Developing a Multilingual Corpus of Wikipedia Biographies

Hannah Devinney[12], Anton Eklund[1], Igor Ryazanov[1], Jingwen Cai[1]
[1] Umeå University, Department of Computing Science
[2] Umeå Centre for Gender Studies
Umeå, Sweden
[hannahd, antone, igorr, jingwenc]@cs.umu.se

Abstract

For many languages, Wikipedia is the most accessible source of biographical information. Studying how Wikipedia describes the lives of people can provide insights into societal biases, as well as cultural differences more generally. We present a method for extracting datasets of Wikipedia biographies. The accompanying codebase is adapted to English, Swedish, Russian, Chinese, and Farsi, and is extendable to other languages.

We present an exploratory analysis of biographical topics and gendered patterns in four languages using topic modelling and embedding clustering. We find similarities across languages in the types of categories present, with the distribution of biographies concentrated in the language's core regions. Masculine terms are over-represented and spread out over a wide variety of topics. Feminine terms are less frequent and linked to more constrained topics. Non-binary terms are nearly non-represented.

1 Introduction

Wikipedia's decentralised organisation and independent communities in different languages have led it to be considered a 'global repository of knowledge' (Callahan and Herring, 2011). Easily and openly accessible, in many language environments it has displaced more traditional and specialised resources as a 'default' encyclopedic source, shaping the language landscape. This effect is noticeable in NLP research and development, where Wikipedia is a staple training data source for language models that imitate Wikipedia when generating texts or making writing suggestions.

When it comes to biographical information, Wikipedia has become a primary source of reference, especially outside of education systems. Because of the near-monopoly that Wikipedia has on public knowledge in these environments, it can define which persons are perceived as notable, which aspects of their lives deserve a mention, and how the persons are presented. The community guidelines of Wikipedia are built around the concept of 'neutrality', but the content is still inevitably shaped by societal biases, such as gender gaps (Hube, 2017). Besides the content of the biographical articles, Wikipedia also shapes the expectations the reader has in terms of format, language, style and inclusion. A 'biography' can come to invoke a Wikipedia-like structure, becoming a commonly accepted way of summarising a person's life.

Most automatically extracted datasets consists of Wikipedia backup dumps or rely heavily on the connection to Wikidata. Using Wikidata, however, makes it harder to modify or update the dataset, or replicate it for another domain. Manually collected datasets, on the other hand, are limited in size. They are almost inevitably biased towards longer, popular or better-categorised articles because poor categorisation prevents other articles to be discovered in the first place. These problems become even more apparent when creating a multilingual dataset. While different editions share a general article structure and templates, they are far from identical. As we discuss further in the paper, straightforward parsing approaches can fail if adapted directly because of the subtle markdown changes. For our biographical dataset, this often results in the omission of less well-documented (often marginalised) people.

This paper contributes an adaptable method for curating a multilingual corpus of Wikipedia biographies. We analyse general statistics and structures of the biographies for corpora in several languages and compare them with existing literature. Finally, we release the code[1] and instructions on how to create a biography dataset from an up-to-date Wikipedia dump adaptable to any language.

[1] https://github.com/antoneklund/wikipedia-biographies

Proceedings of Recent Advances in Natural Language Processing, pages 285–294
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_032

2 Background

2.1 Biography

We define a *biography* as the running text of an article about an individual person and their life or story (rather than a single event, or a more tailored summary of a one's professional life, i.e. a 'short bio'). We define *persons* as animate individuals, and include in this definition both real people and fictional or mythological figures. This does not actively attempt to include animals, but if the biography of an animal meets all other criteria we do not reject them, as their page is likely to also contain their life or story.

2.2 Related Work

Wikipedia is commonly leveraged as a resource in Natural Language Processing, both for its texts and the associated metadata such as edit history (Botha et al., 2018; Faruqui et al., 2018), infoboxes (Wu and Weld, 2010), and hyperlinks (Gemechu et al., 2016). Its multilingual nature makes it appealing for both cross-lingual (Perez-Beltrachini and Lapata, 2021) and translation-based tasks (Coster and Kauchak, 2011; Drexler et al., 2014).

