CONSIDERATIONS FOR MULTILINGUAL WIKIPEDIA RESEARCH

Isaac Johnson & Emily Lescak
Wikimedia Foundation
{isaac,elescak}@wikimedia.org

ABSTRACT

English Wikipedia has long been an important data source for much research and natural language machine learning modeling. The growth of non-English language editions of Wikipedia, greater computational resources, and calls for equity in the performance of language and multimodal models have led to the inclusion of many more language editions of Wikipedia in datasets and models. Building better multilingual and multimodal models requires more than just access to expanded datasets; it also requires a better understanding of what is in the data and how this content was generated. This paper seeks to provide some background to help researchers think about what differences might arise between different language editions of Wikipedia and how that might affect their models. It details three major ways in which content differences between language editions arise (local context, community and governance, and technology) and recommendations for good practices when using multilingual and multimodal data for research and modeling.

1 INTRODUCTION

Wikipedia has been around for over 20 years and has been studied and modeled by researchers for almost as many, but it is not a monolithic resource. By 2005, there were already almost 200 language editions of Wikipedia and, as of 2022, there are over 300.\footnote{https://meta.wikimedia.org/wiki/List_of_Wikipedias} Voss (2005) analyzed the structure and differences between several language editions (German, Japanese, Danish, and Croatian) as early as 2005, notably excluding English Wikipedia due to 36,000 geographic articles that were bot-generated, which “biases some statistics.” Since then, however, the focus has largely been on the textual content of English Wikipedia. It is only recently that the research community has expanded their work to be more multilingual—including more languages, especially non-English and smaller language editions—and multimodal—also studying the images and other multimedia content within Wikipedia and corresponding media repository Wikimedia Commons.

While multilingual and multimodal models are an important step forward, it is likewise important that researchers understand their increasingly multifaceted Wikipedia datasets so that they may use them responsibly. It is likely not necessary to exclude full language editions as Voss (2005) did, but the dynamics that he described (bot-generated content and the skew that that can introduce as well as other aspects described in this work) are still present in many language editions of Wikipedia\footnote{In fact, many of those 36,000 bot-generated articles mentioned by Voss about towns and cities on English Wikipedia remain largely unchanged per Johnson et al. (2016).}.

This paper seeks to support researchers working with multilingual and multimodal Wikipedia content through a survey of findings about the differences in content across languages and how they arise. These findings from the literature are complemented with the authors’ own personal experiences when working with Wikimedia data and conversations with other researchers. Three major sources of differences are identified and detailed with examples: the local context, editor communities and governance, and technologies used by editors. Many of these facets are relevant even when working with a single language edition of Wikipedia, but their relevance becomes more pronounced when trying to handle multiple language editions in a consistent manner. The paper finishes with a few recommendations for good practices when doing this research and open areas of future research.
2 BACKGROUND

We provide context on the role of Wikipedia data in large language models, a brief summary of the importance of understanding data quality, and some of the research that has more deeply explored differences across languages in Wikipedia.

2.1 WIKIPEDIA IN LARGE LANGUAGE MODELS

Many (perhaps all) large language models have made extensive use of text from Wikipedia to help train their models. Wikipedia is a substantive component of The Pile (Gao et al. (2020)) and Common Crawl snapshots (Dodge et al. (2021)) that are used in training models like Megatron-Turing NLG 530B (Smith et al. (2022)), more than half of the data used to train the original BERT model (Devlin et al. (2018)), and was explicitly excluded from the WebText corpora used in training the GPT-2 and GPT-3 models due to its appearance in the other datasets that were used (Radford et al. (2019)). It also figures prominently in image-text datasets such as the Wikipedia-Based Image Text (WIT) Dataset (Srinivasan et al. (2021)) and LAION-400M (Schuhmann et al. (2021)), the latter of which is derived from the Common Crawl and thus almost certainly includes Wikipedia articles. Wikipedia also is an important source of many standard benchmark datasets – e.g., Wikimedia for language modeling (Merity et al. (2016)) and SQuAD for reading comprehension (Rajpurkar et al. (2016)).

