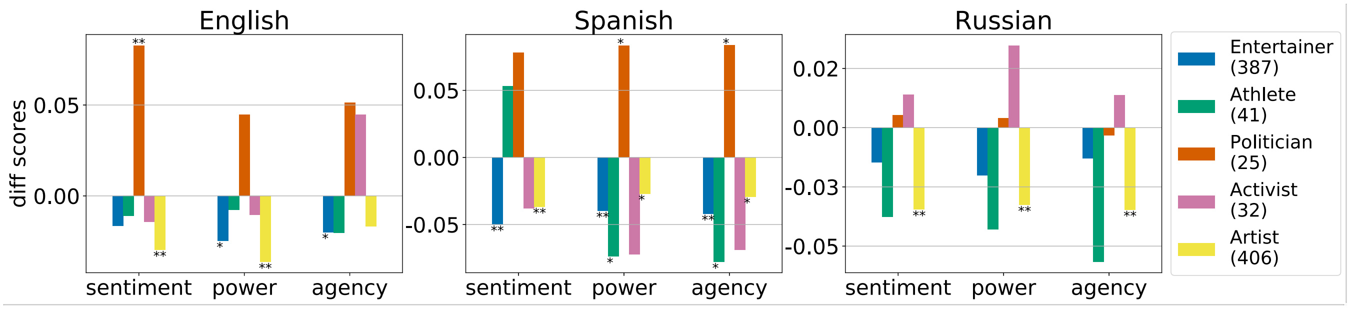

Figure 5: The average diff score of occupation subgroups in the LGBT community across aspects and languages.

born earlier. Because Wikipedia pages contain details of people's lives and accomplishments, we see increasing sentiment, power, and agency in the Wikipedia pages of LGBT people born in different decades. Thus, we find some evidence that differences in portrayal can occur because of biases in society, rather than in Wikipedia pages.

However, for English and Russian, the shift from negative-neutral to positive-neutral occurs for people born after 1960, while in Spanish, the shift is delayed to 1980. To understand this trend, we examined pages of people born in 1960–80 that had the greatest disparities according to our model, e.g., pages that had negative connotations in Spanish but positive ones in Russian and English as compared to controls. The Spanish pages typically listed only brief biographical information, while the English and Russian pages also detailed accomplishments like winning awards (e.g., Jeremy Podeswa and Nilo Cruz). Notably, omissions were more prominent for LGBT people born in 1960–1980 than for their matched controls. Thus, these Spanish articles could benefit from editing to reduce omissions.

Finally, in Figure 5, we subdivide the LGBT corpus by occupation. In all languages, artists have significantly negative portrayals for all connotations dimensions. In contrast, politicians are portrayed more positively than matched controls in both English and Spanish. In examining the most positive articles, we identified sentences that show how LGBT people often receive positive connotations as elected officials. For example, this sentence about Kate Brown, governor of Oregon: *As an openly bisexual woman, Brown has made history several times through electoral success* is scored with positive power, agency, and sentiment. By using multilingual connotation frames to compare across languages, we identify possible biases in Wikipedia that can be addressed through article revisions, and reveal trends reflective of societal changes.

7 Related Work

Our work follows on a series of prior work: Rashkin, Singh, and Choi (2016) introduced sentiment connotation frames, Sap et al. (2017) extended them to power and agency, and Field, Bhat, and Tsvetkov (2019) introduced the CAA framework. Connotation frames have been used to analyze films, newspaper articles, and online stories (Rashkin, Singh, and Choi 2016; Sap et al. 2017; Field, Bhat, and Tsvetkov 2019; Antoniak, Mimno, and Levy 2019). Rashkin et al. (2017) do extend connotation frames to other languages through mapped embeddings, but they do not conduct evaluations against in-language annotations nor provide multilingual annotations.

Our work is generally consistent with existing literature on cross-cultural biases and online biographies. Dong et al. (2019) show that perceptions of social roles differ across cultures, while De-Arteaga et al. (2019) reveal gender bias in online biographies. Other work has examined biases in Wikipedia. Wagner et al. (2015) show that portrayals of men and women differ across languages, and Callahan and Herring (2011) reveal systematic cultural biases, particularly in biography pages.

Several studies in social science literature have analyzed biases and their effects on the LGBTQIA+ community, for example, examining mental health (Almeida et al. 2009) microaggressions (Balsam et al. 2011), and sociopolitical involvement (Harris et al. 2013). With a few exceptions (Schmidt and Wiegand 2017; Fast and Horvitz 2016; Dinakar et al. 2012; Mendelsohn, Tsvetkov, and Jurafsky 2020), biased language about/against LGBTQIA+ community has not been examined and analyzed extensively in automated analyses. The closest study to ours is an examination of gender, race, and LGBT portrayals in 700 popular films (Smith et al. 2015).