Wikipedia biographies have been leveraged for summarisation (Gao et al., 2021) and information extraction (Hogue et al., 2014). Palmero Aprosio and Tonelli (2015) train a supervised classifier to recognise sections of Wikipedia entries as biographies. Most recently Stranisci et al. (2023) presented a task for biographical events detection accompanied with an annotated dataset, as well as intersectional analysis of writers' biographies in English Wikipedia. To extend this to other languages, a new classifier would presumably need to be trained for every target Wikipedia.

Due to its near-monopoly on up-to-date biographical information, Wikipedia is a prime resource for biographical bias studies while also allowing for comparative studies between languages (Callahan and Herring, 2011; Wagner et al., 2015; Field et al., 2022). In particular, the Wikipedia gender gap in biographical coverage and representation is well-studied. Women are less likely to write or be written about in Wikipedia articles, and the events focused on biographies of women are more often constrained to the private sphere (Klein and Konieczny, 2015; Fan and Gardent, 2022; Schmahl et al., 2020; Sun and Peng, 2021; Ferran-Ferrer et al., 2022; Wagner et al., 2015). However, there also exists a 'glass-ceiling effect', where women in Wikipedia are more present among longer and more detailed biographies and more notable, suggesting a higher barrier to entry. There is also evidence of women of non-western background being particularly underrepresented in English Wikipedia (Stranisci et al., 2023). Although there is little research covering trans and nonbinary representation in Wikipedia biographies, similar barriers may exist, and there is more of a focus on the subject's gender identity (Field et al., 2022), which is generally unmarked in biographies about cis people.

3 The Corpus

The code for creating the Wikipedia biographies corpora released with this paper is created with the purpose of exploring cultural and narrative trends, including social bias analysis. The Wikipedia articles that are collected should meet the criteria of being a biography (section 2.1). In this section, we describe the process of identifying biographies and extracting clean text; and present a data card with basic corpora statistics.

3.1 Collecting Articles/Biographies

Biographies are identified and extracted using regular expressions (see Appendix A) directly applied to the markdown (source text) of Wikipedia pages which are obtained from a Wikipedia dump. Using only the Wikipedia dump allows reproducing the dataset without incorporating other data. Not relying on Wikidata connections makes it significantly easier to create analogous datasets for other languages with limited curation.

In practice, we identify articles about persons in two ways. Our main approach checks the categories associated with that article. We look for broad category tags such as 'living people' in English as well as those tags listing birth and death years, such as 'födda 1975' (*born in 1975*). Since the markdown differs between languages despite superficially standard categorisation, manual investigation is necessary to decide which tags to use when adapting to a new language.

For languages where not all categories are explicitly listed in the markdown, there is a risk of severe under-capture, and other methods of identification must be used. For instance, in Chinese, the birth year and either the death year or 'living person' categories are in most cases not specified manually like other, non-standard, categories. In-

stead, a special birth and death date markdown element is inserted at the beginning of the article which adds the appropriate categories to the page. So, for the Chinese Wikipedia, we check for, e.g. 'bd|1239年|6月17日|1307年|7月7日|Edward I' (bd|June 6th, 1239|July 7th, 1307|Edward I; pointing to birth and death dates) instead of '1307年逝世' (*died in 1307*).

There are also languages that, assign the *living people*, *born* and *died* categories fully automatically from Wikidata, without any specific mentions in markdown. This cannot be tracked in text. This applies to Russian: the category for all people – 'Персоналии по алфавиту' (*Personae alphabetically*) – as well as the birth and death categories – 'родившиеся в [YEAR] году' (*born in year [YEAR]*) and 'умершие в [YEAR] году' (*died in year [YEAR]*) – are applied from Wikidata based on the page template.

To capture articles in Russian, we scan for non-category elements in the markdown. We look for specific lines in the infobox, e.g 'Дата рождения' (*Date of Birth*), which should be present only for persons. We expect this approach to miss more articles than using categories because shorter articles may not have an infobox, but these are likely to be rejected anyway because of the minimum length requirement.

3.2 Processing Texts

Following our definition of biography as a running text, we strip all the additional markdown elements, as well as the references. These include infoboxes, illustrations and other media, footnotes, and hyperlinks to other Wikipedia pages. We also strip the sections that consist solely or primarily of external references, such as 'External links' and 'See also'. While processing, we also extract some supplementary information, such as the associated categories and any alternate names.

