A quantitative study of gender representation and authors' gender in a large-market print medium*

Christoph Bartl[1], Sharwin Rezagholi[2,*] and Mareike Schumacher[3]

[1]Department Computer Science, University of Applied Sciences Technikum Wien, Austria
[2]Department Computer Science, University of Applied Sciences Technikum Wien, Austria
[3]Institute of Literary Studies, University of Stuttgart, Germany

Abstract

We analyse gender representation in articles published by the Austrian daily newspaper 'Der Standard' in the years 2021 and 2022. We use named entity recognition and automated gender classification of first names to count the number of female and male persons in articles. The analysis reveals the dominance of male persons in article content. We find that female authors exhibit a significantly higher tendency to mention female persons in their articles.

Keywords

Gender of journalists, gender representation in articles, print newspaper content

1. Introduction

This paper asks whether (i) female persons are less likely to be mentioned by newspaper journalists than male persons, and (ii) whether the propensity to mention female persons differs between female and male journalists. To answer these questions we analyse the entire article output of the Austrian daily newspaper *Der Standard* (https://www.derstandard.at) from the years 2021 and 2022. These articles are publicly available online. We additionally obtained the full names of the authors of all articles from *Der Standard*; this authorship information is non-public for most of the articles. *Der Standard* is the fourth-largest daily newspaper in Austria, having more than 500,000 daily readers [29].

This study employs a binary notion of gender, as does the language policy of *Der Standard*, which prescribes the sole use of feminine and masculine pronouns, effectively prohibiting the use of neopronouns. Therefore this study does not contribute to the abolishment of a binary notion of gender, an aim prominently championed within the digital humanities by Laura Mandell [20].

We employ pretrained models for natural language processing to automatically identify and enumerate female and male persons mentioned in article texts. In particular, we use named entity recognition and automated gender-assignment to first names. The respective methods

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
christoph.bartl@me.com (C. Bartl); sharwin.rezagholi@technikum-wien.at (S. Rezagholi); mareike.schumacher@ur.de (M. Schumacher)
0000-0003-1090-0240 (S. Rezagholi); 0000-0002-7952-4194 (M. Schumacher)

are susceptible to bias. We therefore estimate the error profile of the gender assignment to first names and verify that there is no significant bias that might have distorted our analysis.

We statistically estimate the probability that a journalist mentions a female person when mentioning a person. We find that females are less likely to be mentioned by journalists. This finding holds for male as well as female authors, but we find it to be less pronounced in female journalists. These effects are present to different degrees in different editorial departments, with some departments not exhibiting the imbalance at all. The causal pathways leading to these effects could include statistical 'self-selection', whereby female authors and male authors have a differing propensity to report on certain issues, or female authors could tend to highlight the roles of female persons more than their male colleagues, even when reporting on the same issue.

Literature review. The issue of gender representation and gender inequality is lively discussed in the digital humanities, including computational linguistics [12], digital film studies [3, 32], game studies [31], and computational literary studies [4, 5, 8, 13, 19, 30, 32]. Newspapers and magazines in particular have been studied with respect to gender representation and possible gender bias. Yun et al. found that women were given more space in online than in print journals and that, in a significant fraction of cases, women were portrayed in stereotypical ways [16]. Kian et al. analysed tennis news, finding that female reporters did not write more often about female athletes than their male colleagues but that female reporters tended to use more stereotypical descriptions [17]. Kozlowski et al. [18] analyse the magazines from an Argentinian publisher from 2008 to 2018 using topic modelling and find that the prevalence of thematic areas differs between magazines that target female readers and those that target male ones. They find that this gap is diminishing with respect to certain topics, such as 'family' and 'children', whereas it remains large in others, such as 'fashion' and 'horoscope'. The large-scale analysis of Shor et al. [27], in which more than 20,000 prominent personalities of male and female gender and different (but matching) professions were searched in about 2,000 English-language newspapers, came to the conclusion that the reduced media coverage of women is not in line with the readers' interest, which does not favour prominent men over prominent women. They thus provide some evidence in favour of the hypothesis that newspapers and magazines foster stereotypes and gender bias.

The work most related to ours is due to Mateos de Cabo et al. [21], whose analysis of Spanish online newspapers found that females were more likely to be mentioned in female-authored articles, and to Shor et al. [28], whose analysis of about 2,000 news sources found that the fraction of females in articles increased from 19% in 1983 to 27% in 2008. The latter also found significant differences between editorial departments.

The digital humanities community has formulated a need to further the application of its methods to questions of gender. These voices include Miriam Posner [23], who criticized that gender-related work in the digital humanities does not receive sufficient attention, neither from the scholarly community nor from news outlets. In 2018 Susan Brown stated that a feminist perspective is largely lacking in the digital humanities, going as far as calling 'feminism' the 'f word', suggesting that feminist approaches are effectively silenced [6]. In 2019 Laura Mandell argued that studies on gender within the digital humanities would rather reproduce stereotypes

than analyse them [20]. Coining the term 'data feminism' in 2020, D'Ignazio and Klein drew awareness to gender representation biases in the digital humanities and in other data-driven fields [10]. Although gender bias has been studied in various domains [9, 25, 15, 30, 19, 11, 24, 12], we agree that a sufficient corpus of statistical results remains absent.