For the preprocessing of Wikipedia datasets, there are two high-level questions that must be asked: 1) what articles (or images) are included, and, 2) for articles, what text cleaning steps are used to convert the raw markdown used to write Wikipedia articles (wikitext) into natural language free of extraneous syntax or noisy text. Regarding the first question, almost all Wikipedia datasets for training language models use all articles in the main namespace (regardless of quality and subject). Accordingly, image datasets such as WIT often use images that are included within Wikipedia articles (Srinivasan et al. (2021)). Wikipedia-derived benchmarks, however, such as WIT and SQuAD do filter the articles extensively and often only use high-quality articles (featured articles and articles with high PageRank values respectively). An exception to this is Guo et al. (2020), who released a pre-processed multilingual Wikipedia dataset for use in language models and took the additional step of filtering out pages unlikely to have high-quality text such as disambiguation pages and list articles. The preprocessing pipelines for parsing wikitext also vary greatly – e.g., a simple Perl script used by fastText that removes some basic wikitext syntax, a more fully-fledged Python tool used by BERT that removes some syntax but also expands templates on Wikipedia, a Python script used by HuggingFace that has its own process for cleaning syntax and the process described in Guo et al. (2020) that indicates some of the syntax/objects that it seeks to remove like tables, lists, and reference sections that are unlikely to have high-quality text (though no code is provided).

2.2 IMPORTANCE OF DATA TO MODELING

While datasets are highly important to the functioning of models, in-depth studies of their quality and approaches to improving that quality are often undervalued (Gebru et al. (2021); Sambasivan et al. (2021); Paullada et al. (2021)). For instance, studies that have examined some of the multilingual and multimodal large-scale web-crawled datasets (that often contain Wikipedia as detailed above) find many issues. Kreutzer et al. (2021) provides an in-depth look into the quality of several large-scale multilingual datasets and found substantive errors, including in the parallel sentence dataset WikiMatrix (Schwenk et al. (2021)) that is extracted from Wikipedia articles. For WikiMatrix, the issues seemed to arise due to the automated process used to find parallel sentences but also stemmed from some of the ways in which Wikipedia language editions are divided up. Kreutzer et al. (2021) goes on to describe some of the potential negative downstream effects of these issues such as reduced model quality, representation washing (false inclusion of low-resourced languages), and trust in incorrect “facts” extracted from models due to improper alignment of sentences. To quote them, “data cleaning is no trivial task!”

A quick examination of a sample of the data confirmed this fact.

[https://github.com/facebookresearch/fastText/blob/main/wikifil.pl](https://github.com/facebookresearch/fastText/blob/main/wikifil.pl)

[https://github.com/attardi/wikiextractor](https://github.com/attardi/wikiextractor)

[https://github.com/huggingface/datasets/blob/master/datasets/wikipedia/wikipedia.py](https://github.com/huggingface/datasets/blob/master/datasets/wikipedia/wikipedia.py)
Birhane et al. (2021) examine the images and captions in LAION-400M (Schuhmann et al. 2021) and found many examples of disturbing imagery and highly-problematic stereotypes or captions.

2.3 MULTILINGUAL WIKIPEDIA RESEARCH

Much research (a lot of which referenced below comes from the field of computational social science) has examined differences between different language editions of Wikipedia from the standpoints of content (e.g., Bao et al. (2012); Warncke-Wang et al. (2012); He et al. (2018); Beytía et al. (2022)), readers (e.g., Johnson et al. (2021b); Lemmerich et al. (2019); Arora et al. (2022)), and editors (e.g., Kim et al. (2016); Bipat et al. (2018); Sen et al. (2015)). These multilingual and multimodal Wikipedia studies provide insight into what differences arise across language editions and some of the factors that may cause these differences (further detailed in Section 3). Much of these differences relate back to coverage and quality of text in Wikipedia and thus the quality of machine learning models trained on this content. While some coverage biases on Wikipedia are relatively consistent across language editions such as the gender gap others vary greatly based on what places and cultures are written about in a given language (Miquel-Ribé & Laniado 2019). Better understanding these differences can help researchers and practitioners to construct higher-quality datasets from Wikipedia and understand what is and is not included.

3 FACTORS AFFECTING WIKIPEDIA DATA ACROSS LANGUAGES

Below, we outline some of the sources of major differences in content (and therefore data) between Wikipedia language editions based on an informal survey of research and documentation about the Wikipedia projects. These facets are certainly incomplete but cover a wide range of aspects to be aware of when working with multilingual and multimodal Wikipedia content. The impacts of some of these facets are direct—e.g., access to translation tools that greatly affect what content is created (articles or images that already exist on other language editions, often English, as described in Warncke-Wang et al. (2012)) and the style of writing (the content often starts from machine translation)—but others are far more subtle though no less important—e.g., barriers to new editors that might reduce the diversity of an editor community and therefore its content.