3.3 Data Statement

Curation Rationale - The goal of the dataset was to extract biographies as per our definition in section 2.1. A regular expression per language was used to match data from a Wikipedia dump. The regular expressions were developed by the authors in their first languages who tried to find a small set of categories that would extract most biographies. In general, we use the categories *living people*, *born*, and *died*.

Languages - The languages currently available are English (en), Swedish (sv), Russian (ru), Chinese (zh), and Farsi (fa)[2]. Mentions of Chinese in this paper means the *zhwiki* which consists of both Mandarin and Cantonese. The size of the files and the number of words are in Table 1. More languages can easily be added following the guidelines in the code[3].

Author and Annotator Demographics - The authors of the texts on Wikipedia are not explicitly mentioned due to the open-source nature of Wikipedia. The latest available survey states that contributors are 86.73% male, 12.64% female, and 0.63% other (Glott et al., 2010).

No explicit annotations are included with this work, although we can consider the categories that are collected along with each biography as annotations. These categories are applied by the Wikipedia authors and, hence, annotators can be assumed to be of a similar demographic to authors.

Speech situation - The corpora are written texts intended to give neutral information[4] about people, which are aimed at a general audience. The texts are continuously and asynchronously edited by many contributors and therefore assumed to have a modern speech mode. As the speech mode and content of the articles may change along with societal shifts, it is recommended to download a suitably recent dump when working with this data.

Columns - The following columns are produced by the default biography extractor: *title*, *names*, *categories*, *body*. *Title* is the name of the biography, usually the name of a person. *Names* are the different names that link to the specific biography and may include formal titles, stage names, prior names, etc. *Categories* are the extracted categories that have been given to the biographies by the contributors. The body is the running text which has been stripped of image texts, links, tables and other markdown artefacts.

3.4 Corpora Statistics

The basic statistical analysis of the corpora collected for this paper can be seen in Table 1. The word and character counts, together with the more in-depth distributions shown in Figure 1, give a

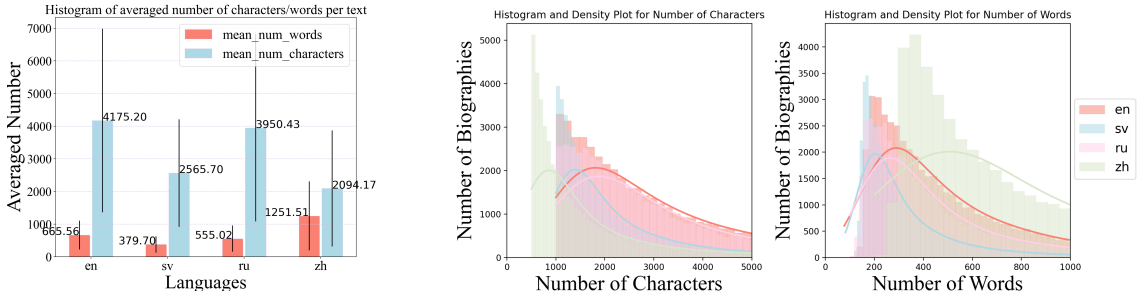
[2]Demo cases are not available for Farsi, as we did not have an L1 speaker available for the analysis.

[3]https://github.com/antoneklund/wikipedia-biographies

[4]Wikipedia enforces a 'neutral point of view' for all encyclopedic content, although in practice editor bias remains Hube (2017).

Language	Biographies	Size	avg. Char. per Article	avg. Words per Article	avg. Categories per Article
English	1,219,516	5.9GB	4,175.20	665.56	10.26
Swedish	107,868	332.9MB	2,565.70	379.70	8.44
Russian	331,655	2.5GB	3,950.43	555.02	5.41
Chinese	92,540	484.6MB	2,094.17	1,251.51	7.75

Table 1: Comparative overview of some basic statistics about our corpora. The averages are calculated from a sample of 50,000 articles in each language.



(a) (b)

Figure 1: (a): Comparisons of averaged characters per article and averaged words per article between different language biographies.(b): Density distributions of the number of characters and words of the sampled data texts.

rough overview of how the biographies manifest in different languages.