2. Data and methods

We start with the text and metadata of 87,032 articles, corresponding to the entire journalistic output of *Der Standard* between January 1, 2021 and December 31, 2022. The metadata of the articles includes authorship, publication date, title, and editorial department. We removed all articles from consideration that were written by a group of journalists or signed by a press agency. This restriction is in line with our interest in the behaviour of individual writers at a newspaper, as opposed to press agencies, group authorship, or anonymous authorship. We retain 36,204 articles.

Named entity recognition and gender-assignment. We use the Python package GenderGuesser 0.4.0 to assign gender to the authors' given names. GenderGuesser uses a database of 40,000 gender-assigned first names to assign gender to a given name [2]. Since some given names, such as 'Andrea', 'Maria' or 'Robin', are not gender-exclusive, GenderGuesser returns 'mostly female', 'mostly male', and 'androgynous' in some cases. We assigned the first two of these categories to 'female' and 'male' respectively. We manually checked the gender-assignments of all 1,571 authors and corrected three cases. We therefore treat the assignment of gender to the names of authors as certainly correct in the remainder of our analysis.

We used the Python package Flair 0.12.2 [1] to recognize personal names in the article texts using the named entity recognition model 'ner-german-large' [26]. In a first step we restrict attention to names consisting of a given name and a family name since the editorial policy of *Der Standard* prescribes the use of the full name at least once per article. The gender of the detected given names was then determined using GenderGuesser. Names such as 'Barack Obama', 'Viktor F.', 'Angela Merkel', 'Nina H.', and 'Luke Skywalker' were identified. In a second step we parse the articles for mentions of the identified persons using only a part of the full name. To clarify our counting method: An article mentioning two persons, the same male person 9 times and a female person once, is considered a text where 90% of mentioned persons are male.

We evaluated our extraction of full names in comparison to the manual counting of the number of female and male full names in 200 randomly selected articles (Table 2). We use binomial estimates to quantify the conditional accuracies of the automated extraction (Table 1). The fairness criterion of predictive parity requires the equality of the true positive rate for male and female cases [7]. We find that these two rates are numerically very similar, approximately 0.87 and 0.88, and that the hypothesis that they are equal can not be rejected (Binomial proportions test, $p = 0.68$).

Generative model. Our main interest is the probability that an author, when mentioning a person, does mention a female person. Our task is complicated by the fact that the automated

Table 1
Performance of gendered name recognition.

	Estimate	St. dev.	95% interval (Wald)
ℙ(classified as male|male)	0.869	0.013	[0.844, 0.893]
ℙ(classified as female|female)	0.881	0.026	[0.831, 0.931]
ℙ(classified as female|male)	0.003	0.002	[0.000, 0.007]
ℙ(classified as male|female)	0.019	0.010	[0.000, 0.040]
ℙ(not detected|male)	0.129	0.013	[0.104, 0.153]
ℙ(not detected|female)	0.100	0.024	[0.054, 0.146]

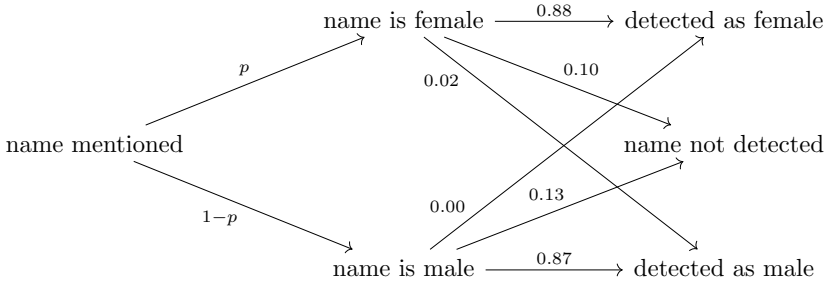

Figure 1: Generative model. Arrows are labelled with estimated conditional probabilities computed on the basis of the manually labelled test set (Table 2).

detection of full names and the automated assignment of gender to the respective first names could be biased. The probability that we have access to is

$$\mathbb{P}(\text{detected as female}|\text{name detected}).$$

We consider a simple generative model for our data (Figure 1), which illustrates that the events 'detected as female' and 'detected as male' do not allow the identification of the parameter of interest, that is p, unless certain assumptions are made. The two necessary assumptions are the absence of mix-ups, that is

$$\mathbb{P}(\text{detected as female}|\text{male}) = \mathbb{P}(\text{detected as male}|\text{female}) = 0, \tag{1}$$

and unbiased non-detection, that is

$$\mathbb{P}(\text{not detected}|\text{female}) = \mathbb{P}(\text{not detected}|\text{male}). \tag{2}$$

If Equations 1 and 2 hold, then

$$\mathbb{P}(\text{fem. name detected}|\text{name detected}) = \mathbb{P}(\text{fem. name mentioned}|\text{name mentioned}).$$

Table 2
Contingency table for gendered name recognition.

	Classified as male	Classified as female	Not recognized	Total
Male	621	2	92	715
Female	3	141	16	160
Total	624	143	108	

It is not necessary to assume that non-detection does not occur, but that non-detection is unbiased. The empirical probabilities corresponding to those in Equation 2 are similar, approximately 0.10 and 0.13, and the hypothesis that they are equal can not be rejected (Binomial proportions test, $p = 0.31$). As Table 1 reports, the empirical probabilities for mix-ups equal approximately 0.00 and 0.02. We feel that these values are sufficiently low to assume that Equation 1 holds. In Table 2 the fraction of female persons among the mentioned persons equals $160/(160 + 715) \approx 0.18$ while the fraction of names classified as female among the classified names equals $143/(143 + 624) \approx 0.19$. This numerically illustrates the absence of bias.