8 Conclusion

Our work provides methodology and datasets that extend the capabilities of affective analysis to multilingual settings. While we focus on Wikipedia, our methodology could be used to conduct analyses in any English, Russian, and Spanish narrative text, which can aid writers in obtaining a neutral point of view and provide insight into social stereotypes, especially when used in combination with other methods. This framework supports the investigation of a wide range of research questions, and offers multiple avenues for future work such as improving the multilingual model, expansion to additional languages, investigation of Wikipedia edit histories, and the incorporation of additional connotations and existing linguistic databases.

A Multilingual Annotation Collection

We provide additional details of data collection to facilitate reproduciblity and fully describe data quality.

Language Choice All annotations were crowdsourced through the Figure 8 platform.[12] Our original target languages for this task were English, Russian, Mandarin Chinese, and French. We constructed three rounds of in-house pilot studies for English and one round for Chinese before we released the annotation tasks. After releasing these tasks, we received almost no annotations in French or Chinese, despite increasing payment, expanding the number of target countries, and relaunching tasks. We ultimately dropped Chinese and French in favor of Spanish, for which we were able to obtain annotations. The English and Spanish annotation tasks both finished within 72 hours of being launched. The Russian tasks took several weeks, involving multiple re-launches and pay increases.

Instructions Task instructions were originally written in English and then translated to other languages by native speakers. For each language, a second native speaker checked the translation. The same native speakers also examined the data samples to be annotated, in order to ensure that contexts were grammatical and representative. Heuristics for generating samples were revised according to their feedback.

While power and sentiment are generally well-known terms, agency is unfamiliar to most people and can be difficult to define. Additionally, our Russian, Chinese, and French translators determined that there is no single-word translation for "agency" in these languages, and they instead used combinations of other words to define the concept. Thus, in the annotation task, following Sap et al. (2017) we provided three agency "priming questions" to the annotators.

Task Settings We placed several restrictions on the pool of annotators in order to ensure annotation quality.

Each annotation task included five to eight examples in the task instructions, which were then turned into "quiz questions". Annotators needed to answer an initial eight quiz questions with 70% accuracy to begin the task. As the task proceeded, an additional 10 quiz questions were interspersed with examples to be annotated, and annotators needed to maintain a 70% score on these questions in order to continue the task. We made our questions extremely similar to the examples given in the task instructions, because affective connotations can be subjective, and it is difficult to construct quiz questions that we can expect all high-quality annotators to consistently answer correctly. Instead, our quiz questions ensured that annotators read and understood the task instructions.

We released data in batches of 100-200 annotation examples, as we found that larger batch sizes did not complete. Small batches also allowed us to block annotators who failed quiz questions in earlier batches from attempting to complete later batches. Additionally, we disabled the chrome translation plugin for all non-English tasks.

For Spanish, we restricted annotators to people from nine South America countries: Argentina, Bolivia, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay and Venezuela. We restricted English annotators to the United States and Russian annotators to Russia. Payment for annotation tasks was set based on the time taken to complete the task during pilot studies and the minimum wage of target countries. We also adjusted pay based on survey feedback from early batches. The final rates are in Table 6. In general we paid agency substantially higher than the other two tasks because we had three additional priming questions that annotators need to annotate for each instance.

Task	English	Russian	Spanish
Power	20	4	5
Agency	40	8	8
Sent	20	6	5

Table 6: Task pay rates in cents per five instances

Finally, as mentioned in §2, we screened out annotations from lower-quality annotators after data collection. For each annotator, we computed how often that annotator judged an instance differently than the other two annotators who judged that instance. We then removed annotations from any annotators whose disagreement rate was greater than one standard deviation away from the mean disagreement rate. In our final dataset, we keep only instances that have at least two judgements after removing these annotators. Treftab:agreement shows the full agreement scores for each annotation task after post-processing and Table 8 shows the number of annotated instances for each language.

[12] Figure 8 is now called Appen (https://appen.com/)

	English	Russian	Spanish
Power	0.27	0.33	0.25
Agency	0.20	0.23	0.24
Sent(subj)	0.20	0.27	0.22
Sent(obj)	0.22	0.39	0.31

Table 7: Krippendorff's Alpha per task, after post-processing.

	English	Russian	Spanish
Power	837	880	877
Agency	888	879	888
Sent(subj)	860	868	808
Sent(obj)	860	868	808

Table 8: Number of annotated instances, after post-processing.