We are interested in the biographies mainly for their potential, among other use cases, in studying the narrative structure and social biases. Therefore, in the statistical analysis, only articles where the running text is sufficiently long were used. What is considered a sufficient length, is adjusted as appropriate for different languages. In this paper, for English, Swedish, and Russian, a minimum of 1000 characters of running text was used. For Chinese, a minimum of 500 characters were used because, in most cases, written Chinese uses fewer characters to represent the same amount of information as the other languages. Also, based on our statistical analysis, it is evident that Chinese biographies have an average number of characters lower than the other languages, and hence, the limit is adjusted accordingly. For other applications, scopes and languages, we suggest adjusting the character limit as appropriate.

4 Demo Cases

Two demo cases were designed to demonstrate some general usage of the corpora and to acquire more latent information about their contents. The corpora are well-situated to study societal structures and how information is relayed, and make comparisons across languages. The first demo case (section 4.1) is a study on topical differences between languages with a focus on gendered themes. The second demo case (section 4.2) is a cluster analysis of the corpora to visualise how the biographies are broadly divided into clusters depending on their running text.

4.1 LDA Topic Modelling

One main strength of the corpora is their multilinguality. A natural first study is to compare the content of the corpora for all the languages, looking for regional differences. We focus on how gendered terms are used in the biographies using a pared-down and fully unsupervised variant of the methodology described by Devinney et al. (2020).

We use gensim[5] Latent Dirichlet Allocation (LDA, Blei et al. (2003)) to model topics in the data. We use a sample of 50,000 articles per language and pre-process the articles with lemmatisation (except for Chinese) and removal of stop-words. The stop-word list for each language was modified to allow gendered words like *he*, *she*, and *they* in the text, as we want to study the occurrences of these words in the generated topics. The Chinese stop-word list was extended to include both simplified and traditional forms. Lemmatising was done with nltk WordNet[6] (English), efselab[7] (Swedish), and

[5] https://radimrehurek.com/gensim/
[6] https://www.nltk.org/
[7] https://github.com/robertostling/

pymystem3[8] (Russian). We use the jieba[9] package to pre-process the Chinese corpus.

We generate 50 topics and used the top 30 highest-weighted terms for each topic to label them with their apparent themes (e.g. *Chinese history*, *Education/academia*). Topics were labelled by an L1 user, with an L2 user checking for agreement where possible. From these, we created 20 general themes to allow for better comparison across languages. The breakdown of general themes for each language can be seen in Table 2.

From this analysis, we can see that *Sports* and *Entertainment* make up a significant number of topics in all samples. The topics with *History*, *War/military*, *Politics/government* and *Places* account for most of the rest. The Chinese sample notably has more topics around *Entertainment* and *Sports* and only one about *Places*. The *Places* theme is more common in other languages and the English corpus has by far the most. We suspect that this theme is intermixed with *History*.

When we subdivide *History* into *History (local)* and *History (foreign)*, we can see that there is a greater number of history topics local to a language, indicating there is likely more detail or nuance latent in the data. Furthermore, the foreign history topics remain focused on history that is 'close to home', with English, Swedish, and Russian remaining quite heavily focused on European history and places. Chinese, while still including European history, has more East Asian topics.

The number of *No clear theme* topics is similar between the models. These include the captured structural elements such as tables and language artefacts foreign to a particular Wikipedia (e.g. English terms in the Russian sample). This may indicate that other cleaning choices (e.g. more thoroughly removing the tables) may be preferable depending on the task.

4.1.1 Gendered Analysis

We take a closer look at some of the gendered patterns made evident by topic modelling. We identify topics where gendered pronouns or other lexically-gendered terms are highly weighted and relate the general themes of the topics to these gendered associations.

For the English sample, masculine pronouns (e.g. *he*, *his*) appear frequently in topics related to politics, war, and inheritance, although they also appear in a number of other topics across a wide range of subjects. Feminine pronouns (e.g. *she*, *her*), in contrast, are highly weighted in only one topic: family and relationships. We find similar patterns in Russian and Swedish, where masculine terms appear in a wide range of topics and feminine terms are confined to only one or two, with a focus on the domestic sphere of family and/or romantic relationships.

For the Chinese sample, masculine terms (e.g. 他- he, 男子- male) appear frequently in topics related to sports, history, and politics, while feminine terms (e.g. 她- she, 女子- female) are more common in topics of TV series and music, a notable departure from our other three samples. Although women are mentioned in sports-related topics, they are almost absent in the top 30 most frequently mentioned keywords of political topics.