Statistical model. The details of our statistical approach are presented in Appendix A. Our basic modelling assumption is that the number of persons in an article is predetermined, but the respective journalist 'chooses' the gender of the mentioned persons independently from a Bernoulli distribution. The parameter of this Bernoulli distribution is specific to the subset of the data for which an estimate is desired. We believe that this model is sufficient to organise the data in a tractable and intuitive fashion. To be concrete: We estimate the probability that a journalist uses the name of a female person when using the name of a person. Note that our estimate does not equal the fraction of female persons in a certain subset of articles. Our estimate is descriptive of authors' behaviour, not of article output (see Appendix A). It is important to note that every author, regardless of the number of persons mentioned in their respective articles, has equal importance for our point estimates, but the differing degrees of uncertainty for different authors, caused by the different quantities of mentioned persons, are reflected in the interval estimates, that is in the confidence intervals. When we estimate probabilities specific to editorial departments, we employ a weighting scheme whereby the degree of membership of an author in an editorial department is taken into account. Consult Appendix A for details on our statistical approach.

3. Descriptive analysis

Of the 36,204 articles, 12,736 (35%) were written by a female author and 23,468 (65%) were written by a male author. Among the 1,571 unique authors 606 (39%) are female. The average author has contributed 23 articles to the dataset. The articles are of varying length, the median length being 3,859 characters (1st quartile = 2,560, 3rd quartile = 5,248).

The mean of the fraction of females in an article equals 25%. The respective mean for the articles written by male writers equals 20%. On the other hand, the mean for the articles authored

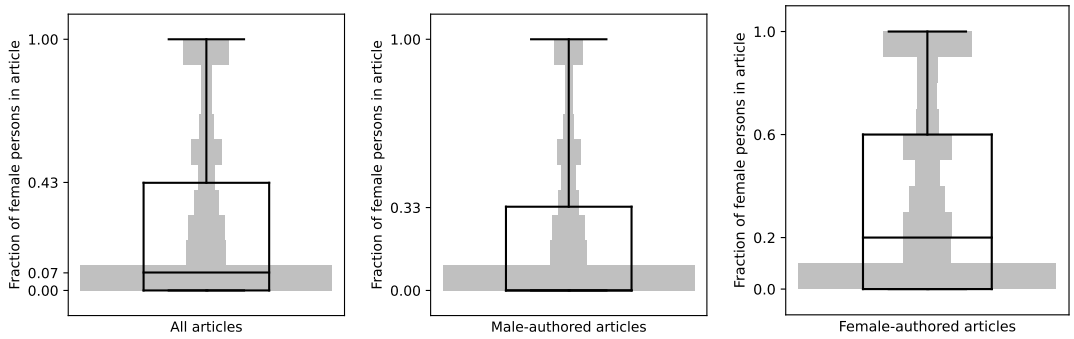

Figure 2: Empirical distributions of the fraction of females in an article, quartiles on the vertical axes.

by female writers equals 33%. The empirical distributions of the fraction of female persons in an article are visualized in Figure 2. Note that we only consider articles that do mention at least one person. It is apparent that many articles that do mention persons do not mention females at all. This is true for female- as well as male-authored articles. It is evident that the distributions for female- and male-authored articles differ. The distribution for female-authored articles stochastically dominates the distribution for male-authored articles (McFadden's test [22], $p = 0.00$). The distributions exhibit concentrations at the extremes. This illustrates that many articles solely mention persons of one gender.

Differentiating with respect to the editorial departments of *Der Standard* (Table 3), one finds that the department Family produces the smallest number of articles but has the largest fraction of female authorship. The largest number of articles is produced by the department Culture, which corresponds to roughly 13% of our data.

4. Statistical estimation

Our first interests are to estimate the probabilities that (i) an author who mentions a person does mention a female person, (ii) a female author who mentions a person does mention a female person, and (iii) a male author who mentions a person does mention a female person. The respective estimated probabilities are reported in Table 4 and visualised in Figure 3. While the probability that an author mentions a female person when mentioning a person is estimated to be roughly 27%, the respective estimate for female authors is roughly 46% and roughly 14% for male authors. The differences are highly statistically significant ($p = 0.00$). We conclude that there is strong evidence, at least in our data, that female journalists are more likely to mention female persons in their articles compared to male journalists.

To elucidate the differences between authors from different editorial departments, we estimate the probability that a journalist from a given department mentions a female person when mentioning a person. These estimates are reported in Table 5 and visualized in Figure 4. There are editorial departments whose output is likely to mention females, such as Female *Standard* and Family, and departments whose writers are unlikely to mention females, such as Automo-

Table 3
Data composition and author's gender in terms of editorial departments.