B Additional Information for Reproducibility

For the computing infrastructure, we used a single GPU server to extract XLM embeddings of verbs. For 900 contexts, it took about a minute to extract each verb embedding. It took three seconds to train one single logistic regression classification model, which has 3,075 parameters. We did a grid search over the weight given for each class; We binned the range $[0, 1]$ into 20 bins for two classes, which resulted in 400 trials. We chose the final weights based on the validation set F1 score. , and reported the test set performance in the paper. It took about two hours to process the 800k sentences in LGBTBio and label them using the trained classification models. All trained models and their hyperparameter configuration can be found in our codebase under the directory named `saved_models`.

C Classifier Evaluation

Table 9 validates that the performance of our XLM-based model is comparable to prior work by evaluating our model on the same data used in Rashkin, Singh, and Choi (2016) and Field, Bhat, and Tsvetkov (2019). In this setting where we evaluate our model on uncontextualized annotations, we decontextualize the XLM embeddings in the same way as Field, Bhat, and Tsvetkov (2019). Thus, the primary difference between our model and theirs is the use of XLM embeddings instead of ELMo embeddings. Using the exact same training, validation, and test split as prior work, our model performs similarly with Rashkin, Singh, and Choi (2016). Although our model is slightly worse than Field, Bhat, and Tsvetkov (2019), this drop is not surprising and considered a cost of making our model language agnostic.

	Sent$_{subj}$	**Sent$_{obj}$**	**Power**	**Agency**
Majority baseline	26.2	28.7	27.4	29.5
Rashkin, Singh, and Choi (2016)	66.6	37.4	51.8	46.5
Field, Bhat, and Tsvetkov (2019)	61.1	40.4	56.0	48.8
Our model	54.6	45.0	47.4	45.0

Table 9: Classification evaluation results of baselines and our model in macro F1 score. Majority baseline always outputs the most frequent label in the data. Numbers in the Field, Bhat, and Tsvetkov (2019) row are directly borrowed from the original paper.

D LGBTBio Corpus Construction

As described in §6.1, we collected Wikipedia biography pages about LGBT people using lists of LGBT people[13] from Wikipedia. We removed people who do not have pages in all target languages (English, Spanish, Russian) and who have < 2 sentences to be analyzed in any language. We define sentences to analyzed as ones containing the person's name or pronoun as a subject or object. We automatically inferred pronouns based on whether "he" or "she" is more frequent in the article text, which we expect to be effective even for transgender people because of Wikipedia's Manual of Style/Gender identity . After filtering, the LGBT-half of the LGBTBio corpus contains $1,044$ biography pages (L: 174; G: 570, B: 254 T: 43).

We identify matched control biography pages using the algorithm from Anonymous (2020). This algorithm using TF-IDF vectors constructed from biography page categories as matching features. The vectors contain a pivot-slope correction term, which is intended to prevent the method from favoring pages with fewer categories (Singhal, Buckley, and Mitra 2017). Following the recommendations in Anonymous (2020), we set the pivot to the average number of categories per article in our data set. We then tune the slope until the LGBT-half and the matched controls have the same number of average categories (excluding LGBT-specific categories). We try pivot values in 0.1 increments between [0, 0.5], and fix the pivot as 0.1. In Table 10 we provide examples of pairs constructed by this method.

LGBT (Non-LGBT pair)	Common Categories (sampled three)
Tim Cook (Steve Jobs)	*Apple_Inc._executives* *American_computer_businesspeople* *21st-century_American_businesspeople*
Plato (Aristotle)	*4th-century_BC_philosophers* *Ancient_Greek_political_philosophers* *4th-century_BC_writers*
Lily Allen (Dua Lipa)	*English_female_singer-songwriters* *Brit_Award_winners* *Electropop_musicians*
Tom Ford (Anna Sui)	*American_fashion_businesspeople* *Luxury_brands* *Parsons_School_of_Design_alumni*

Table 10: A sampled list of paired-people from our LGBT and non-LGBT corpora.

References

Almeida, J.; Johnson, R. M.; Corliss, H. L.; Molnar, B. E.; and Azrael, D. 2009. Emotional distress among lgbt youth: The in-

[13] https://en.wikipedia.org/wiki/List_of_gay,_lesbian_or_bisexual_people

https://en.wikipedia.org/wiki/List_of_transgender_people

fluence of perceived discrimination based on sexual orientation. *Journal of Youth and Adolescence* 38(7):1001–1014.

Anonymous. 2020. An algorithm for controlled text analysis on wikipedia. anonymous preprint.

Antoniak, M.; Mimno, D.; and Levy, K. 2019. Narrative paths and negotiation of power in birth stories. volume 3, 88. ACM.

Balsam, K. F.; Molina, Y.; Beadnell, B.; Simoni, J.; and Walters, K. 2011. Measuring multiple minority stress: The lgbt people of color microaggressions scale. *Cultur Divers Ethnic Minor Psychol.* 17(2):163–174.