From counting pronoun frequency in our samples (Figure 2,) we know that masculine pronouns vastly outweigh feminine pronouns[10]; and nonbinary pronouns (e.g. *ze*, *hir*) are extremely rare (where they can be clearly disambiguated from neutral or plural pronouns). The distribution of the number of gender-associated topics (masculine more frequent than feminine; nonbinary excluded) can somewhat be expected based on these term distributions. However, both are evidence of the hierarchical relationship, where men are 'more' – talked about, present in the data, valued – than

Themes	en	sv	ru	zh
Entertainment	4	3	2	6
Sports	9	9	7	12
Music	2	3	2	2
Art	1	2	1	1
Literature	1	3	1	1
Journalism	1	0	0	0
Business	0	0	1	1
Science/Technology	2	0	2	0
Education/Academia	2	2	0	2
History (local)	2	2	4	6
History (foreign)	0	2	3	6
Places	13	5	7	1
Religion	2	3	1	1
War/Military	2	2	4	1
Politics/Government	3	4	2	4
Crime	1	1	1	0
Family	1	1	1	0
General Biography	0	2	1	3
(No clear theme)	2	3	4	3

Table 2: Summary of themes found for unsupervised LDA with 50 topics, run on samples of 50k biographies.

efselab

[8] https://pypi.org/project/pymystem3/

[9] https://github.com/fxsjy/jieba

[10] In the case of Russian, this may be in part due to language-specific behaviour of grammatical gender.

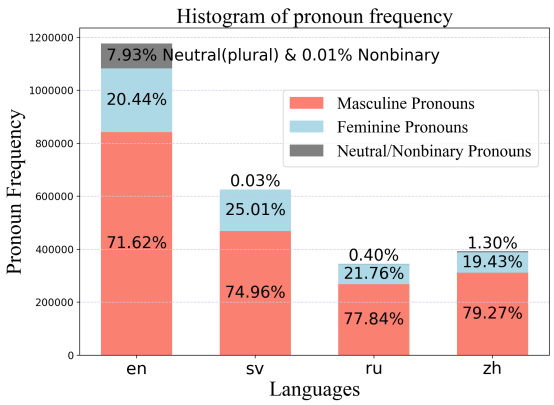


Figure 2: Pronoun frequency in each sampled corpus by gender, calculated after preprocessing.

women. We also see evidence that these hierarchies surface differently in our different samples, according to the different cultural hegemonies. The European languages relegate women to the private sphere, whereas men take up the public sphere and are treated as the unmarked norm (meaning they can 'be' almost anything). The Chinese sample puts men in 'serious' or important topics, and women in those related to entertainment and other less serious pursuits. Our findings correlate well with other research on gender bias in Wikipedia, e.g. (Sun and Peng, 2021; Schmahl et al., 2020).

4.2 Cluster Analysis

To look for writing-style patterns in the biography texts we use the BERTopic pipeline (Grootendorst, 2022) to create clusters of biographies. A random sample of $50,000$ biographies was used for each language. The text is vectorised with the multilingual model XLM-RoBERTa[11] (Conneau et al., 2020). Then, the vectors are projected to two dimensions using UMAP (McInnes et al., 2018) and then clustered with HDBSCAN (Campello et al., 2013). This results in 2D plots for each language where the clusters in theory represent biographies that are similar to each other. The plots can be seen in Figures 3(a)–3(d). The keywords are extracted using c-TF-IDF that was introduced in Grootendorst (2022) with an extended stop word list for cluster visualisation.

The structure of the vector space reveals clear clusters that have been formed for all languages. English, Swedish, Russian, and Chinese have six, six, five, and nine clusters respectively, with the model set to find coarse-grained clusters. The largest clusters could be generally labelled as being about people from the core regions of the language. E.g. an English cluster about Americans and a Swedish cluster about Swedes. The smaller clusters have more informative keywords about a specific group of biographies. This could be a topic about hockey players which are found in English, Swedish, and Russian, or other sports and TV series that were found in Chinese. Smaller clusters reveal more distinct themes such as the *Communist Party of China* or *Theatre*. This indicates that a deeper analysis with finer-granular clusters would probably reveal more interesting structures.