Department	Article count	Article percent	Female authorship count	Fraction female	Unique authors	Unique female authors	Fraction female
Wirtschaft (Economy)	3131	8.65	1462	0.47	134	44	0.33
Karriere (Career)	724	2.00	584	0.81	78	40	0.51
Recht (Law)	704	1.94	148	0.21	131	36	0.27
Die Standard (Female *Standard*)	367	1.01	335	0.91	54	46	0.85
Gesundheit (Health)	657	1.81	603	0.92	47	30	0.64
Automobil (Automobile)	644	1.78	44	0.07	35	9	0.26
Web	3734	10.31	44	0.01	46	9	0.20
Reisen (Travel)	319	0.88	140	0.44	52	23	0.44
Meinung (Opinion)	4205	11.61	1472	0.35	692	226	0.33
International (International)	3812	10.53	1360	0.36	150	54	0.36
Inland (Domestic)	1972	5.45	638	0.32	88	41	0.47
Lifestyle	1588	4.39	833	0.52	112	60	0.54
Etat (Government)	1902	5.25	716	0.38	98	37	0.38
Immobilien (Realty)	853	2.36	364	0.43	39	17	0.44
Bildung (Education)	466	1.29	335	0.72	64	38	0.59
Zukunft (Future)	628	1.73	145	0.23	44	25	0.57
Sport	1375	3.80	28	0.02	61	14	0.23
Familie (Family)	238	0.66	230	0.97	32	25	0.78
Kultur (Culture)	4625	12.77	1691	0.37	294	131	0.45
Panorama	2330	6.44	858	0.37	177	70	0.40
Wissenschaft (Science)	1930	5.33	706	0.37	156	80	0.51

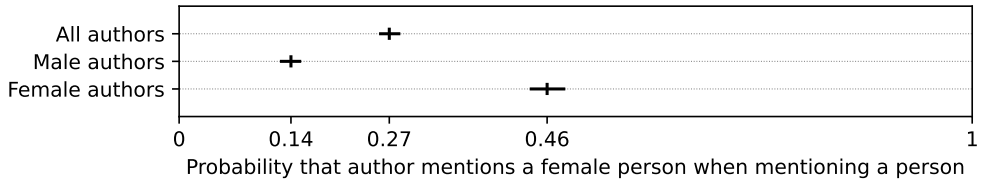

Figure 3: Estimated probabilities that an author, when mentioning a person, does mention a female person (with 95%-confidence intervals).

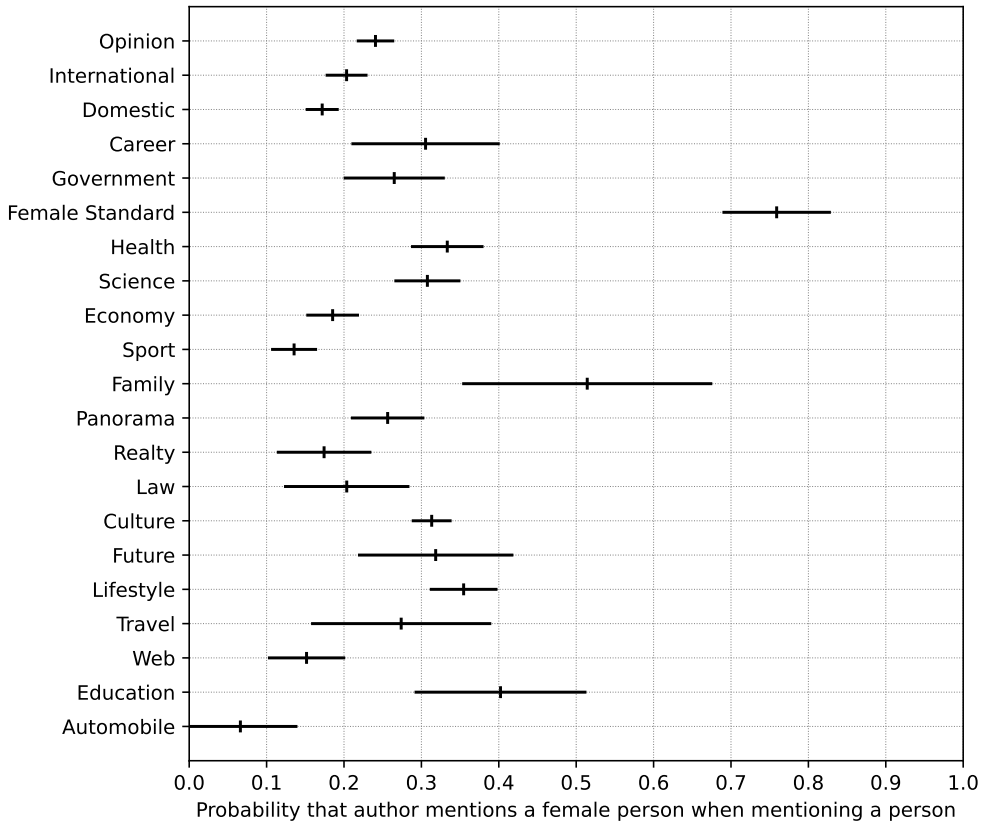

Figure 4: Estimated probability that an author from the respective department mentions a female person when mentioning a person (with 95%-confidence intervals).

bile and Sport. These findings are in line with previous studies from different countries and languages [17, 28, 18].