3.1 LOCAL CONTEXT

The first set of facets touches on the offline context for editors of a particular language edition.

3.1.1 GEOGRAPHY AND CULTURE

The Wikipedia projects are divided up by language and not geography. This simple fact has two major implications.

First, the different language editions of Wikipedia serve very different communities (Johnson et al. 2021b; Lemmerich et al. 2019) and thus often cover very different topics (Bao et al. 2012; Hecht & Gergle 2009; Miquel-Ribé & Laniado 2020). The resulting variation in quality and quantity of content about different topics (Lewoniewski et al. 2017b) presumably would affect the resulting vocabulary and ability of language models trained on Wikipedia to accurately handle different topics. Even when language editions cover the same topics, editors often choose culturally-relevant imagery to include (He et al. 2018).

Second, not only are different language editions contextualized within different cultures (Chelsy Xie et al. 2019), but the diversity of these cultures varies greatly as well. On one end are languages like English, French, or Arabic that have been associated with colonialism or large diasporas and thus are home to editors from many different countries and social contexts. On the other end are languages like Croatian or Japanese that are much more specific to a single country and social context. The position on this spectrum can have trivial effects on the content—e.g., variation in dialects or spellings of words such as American English and British English—that could affect

For a fuller accounting of the different gaps across the Wikimedia projects, see Redi et al. (2020).

https://humaniki.wmcloud.org/
https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style#National_varieties_of_English
language models—as well as much larger impacts on the content. For an example of the latter, community diversity appears to be an important factor in creating high-quality content (Shi et al. (2019)) maintaining Wikipedia’s core content policy of Neutral Point of View,

which if not followed, can lead to highly problematic content, especially around history or political topics (e.g., Wikimedia Foundation (2021a); Satô (2021)).

### 3.1.2 Source Availability

As a tertiary source that itself depends on reliable sources to support what is written in it, Wikipedia reflects the world and its biases (Maher (2018)). While this helps maintain the high quality and verifiability of content, it also introduces biases based on what topics or people are themselves well-documented. For languages with fewer resources, especially when they are not digital or accessible (Gill (2021)), writing high-quality content can be much more difficult and the perspectives presented can be drawn from non-local viewpoints based on the sources available (Sen et al. (2015); Lewoniewski et al. (2017a)).

### 3.2 Editor Community and Governance

The second set of facets focuses on aspects of governance that affect what content is contributed and by whom to a given Wikipedia.

#### 3.2.1 Rules

The most concrete aspect of governance are the rules that are hard-coded into different Wikipedia language editions and govern who can contribute and how. These rules range from governing who can edit articles—e.g., while no accounts are required to edit content on most Wikipedia language editions, Portuguese Wikipedia requires an account to edit (Wang et al. (2021)), which is known to discourage new editors from contributing (Mako Hill et al. (2019))—to whether articles can easily be generated based on translations of other language editions—e.g., an option only available for a select few editors on English Wikipedia but that is widely available on many smaller language editions (Ozurumba (2021))—to whether the Wikipedians who are able to do many core moderation tasks are native speakers of the language.

#### 3.2.2 Policies

Along with rules, Wikipedia language editions also have core policies that guide what content can be included in that language edition. While the core content policies are common across language editions, other policies can differ substantially (Wikimedia Foundation (2021b)). For example, English Wikipedia accepts fair-use imagery that is not allowed in Wikimedia Commons (the main image repository for Wikipedia). As a result, English Wikipedia contains many images of company logos, movie posters, and other otherwise protected content that would not be found on other language editions and would greatly affect the types of media and captions in datasets built from English Wikipedia as opposed to Wikimedia Commons. What constitutes a reliable source also can vary greatly by language edition (Berson et al. (2021)), which has massive implications for what topics are notable enough for inclusion and what points-of-view can be expressed in the content.