In general, many clusters are about sports for all languages. This indicates that there are many athlete biographies, which may follow a structure of writing that is distinctly different from those of other persons. These writing patterns are revealed by the clustering system which shows multiple sports clusters while the other biographies are in a larger shared cluster. This indicates that there is a writing pattern in how people related to sports differ from many other categories of biographies. These other categories, such as themes of *History* or *Entertainment* seem to share a common writing style for biographies.

5 Limitations

The aim of this work was to collect as many Wikipedia articles as possible that fit the criteria for being a biography. While the corpora presented in this study are largely biographies, there are articles that evade the filter, e.g. the Wikipedia category 'mountaineering deaths' includes both biographies and articles about accidents. Although these articles are easy to manually identify, we do not remove them as they must be considered individually and we found them to be extremely uncommon.

False negatives are harder to identify. We generally assume that all biographies have at least one of the patterns: 'born', 'death', or 'living'. In cases where a person does not have these, it may be the case that they have a sufficiently mythological status (for example, the Buddha is not captured in our English corpus). More likely, however, it could be due to human error when editing the page. We recommend making manual checks with samples of biographies expected to be in the data. This is especially important when considering biographies of people belonging to marginalised groups, who

[11] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

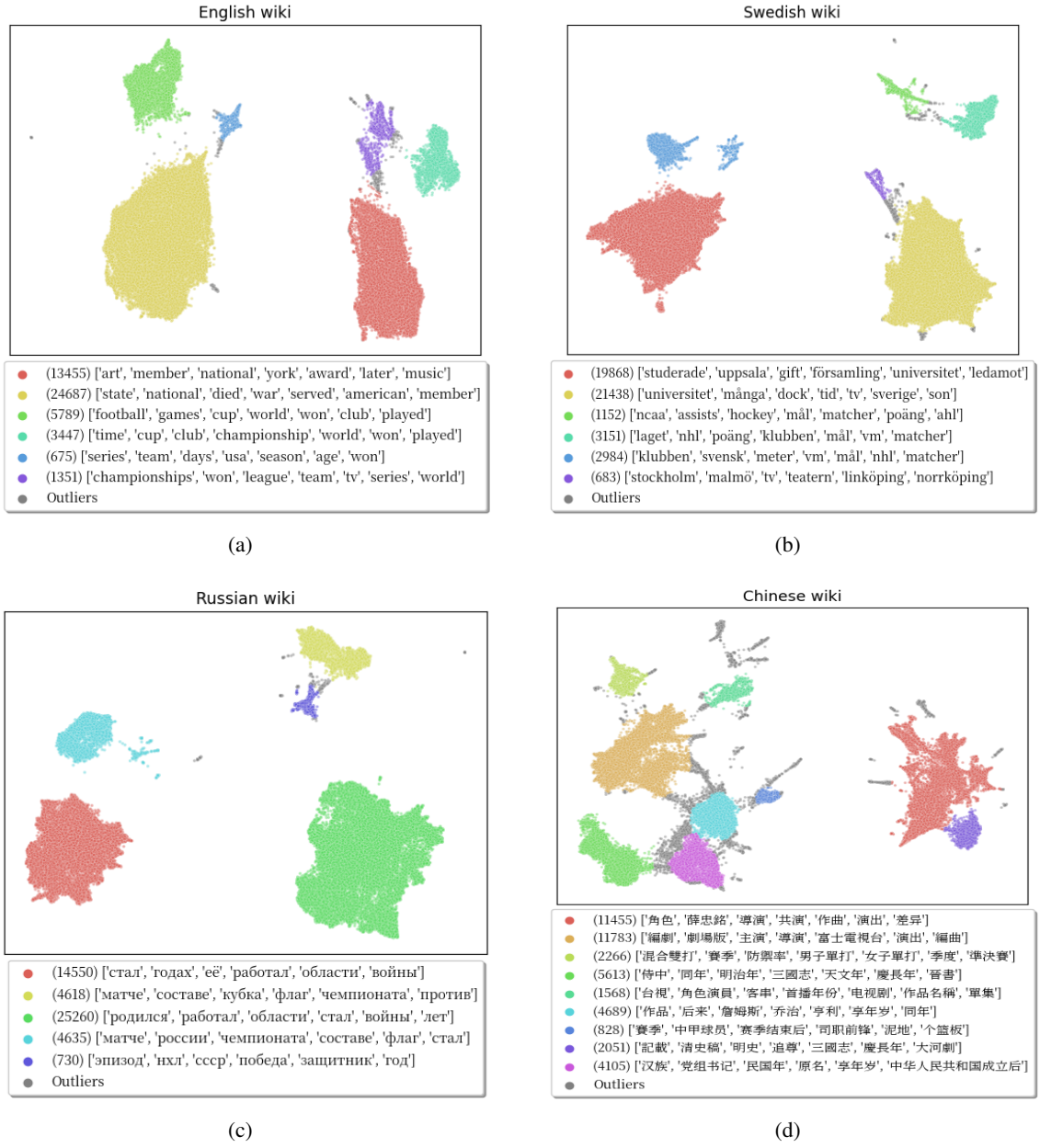


Figure 3: 2D plots of the English (a), Swedish (b), Russian (c), and Chinese (d) biographies. The larger groups formed have the keywords shown in the legend.

may be less likely to be seen as 'significant' and thus not be properly curated.

We attempted to strip the text from all links, tables and other clutter to only have the running text of the biography easily accessible in the dataset. We can not guarantee that the texts do not contain any errors from the cleaning, and some NLP applications may require different information (such as extracting links) which we do not provide.

While analysing underlying gendered patterns can be done through topic modelling, this technique is not well-suited to languages where grammatical and social gender overlap. It also fails in data-sparse contexts, such as for gender-diverse populations. Our demo is provided as a proof-of-concept of the utility of the corpora for social bias analysis, and as a warning that they should not be used uncritically: more detailed analysis and mitigation, tailored to specific use cases, will be necessary for all social biases.

Finally, these corpora should not be used as a benchmark for pre-trained models that where constructed using Wikipedia in their training data.