To disentangle the effect of authors' gender and editorial departments, we stratify our analysis with respect to both. These estimates are visualized in Figure 5 and reported in Table 6, including hypothesis tests for the null-hypothesis that female and male authors behave identically. We obtain significantly different estimates for female and male authors for many edito-

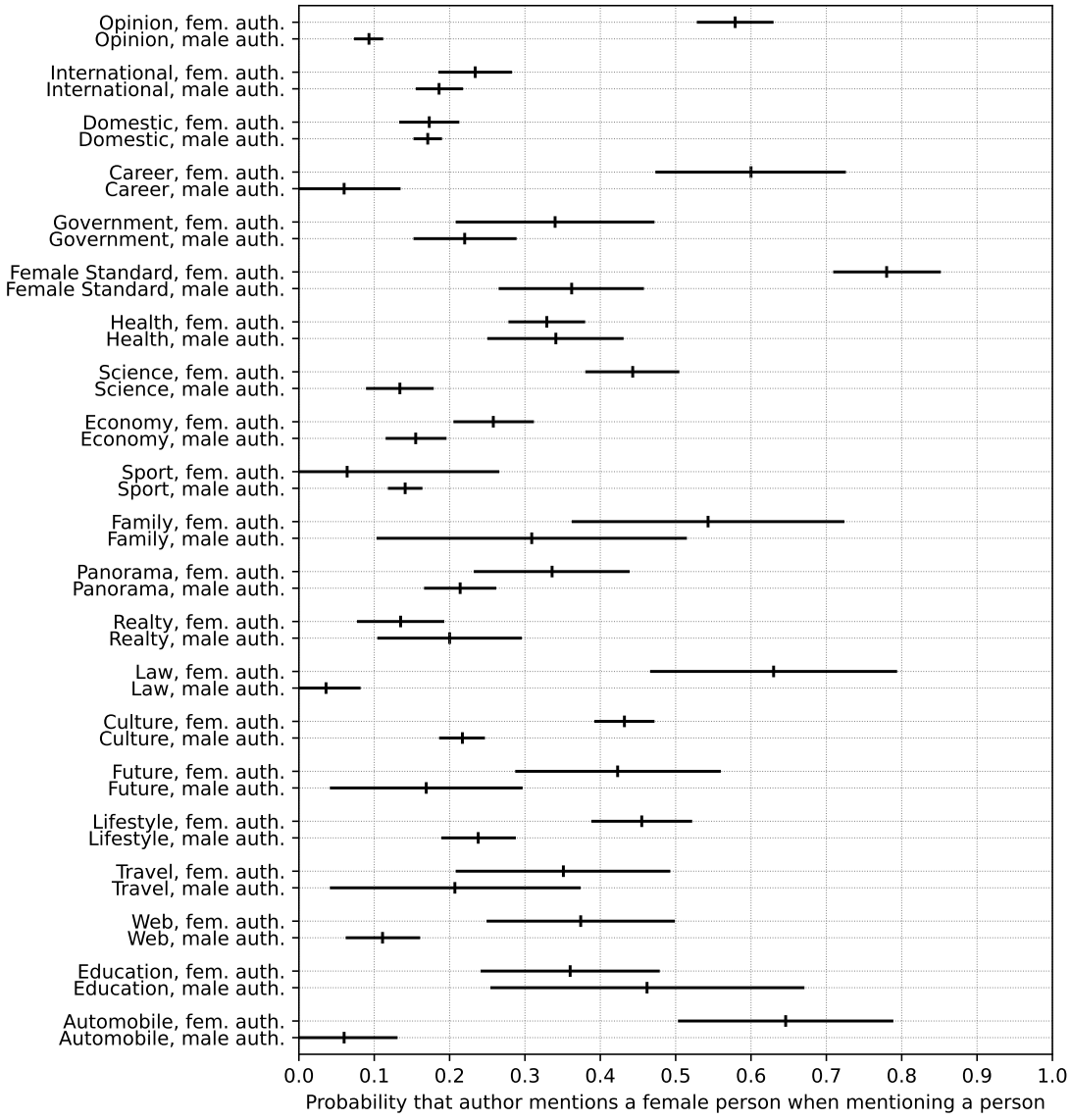

Figure 5: Estimated probability that a journalist from the respective department and of the respective gender mentions a female person when mentioning a person (with 95%-confidence intervals).

rial departments. The departments Opinion, Career, Female *Standard*, Science, Economy, Law, Culture, Lifestyle, Web, and Automobile exhibit highly significant differences with respect to author gender ($p = 0.00$). The departments Domestic, Government, Health, Realty, Travel, and Education do not exhibit statistically significant gender differences ($p > 0.1$).

5. Conclusion and caveats

We present some statistical evidence for the hypothesis that female and male journalists have differing propensities to mention female persons in their writing. This effect varies between editorial departments, at least in our data, and is stronger within certain editorial departments than across editorial departments. Further research is needed to elucidate whether the findings of the present study are driven by a mechanism through which female journalists write about different topics than male journalists. Even if this were true, it would remain ambiguous whether the ultimate cause is the issue-assignment policy within newsrooms, or a desire by journalists to write on topics featuring persons of a certain gender. The latter could also correspond to a conscious attempt by female journalists to highlight female persons in an attempt to counteract existing gender imbalances. We deem it an interesting avenue for further research to quantitatively elucidate the relationship between journalistic topics, gender representation, and authorship. If the observed differences were caused by different propensities in mentioning female persons even when reporting on the same topic, this would indicate gender-specificity in journalist's viewpoints. Finally we want to highlight that the present study is most certainly marred by the problem that many relevant and potentially confounding factors, such as topics, have not yet been taken into account. Therefore the present study is but an empirical quantification of a status quo.