#### 3.2.3 Norms

Even with similar rules and policies, different language editions likely will have different norms about how content is created. How work is coordinated—e.g., discussions on talk pages or Wikiprojects as opposed to off-wiki groups on Facebook, email listservs, or chat apps like Telegram—directly impacts any conversational data extracted from Wikipedia (e.g., [Hua et al. (2018)]) and met-
rics related to collaboration but also might impact the resulting quality of articles (Morgan et al. 2013). How articles are improved—e.g., in draft workspaces, through fewer but larger edits, through many incremental edits—would affect edit datasets and tools such as vandalism detection that aim to evaluate edits as well as datasets that aim to capture cleaned-up versions of articles created via translation.

3.2.4 SIZE, AGE, AND COMPOSITION

All three prior facets (rules, policies, and norms) are further mediated by the size, age, and composition of a given Wikipedia project. Smaller and younger language editions generally lack tools to support their work. They also can be less of a target of vandalism and less entrenched in their norms (TeBlumhuis et al. 2013) but might display more idiosyncrasies in terms of their content due to more relaxed policies (Keegan & Fiesler 2017) or based on early editors’ interests and abilities to generate content via bots or semi-automated means (Guldbrandsson 2013). While there are large gaps in the diversity of all Wikipedia editor communities (Redi et al. 2020), smaller communities almost by definition will have less diversity.

3.3 TECHNOLOGIES

The third set of facets deals with the technologies used by Wikipedians to contribute to Wikipedia, which can vary substantially by language edition.

3.3.1 INTERFACES AND TOOLS

While every language edition uses the same core MediaWiki software to edit their version of Wikipedia, much of the tooling and interfaces used by editors can be configured in various ways or exists outside of that core software in extensions (Geiger 2014). A salient example of this is an extension called Structured Discussions (or “Flow”) that, as the name suggests, provides more structure around talk page discussions. The increase in structure around conversations can have a substantive impact on the nature of discussions (Aragón et al. 2017). Other major extensions with mixed uptake include Flagged Revisions (determines whether a revision is shown to a reader before review), Content Translation (discussed in Section 3.2.1), Newcomer Tasks (edit recommendations for new editors that include tasks such as adding images to articles) and PageAssessments (simplifies tagging of content by Wikiprojects with topic, quality, and importance annotations).

3.3.2 BOTS AND FILTERS

Bots and filters are also examples of tools that vary widely and have an even more direct relationship to edits on Wikipedia. Taking the example of vandalism detection, English Wikipedia has an extensive AbuseFilter configuration that catches many bad edits before they are even published, automated bots such as ClueBot NG and RecentChanges filter based on machine learning models for detecting bad-faith and damaging edits (Halfaker & Geiger 2020). This suite of tooling both reduces the amount of vandalism that reaches Wikipedia and greatly speeds up the response time of editors to vandalism by helping them to quickly identify problematic edits (Geiger & Halfaker 2013). English Wikipedia, as the largest language community, has many specialized tools that blur
the lines between human and bot editing, such as AutoWikiBrowser or Twinkle and, as a result, has a very high rate of small edits done through these tools as a part of routine maintenance tasks.

Awareness of the origin of content and edits can help greatly in identifying what might be more natural language as opposed to large-scale, rule-based language generation. While bot-based generation of entire articles is less common on English Wikipedia now, many articles still have their origins as bot-generated and have been minimally edited since then (Johnson et al., 2016). This content is largely standard text with the key facts differing between articles—e.g., the population of a town or its geographic neighbors, all of which can be automatically extracted from databases. Tool-based editing to, for example, fix common misspellings, could skew grammatical error correction corpora built from Wikipedia (e.g., Lichtarge et al., 2019). Tools exist to identify these patterns (e.g., Alshomary et al., 2019) and a high-quality language model might seek to deduplicate this content or downweight it in training.

3.3.3 TRANSCLUDED CONTENT

Mitrevski et al. (2020) and Johnson (2020) have documented a growing trend on certain Wikipedia language editions in the shift from content being written within the wikitext of an article to content being transcluded from other pages or via complicated logic—e.g., copying a standard table of links for all articles about a particular region, auto-filling infoboxes with standard facts such as a town’s population or a country’s flag. This brings consistency for readers and eases the maintenance burden for editors (only have to update content in one place, not in all the articles in which it is used), but it also leads to a growing disconnect between the content directly expressed in an article’s wikitext and the parsed content that comprises the article. As with the discussion of bots and filters above, filtering out this highly-structured, templated content may be desirable for language models in some circumstances. However, for research that relies on extracting facts from infoboxes or using links to build a graph representation of Wikipedia, a large difference would be seen between the data extracted from wikitext and the data extracted from the final HTML of a page.