6 Conclusions

The Multilingual Wikipedia Biographies is mainly a method for extracting an up-to-date high-quality dataset from the Wikipedia dump. The method is easily adaptable to other languages including those with low resources. Some general structures were common between languages such as masculine terms being generally more prevalent in topics compared to feminine, which were constrained to more specific topics. The distribution of biographies is naturally highest in the language's core region and gradually declines as it extends outward. The corpora allow for comparing the structures and composition of biographies in multiple languages which is important for understanding how Wikipedia biographies shape how information about individuals, and society, as a whole is shared.

Ethical Considerations

In this paper, we explore a very surface-level understanding of gender bias, focusing on how the potential for representational harms can be seen for groups and individuals of different genders (studying masculine, feminine, and neutral/nonbinary representation). In our case, representational harm is concerned with stereotyping (e.g. women are most associated with home/family) and erasure (e.g. nonbinary people are largely not present in the samples). Although we do not explore other biases, such as race or class, as well as intersectional biases; we expect these representational harms to also be present and discoverable in the corpora, as Wikipedia is not written or curated in e.g. specifically anti-racist ways. We do not attempt to mitigate any of these harms, because we believe they provide valuable data about cultural and societal norms and attitudes, which may be important for research. However, this also means that there is an additional risk of perpetuating and even amplifying stereotypes or erasure if the data are used uncritically.

This dataset contains publicly available information about living people. Crucially, this information may go (or already be) out of date and we encourage the use of the provided code on a recent Wikipedia dump when appropriate.

Although this dataset and de facto annotations (in the form of category tags) are publicly available and can be used and shared for research under Wikipedia's Creative Commons by Share Alike license, it is still worth acknowledging that we are collecting other people's words and labour.

Statement on the use of AI tools. No parts of the text in this paper were written with the help of any generative AI.

Acknowledgements

We would especially like to thank our colleagues, Arezoo Hatefi, for her translations, quality checking, and general expertise in developing the Farsi corpus, and Henrik Björklund, for his support and feedback on our ideas.

The authors and this research are supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), the Marianne and Marcus Wallenberg Foundation, the Swedish Research Council, the Swedish Foundation for Strategic Research, Adlede, and Umeå Centre for Gender Studies.