Acknowledgments

The authors thank Martin Kotynek and Werner Weichselberger from *Der Standard* for their support.

References

[1] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. "FLAIR: An easy-to-use framework for state-of-the-art NLP". In: *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2019), pp. 54–59.

[2] D. Arcos. *Gender-guesser*. 2016. URL: https://github.com/lead-ratings/gender-guesser.

[3] D. Bamman, B. O'Connor, and N. A. Smith. "Learning Latent Personas of Film Characters". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2013, pp. 352–361. URL: https://aclanthology.org/P13-1035.

[4] O. Baylog, L. Dimmit, T. Heller, G. Kirilloff, S. Smith, G. Thomas, C. Warren, and J. Wehrwein. *"More than Custom has Pronounced Necessary": Exploring the Correlation between Gendered Verbs and Character in the 19th Century Novel Nebraska Literary Lab*. 2016.

[5] J. Bergenmar and K. Leppänen. "Gender and Vernaculars in Digital Humanities and World Literature". In: *NORA - Nordic Journal of Feminist and Gender Research* 25.4 (2017), pp. 232–246. DOI: 10.1080/08038740.2017.1378256.

[6] S. Brown. "Delivery Service: Gender and the Political Unconscious of Digital Humanities". In: *Bodies of Information: Intersectional Feminism and the Digital Humanities*. University of Minnesota Press, 2018, pp. 261–286.

[7] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12.4209 (2022).

[8] J. Cheng. "Fleshing Out Models of Gender in English-Language Novels (1850–2000)". In: *Journal of Cultural Analytics* 5.1 (2020). DOI: 10.22148/001c.11652.

[9] M. Conroy. "Quantifying the Gap: The Gender Gap in French Writers' Wikidata". In: *Journal of Cultural Analytics* 8.2 (2023). DOI: 10.22148/001c.74068.

[10] C. D'Ignazio and L. F. Klein. *Data Feminism.* MIT Press, 2020.

[11] M. Flüh, J. Horstmann, and M. Schumacher. "Genderaspekte in Fantasy-Jugendromanen von 2008 bis 2020: Distant Gender Reading". In: *Gender in der deutschsprachigen Kinder- und Jugendliteratur*. De Gruyter, 2022, pp. 457–482. DOI: 10.1515/9783110726404-025.

[12] C. Freitas and D. Santos. "Human Depiction in Portuguese. Distant reading Brazilian and Portuguese literature". In: *Journal of Computational Literary Studies* 2 (2024).

[13] S. Hota and S. Argamon. *Performing gender: Automatic stylistic analysis of Shakespeare's characters*. 2006.

[14] F. Hu and J. Zidek. "The weighted likelihood". In: *Canadian Journal of Statistics* 30.3 (2002), pp. 347–371. DOI: https://doi.org/10.2307/3316141.

[15] M. Jockers and G. Kirilloff. "Understanding Gender and Character Agency in the 19th Century Novel". In: *Journal of Cultural Analytics* 2.2 (2016). DOI: 10.22148/16.010.

[16] H. Jung Yun, M. Postelnicu, N. Ramoutar, and L. Lee Kaid. "Where Is She?: Coverage of women in online news magazines". In: *Journalism Studies* 8.6 (2007), pp. 930–947. DOI: 10.1080/14616700701556823.

[17] E. M. Kian, J. S. Fink, and M. Hardin. "Examining the Impact of Journalists' Gender in Online and Newspaper Tennis Articles". In: *Women in Sport and Physical Activity Journal* 20.2 (2011), pp. 3–21. DOI: 10.1123/wspaj.20.2.3.

[18] D. Kozlowski, G. Lozano, C. Felcher, F. Gonzalez, and E. Altszyler. *Gender bias in magazines oriented to men and women: A computational approach*. 2020.

[19] E. Kraicer and A. Piper. "Social Characters: The Hierarchy of Gender in Contemporary English-Language Fiction". In: *Journal of Cultural Analytics* 3.2 (2019). DOI: 10.22148/16.032.

[20] L. Mandell. "Gender and Cultural Analytics: Finding or Making Stereotypes?" In: *Debates in the Digital Humanities 2019*. University of Minnesota Press, 2019, pp. 3–26.

[21] R. Mateos de Cabo, R. Gimeno, M. Martínez, and L. López. "Perpetuating Gender Inequality via the Internet? An Analysis of Women's Presence in Spanish Online Newspapers". In: *Sex Roles* 70.1 (2014), pp. 57–71. DOI: 10.1007/s11199-013-0331-y.

[22] D. McFadden. "Testing for Stochastic Dominance". In: *Studies in the Economics of Uncertainty*. Springer, 1989, pp. 113–134.

[23] M. Posner. "What's Next: The Radical, Unrealized Potential of Digital Humanities". In: *Debates in the Digital Humanities 2016*. University of Minnesota Press, 2016, pp. 32–41.

[24] T. Schmidt, I. Engl, J. Herzog, and L. Judisch. "Towards an Analysis of Gender in Video Game Culture: Exploring Gender specific Vocabulary in Video Game Magazines". In: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)* (2020), pp. 333–341.