4 RECOMMENDATIONS

The above sections identified reasons why the content in different Wikipedia language editions can vary in quality, topic, and appropriateness for various multilingual and multimodal tasks. Below, we offer recommendations for steps to address some of these concerns and better tune Wikipedia-based datasets to a particular modeling task.

4.1 SITUATED RESEARCHERS

The first and perhaps most important recommendation is to have researchers who are situated within the communities being studied. This is not specific to Wikipedia research, with calls for this level of inclusion and participatory research such as for natural language processing (NLP) research more broadly (Nekoto et al., 2020) and data governance (Carroll et al., 2020). Having a research team that includes Wikipedians—i.e. editors—from that language community, native speakers, and folks who are long-time readers of a language edition is an important step beyond e.g., using machine-translated versions of content for inspecting your data. These situated researchers (e.g., Hickman et al., 2021; Bipat et al., 2021; Berson et al., 2021) can help the broader team to understand nuances in the language and identify factors like the local context, governance, and technologies above.

For researchers without existing ties to Wikimedians, there are still many ways to connect with the editor communities. While care should be taken to respect Wikipedia (as with research involving

---

26 https://en.wikipedia.org/wiki/Wikipedia:AutoWikiBrowser and https://en.wikipedia.org/wiki/Wikipedia:Twinkle respectively. For more tools, see https://en.wikipedia.org/wiki/Special:Tags

27 The final HTML of articles had long only been available via scraping or APIs, but complete HTML dumps for Wikipedia are now available to researchers: https://dumps.wikimedia.org/other/enterprise_html/

28 Using this term broadly to not just mean academics but anyone involved in understanding and analyzing the content or data
any online community) research with clear benefits to Wikipedia is generally welcomed. Beyond individual Wikimedians, there are also more formal organizations of Wikimedians ranging from very simple user groups to the more organized chapters to the Wikimedia Foundation. As laid out by Voss (2005), there are many convenings for sharing work and meeting editors, organizers, and other folks connected to the Wikimedia movement. There are also several forms of support that are available to researchers including mailing lists and technical resources to support tool-building or Wikimedia research.

4.2 Language-agnostic metrics

Language model performance is strongly tied to the amount of data available for that language—e.g., Wu & Dredze (2020)’s analysis of multilingual BERT on many low-resourced languages. When models are trained for Wikipedia-specific tasks, however, there are many language-agnostic features that can potentially be used to boost performance in lower-resourced languages. For example, in training topic models for Wikipedia, Piccardi & West (2021) and Johnson et al. (2021a) relied not on the words in an article but the links. Article links can be mapped to language-agnostic Wikidata IDs such that e.g., a link on English Wikipedia to the article for poblano peppers will be represented identically as a link to chiles poblano. This shared vocabulary means that lower-resourced languages can benefit from training data in better-resourced languages as long as they have corresponding articles and greatly reduces the amount of language-specific data needed. Other language-agnostic approaches include graph-based modeling (using links and Wikidata IDs as described for topic modeling) or taking advantage of other semi-structured components of articles—e.g., templates, sections, categories, Wikidata statements—to generate features that can be represented in a language-agnostic manner (e.g., Lewoniewski et al. (2017b); Beytía et al. (2022)).

4.3 Matched corpora

As detailed above, there are many outliers that can skew datasets generated from Wikipedia. Sometimes these outliers are of interest but oftentimes they can just add noise or otherwise obscure what the researcher would like to model. Building matched corpora can help reduce the impact of this skew in training or evaluating models. For example, Field et al. (2020) studied the relationship between an individual’s identity and how they were written about on Wikipedia. They built a matched corpora in which each article about a woman was matched with a similar article about a man (based on article categories) and then computed their metrics of interest. This matched corpora was important to isolating the specific effect of an individual’s gender while holding constant their notability or field of study. For multilingual analyses, an obvious matched corpora is one in which the same articles are included across language editions. While this could entail a large loss of data, particularly around locally-important topics that might not appear in other language editions, it would bring important consistency to e.g., benchmark datasets. Care must be taken when building matched datasets that are more fine-grained than article-level, as e.g., with the parallel sentence dataset WikiMatrix as described in Kreutzer et al. (2021).