Contributions

H.D. conceived the idea which was developed together with I.R. and A.E. H.D. and A.E. implemented the code base with contributions from I.R. H.D. and I.R. provided qualitative motivations and background. H.D. analysed the corpus in English in discussions with I.R., A.E. and J.C. A.E. and H.D. analysed the corpus in Swedish. I.R. analysed the corpus in Russian. J.C. analysed the corpus in Chinese. A.E. provided the BERTopic analysis. J.C. provided quantitative statistics and visualisation. H.D. outlined the paper and all authors collaboratively contributed to the final manuscript.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21577.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based

on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish. In *2nd Workshop on Gender Bias in Natural Langauge Processing*, pages 79–92.

Jennifer Drexler, Pushpendre Rastogi, Jacqueline Aguilar, Benjamin Van Durme, and Matt Post. 2014. A Wikipedia-based corpus for contextualized machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3593–3596, Reykjavik, Iceland. European Language Resources Association (ELRA).

Angela Fan and Claire Gardent. 2022. Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Núria Ferran-Ferrer, Marc Miquel-Ribé, Julio Meneses, and Julià Minguillón. 2022. The gender perspective in wikipedia: A content and participation challenge. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 1319–1323, New York, NY, USA. Association for Computing Machinery.

Anjalie Field, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. Controlled Analyses of Social Biases in Wikipedia Bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635, Virtual Event, Lyon France. ACM.

Shen Gao, Xiuying Chen, Chang Liu, Dongyan Zhao, and Rui Yan. 2021. BioGen: Generating biography summary under table guidance on Wikipedia. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4752–4757, Online. Association for Computational Linguistics.

Debela Tesfaye Gemechu, Michael Zock, and Solomon Teferra. 2016. Combining syntactic patterns and Wikipedia's hierarchy of hyperlinks to extract meronym relations. In *Proceedings of the NAACL Student Research Workshop*, pages 29–36, San Diego, California. Association for Computational Linguistics.

Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group*, 8:1158–1178.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Alexander Hogue, Joel Nothman, and James R. Curran. 2014. Unsupervised biographical event extraction using Wikipedia traffic. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 41–49, Melbourne, Australia.

Christoph Hube. 2017. Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 717–721, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Max Klein and Piotr Konieczny. 2015. Wikipedia in the world of global gender inequality indices: What the biography gender gap is measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*, OpenSym '15, New York, NY, USA. Association for Computing Machinery.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Alessio Palmero Aprosio and Sara Tonelli. 2015. Recognizing biographical sections in Wikipedia. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Lisbon, Portugal. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katja Geertruida Schmahl, Tom Julian Viering, Stavros Makrodimitris, Arman Naseri Jahfari, David Tax, and

Marco Loog. 2020. Is Wikipedia succeeding in reducing gender bias? assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 94–103, Online. Association for Computational Linguistics.

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

A Categories and Regular Expressions by Language

English (en) The English corpus regular expression captures biographies only by categories, and checks for the presence of any of the following: living people, births, and deaths.

```
\[\[Category:(Living people|
.*deaths|.*births)
```

Swedish (sv) The Swedish corpus regular expression captures biographies only by categories, and checks for the presence of any of the following: Living People, Births, and Deaths.

```
\[\[Kategori:(Levande personer|
Födda.*|Avlidna.*)
```

Russian (ru) The Russian corpus regular expression captures biographies by categories (personae alphabetically, births, and deaths) and common lines from infoboxes which we expect only to be present for persons: date of birth, date of death, place of birth, and place of death. While the birth year and death year categories are automatically added and, therefore, not captured in most cases, they are included for redundancy. The same mask also captures non-automated categories related to birth and death places, causes, etc.

```
\| * [Дд]ата рождения |\| * [Дд]ата
смерти |\| * [Мм]есто рождения
|\| * [Мм]есто смерти |\[\[ Катего-
рия:(Персоналии по алфавиту|Родившиеся.*
|\
```

Chinese (zh) The Chinese corpus regular expression captures biographies by both the `bd` template (which automatically generates births and deaths categories) and the following categories: living people, births, and deaths.

```
(\[\[(Category|分类):(在世人物|.*逝
世|.*出生))|(\{\{bd\|.*\}\})
```

Farsi (fa) The Farsi corpus regular expression captures biographies only by categories, and checks for the presence of any of the following: living people, births, and deaths.

```
\[\[(*. زنده افراد |*. زادگان |*.
رده : ( درگذشتگان \]\]
```