[25] M. Schumacher and M. Flüh. "Made to Be a Woman: A case study on the categorization of gender using an individuation-based approach in the analysis of literary texts". In: *Digital Humanities Quarterly* 17.3 (2023). URL: https://www.digitalhumanities.org/dhq/vol/17/3/000728/000728.html.

[26] S. Schweter and A. Akbik. "FLERT: Document-level features for named entity recognition". 2020.

[27] E. Shor, A. van de Rijt, and B. Fotouhi. "A Large-Scale Test of Gender Bias in the Media". In: *Sociological Science* 6 (2019), pp. 526–550. DOI: 10.15195/v6.a20.

[28] E. Shor, A. van de Rijt, A. Miltsov, V. Kulkarni, and S. Skiena. "A Paper Ceiling: Explaining the Persistent Underrepresentation of Women in Printed News". In: *American Sociological Review* 80.5 (2015), pp. 960–984. DOI: 10.1177/0003122415596999.

[29] Statista. *Österreich Tageszeitungen nach Anzahl der Leser*. 2022. URL: https://de.statista.com/statistik/%20daten/studie/307114/umfrage/tageszeitungen-in-oesterreich-nach-anzahl-der-leser/.

[30] T. Underwood, D. Bamman, and S. Lee. "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics* 3.2 (2018). DOI: 10.22148/16.019.

[31] T. Unterhuber. *Männlich codiert?: Annäherung an eine Medien- und Geschlechtergeschichte des Computerspiels*. 2021.

[32] E.-M. Venzmer. *"Oh, the [digital] humanities¡' – Eine quantitative Gender-Analyse von The Big Bang Theory*. 2023.

A. Appendix

Maximum likelihood estimate. We consider the weighted likelihood

$$L(p_1, ..., p_l) \propto \prod_{j=1}^{l} \left(\prod_{i \in S_j} \left(p_j^{k_i} (1 - p_j)^{m_i - k_i} \right)^{w_i} \right), \tag{3}$$

where k_i denotes the number of persons detected as female in article i, m_i denotes the number of detected persons in article i, $\{S_1, ..., S_l\}$ are disjoint subsets of the data $\{1, ..., n\}$, and the weight w_i equals the reciprocal of the number of natural persons detected in articles written by the respective journalist a_i, that is

$$w_i = \frac{1}{\sum_{j:a_i=a_j} m_j}. \tag{4}$$

Note that we have discarded multiplicative constants from Equation 3. Weighted likelihood estimation is a well-established method in several circumstances [14]. The likelihood (Equation 3) is maximized at the parameter-values $\{\hat{p}_1, ..., \hat{p}_l\}$ given by

$$\hat{p}_j = \frac{\sum_{i \in S_j} w_i k_i}{\sum_{i \in S_j} w_i m_i}. \tag{5}$$

Under our choice of weighting (Equation 4), the maximum-likelihood estimates according to Equation 5 can be written as

$$\hat{p}_j = \frac{\sum_{a \in A} \frac{k_{a,j}}{m_a}}{\sum_{a \in A} \frac{m_{a,j}}{m_a}} = \frac{1}{\sum_{a \in A} \frac{m_{a,j}}{m_a}} \sum_{a \in A} \frac{m_{a,j}}{m_a} \frac{k_{a,j}}{m_{a,j}},$$

where A denotes the set of unique authors, $m_{a,j}$ denotes the number of persons detected in texts of author a in subset S_j, m_a denotes the number of persons detected in texts of author a, and $k_{a,j}$ denotes the number of female persons detected in articles by author a in subset S_j. This is but the weighted mean of the naive per-author estimates for the subset, that is $k_{a,j}/m_{a,j}$, weighted by the 'degree of membership' of author a in subset S_j, that is by $m_{a,j}/m_a$. Note that this estimator is such that multiplying all data from a certain author by a constant does not change the estimate. In the special case of a single subset equal to the entirety of the data, the estimator takes the form

$$\hat{p} = \frac{1}{|A|} \sum_{a \in A} \left(\frac{\sum_{i:a_i=a} k_i}{\sum_{i:a_i=a} m_i} \right),$$

which is but the arithmetic average of the per-author relative frequencies.

Confidence intervals. The variance of $\hat{p}_j$ equals

$$V(\hat{p}_j) = \frac{1}{\left(\sum_{i \in S_j} w_i m_i \right)^2} \sum_{i \in S_j} w_i^2 V(k_i)$$

where $k_i \sim \text{binomial}(m_i, p_j)$ and hence $V(k_i) = p_j(1 - p_j)m_i$. Therefore the plug-in estimator for the variance of $\hat{p}_j$ is

$$V(\hat{p}_j) \approx \hat{p}_j(1 - \hat{p}_j) \frac{\sum_{i \in S_j} w_i^2 m_i}{\left(\sum_{i \in S_j} w_i m_i \right)^2} \tag{6}$$

This enables us to use a normal approximation to the distribution of $\hat{p}_j$ to construct confidence intervals.

Hypothesis tests. To test null-hypotheses of the form $p_j = p_{j'}$, we construct a test using the test statistic

$$\hat{p}_j - \hat{p}_{j'} \sim N\Big(\hat{p}_j - \hat{p}_{j'}, V(\hat{p}_j) + V(\hat{p}_{j'})\Big),$$

where the variances are computed according to Equation 6.