Bamman, D.; O'Connor, B.; and Smith, N. A. 2013. Learning latent personas of film characters. In *Proc. of ACL*, 352–361. Sofia, Bulgaria: Association for Computational Linguistics.

Barrault, L.; Bojar, O.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Koehn, P.; Malmasi, S.; Monz, C.; Müller, M.; Pal, S.; Post, M.; and Zampieri, M. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proc. of WMT*, 1–61. Florence, Italy: Association for Computational Linguistics.

Callahan, E. S., and Herring, S. C. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology* 62(10):1899–1915.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):31.

Conneau, A., and Lample, G. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 7057–7067.

De-Arteaga, M.; Romanov, A.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proc. of FAT*, 120–128. ACM.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Doi, K., and Stewart, P. 2019. Interview: The invisible struggle of japan's transgender population. *Human Rights Watch*.

Dong, M.; Jurgens, D.; Banea, C.; and Mihalcea, R. 2019. Perceptions of social roles across cultures. In *International Conference on Social Informatics*, 157–172. Springer.

Eckert, P. 2000. *Language variation as social practice: The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Eom, Y.-H.; Aragón, P.; Laniado, D.; Kaltenbrunner, A.; Vigna, S.; and Shepelyansky, D. L. 2015. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. *PloS one* 10(3).

Fast, E., and Horvitz, E. 2016. Identifying dogmatism in social media: Signals and models. In *Proc. of EMNLP*, 690–699. Austin, Texas: Association for Computational Linguistics.

Fellbaum, C. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Field, A., and Tsvetkov, Y. 2019. Entity-centric contextual affective analysis. In *Proc. of ACL*, 2550–2560. Florence, Italy: Association for Computational Linguistics.

Field, A.; Bhat, G.; and Tsvetkov, Y. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *Proc. of ICSWM 2019*.

Fournier, M. A.; Moskowitz, D.; and Zuroff, D. C. 2002. Social rank strategies in hierarchical relationships. *Journal of Personality and Social Psychology* 83(2):425.

Hall, J. A., and Braunwald, K. G. 1981. Gender cues in conversations. *Journal of Personality and Social Psychology* 40(1):99.

Harris, A.; Battle, J.; Pastrana, Antonio (., J.; and Daniels, J. 2013. The sociopolitical involvement of black, latino, and asian/pacific islander gay and bisexual men. *Journal of Men's Studies* 21(3):236–254.

Lisitza, A. 2017. History of pride parades in the u.s. *TeenVogue*.

Mendelsohn, J.; Tsvetkov, Y.; and Jurafsky, D. 2020. A framework for the computational linguistic analysis of dehumanization. *Front. Artif. Intell.*

Miller, G. A. 1995. Wordnet: A lexical database for english. *Commun. ACM* 38(11):39–41.

Mohammad, S. M., and Turney, P. D. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3):436–465.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proc. of ACL*, 174–184. Melbourne, Australia: Association for Computational Linguistics.

Morris, B. J. 2019. History of lesbian, gay, bisexual and transgender social movements. *American Psychological Association*.

Rashkin, H.; Bell, E.; Choi, Y.; and Volkova, S. 2017. Multilingual connotation frames: A case study on social media for targeted sentiment analysis and forecast. In *Proc. of ACL*, 459–464. Vancouver, Canada: Association for Computational Linguistics.

Rashkin, H.; Singh, S.; and Choi, Y. 2016. Connotation frames: A data-driven investigation. In *Proc. of ACL*, 311–321. Berlin, Germany: Association for Computational Linguistics.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proc. of ACL*, 1650–1659. Sofia, Bulgaria: Association for Computational Linguistics.

Sap, M.; Prasettio, M. C.; Holtzman, A.; Rashkin, H.; and Choi, Y. 2017. Connotation frames of power and agency in modern films. In *Proc. of EMNLP*, 2329–2334. Copenhagen, Denmark: Association for Computational Linguistics.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proc. of SocialNLP*, 1–10. Valencia, Spain: Association for Computational Linguistics.

Singhal, A.; Buckley, C.; and Mitra, M. 2017. Pivoted document length normalization. In *ACM SIGIR Forum*, volume 51, 176–184. ACM New York, NY, USA.

Smith, S.; Choueiti, M.; Pieper, K.; Gillig, T.; Lee, C.; and DeLuca, D. 2015. Inequality in 700 popular films: Examining portrayals of gender, race, & LGBT status from 2007 to 2014.

Stuart, E. 2010. Matching methods for causal inference: A review and a look forward. *Stat Sci* 25(1):1–21.

Tannen, D. 1994. *Gender and discourse*. Oxford University Press.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*.