Table 4
Estimated probabilities that an author, when mentioning a person, does mention a female person.

Model	Article count	Estimate	St. dev.	95% interval
All authors	30,099	0.265	0.006	[0.254,0.277]
Male authors	19,444	0.141	0.006	[0.129,0.152]
Female authors	10,655	0.464	0.011	[0.444,0.485]

Table 5
Estimated department-specific probability that an author mentions a female person when mentioning a person.

Model	Article count	Estimate	St. dev.	95% interval
Opinion	3293	0.241	0.011	[0.218,0.263]
International	3577	0.203	0.013	[0.178,0.229]
Domestic	1934	0.172	0.010	[0.152,0.191]
Career	502	0.305	0.048	[0.211,0.399]
Government	1780	0.265	0.032	[0.202,0.328]
Female *Standard*	312	0.759	0.035	[0.691,0.827]
Health	577	0.333	0.023	[0.289,0.378]
Science	1803	0.308	0.021	[0.267,0.348]
Economy	2858	0.185	0.016	[0.153,0.217]
Sport	1333	0.136	0.014	[0.108,0.163]
Family	139	0.514	0.081	[0.355,0.674]
Panorama	2062	0.256	0.023	[0.211,0.302]
Realty	685	0.174	0.030	[0.115,0.233]
Law	374	0.204	0.040	[0.124,0.283]
Culture	4400	0.313	0.012	[0.289,0.337]
Future	484	0.318	0.050	[0.220,0.417]
Lifestyle	1138	0.355	0.021	[0.313,0.396]
Travel	161	0.274	0.058	[0.159,0.388]
Web	1960	0.152	0.024	[0.104,0.200]
Education	390	0.402	0.056	[0.293,0.511]
Automobile	337	0.066	0.037	[0.000,0.138]

Table 6
Estimated department- and gender-specific probabilities that a journalist mentions a female person when mentioning a person.

Model	Art. ct.	Estimate	St. dev.	95% interval	p-value, $H_0 : p_m = p_f$
Opinion, female authors	1030	0.579	0.025	[0.530,0.628]	0.00
Opinion, male authors	2263	0.093	0.009	[0.075,0.110]	
International, female authors	1279	0.234	0.024	[0.187,0.281]	0.09
International, male authors	2298	0.186	0.015	[0.157,0.216]	
Domestic, female authors	631	0.173	0.019	[0.135,0.211]	0.92
Domestic, male authors	1303	0.171	0.009	[0.154,0.188]	
Career, female authors	398	0.600	0.064	[0.475,0.724]	0.00
Career, male authors	104	0.060	0.037	[0.000,0.133]	
Government, female authors	662	0.340	0.066	[0.210,0.470]	0.11
Government, male authors	1118	0.220	0.034	[0.154,0.287]	
Female *Standard*, female authors	289	0.780	0.035	[0.711,0.850]	0.00
Female *Standard*, male authors	23	0.362	0.048	[0.267,0.456]	
Health, female authors	526	0.329	0.025	[0.280,0.378]	0.82
Health, male authors	51	0.341	0.045	[0.252,0.429]	
Science, female authors	673	0.443	0.031	[0.382,0.503]	0.00
Science, male authors	1130	0.134	0.022	[0.091,0.177]	
Economy, female authors	1332	0.258	0.026	[0.207,0.310]	0.00
Economy, male authors	1526	0.155	0.020	[0.117,0.194]	
Sport, female authors	23	0.064			
Sport, male authors	1310	0.141	0.011	[0.120,0.162]	
Family, female authors	134	0.543	0.091	[0.364,0.722]	
Family, male authors	5	0.309			
Panorama, female authors	766	0.336	0.052	[0.234,0.437]	0.03
Panorama, male authors	1296	0.214	0.023	[0.168,0.260]	
Realty, female authors	271	0.135	0.029	[0.079,0.191]	0.25
Realty, male authors	414	0.200	0.048	[0.106,0.294]	
Law, female authors	76	0.630	0.083	[0.468,0.792]	0.00
Law, male authors	298	0.036	0.023	[0.000,0.080]	
Culture, female authors	1591	0.432	0.019	[0.394,0.470]	0.00
Culture, male authors	2809	0.217	0.015	[0.188,0.245]	
Future, female authors	93	0.423	0.068	[0.289,0.558]	0.01
Future, male authors	391	0.169	0.064	[0.043,0.295]	
Lifestyle, female authors	537	0.455	0.033	[0.390,0.520]	0.00
Lifestyle, male authors	601	0.238	0.024	[0.191,0.286]	
Travel, female authors	40	0.351	0.072	[0.210,0.491]	0.19
Travel, male authors	121	0.207	0.084	[0.043,0.372]	
Web, female authors	21	0.374	0.063	[0.251,0.497]	0.00
Web, male authors	1939	0.111	0.024	[0.064,0.159]	
Education, female authors	268	0.360	0.060	[0.243,0.477]	0.40
Education, male authors	122	0.462	0.105	[0.256,0.669]	
Automobile, female authors	15	0.646	0.072	[0.505,0.787]	0.00
Automobile, male authors	322	0.060	0.035	[0.000,0.129]	