4.4 Future research and opportunities

While this work seeks to lay out some factors to be aware of when working with Wikipedia data and good practices to follow, it cannot yet provide best practices. Future work and convenings such as Wiki-M3L or the Wikimedia community events mentioned in Section 4.1 will be necessary to build these tools and norms. As identified in Section 2.1, a good starting place would be shared code

---

29 https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_a_laboratory
30 https://meta.wikimedia.org/wiki/Wikimedia_movement_affiliated
31 https://wikimediafoundation.org/
32 https://meta.wikimedia.org/wiki/Events
33 https://lists.wikimedia.org/postorius/lists/wiki-research-l.lists.wikimedia.org/
34 https://wikiworkshop.org
35 https://wikitech.wikimedia.org/wiki/Help:Cloud_Services_Introduction
36 Q897746: https://www.wikidata.org/wiki/Q897746
37 https://meta.wikimedia.org/wiki/Wiki-M3L
and parameterizations for effectively preprocessing Wikipedia text. These practices should be supported by robust methodological research like Mitrevski et al. (2020) (who examined the difference between wikitext and parsed HTML of Wikipedia articles) or Hill & Shaw (2014) (who examined how including Wikipedia redirect pages affects studies of readership).

Researchers should also consider how their work can contribute back to the Wikimedia communities that have generated so much data that has supported progress in many fields of research. Challenges like the Wikipedia Image/Caption Matching content are a good way to participate in research with clear benefits for the Wikimedia community. Machine translation researchers might find opportunities to expand the models available to Wikipedians for translating articles into their language. In general, building connections with the Wikimedia community as described in Section 4.1 will help to identify opportunities for technical contributions to the Wikimedia projects.

ACKNOWLEDGMENTS

We would like to thank Taryn Bipat and Jacob Thebault-Spieker, whose research and presentations at the Wikimedia Research Showcase motivated this work. We also would like to thank the Wikimedia Foundation Research Team for their input and discussions as well as the many researchers and Wikimedians whose work and discussions we are building on.

REFERENCES

Milad Alshomary, Michael Völkske, Tristan Licht, Henning Wachsmuth, Benno Stein, Matthias Hagen, and Martin Potthast. Wikipedia text reuse: Within and without. In European Conference on Information Retrieval, pp. 747–754. Springer, 2019.

Pablo Aragón, Vicenç Gómez, and Andreaks Kaltenbrunner. To thread or not to thread: The impact of conversation threading on online discussion. In Proceedings of the International AAAI Conference on Web and Social Media, volume 11, pp. 12–21, 2017.

Akhil Arora, Martin Gerlach, Tiziano Piccardi, Alberto García-Durán, and Robert West. Wikipedia reader navigation: When synthetic data is enough. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM ’22, pp. 16–26, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498496. URL https://doi.org/10.1145/3488560.3498496

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: bridging the wikipedia language gap. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1075–1084, 2012.

Amber Berson, Monika Sengul-Jones, and Melissa Tamani. Unreliable guidelines: Reliable sources and marginalized communities in french, english and spanish wikipedias, 2021. URL https://artandfeminism.org/initiatives/current-initiatives/reading-together/

Pablo Beytía, Pushkal Agarwal, Miriam Redi, and Vivek K Singh. Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages. Sixteenth International AAAI Conference on Web and Social Media, 2022.

Taryn Bipat, David W McDonald, and Mark Zachry. Do we all talk before we type? understanding collaboration in wikipedia language editions. In Proceedings of the 14th International Symposium on Open Collaboration, pp. 1–11, 2018.

Taryn Bipat, Negin Alimohammadi, Yihan Yu, David W McDonald, and Mark Zachry. Wikipedia beyond the english language edition: How do editors collaborate in the farsi and chinese wikipedias? Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):1–39, 2021.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963, 2021.

38 https://ai.googleblog.com/2021/09/announcing-wit-wikipedia-based-image.html

39 https://www.mediawiki.org/wiki/Content_translation#How_to_participate
Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodriguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroa, Desi Rodriguez-Lonebear, Robyn Rowe, et al. The care principles for indigenous data governance. *Data Science Journal* **19**(1), 2020.

Xiaoxi Chelsy Xie, Isaac Johnson, and Anne Gomez. Detecting and gauging impact on wikipedia page views. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 1254–1261, 2019.

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

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021.

Anjali Field, Chan Young Park, and Yulia Tsvetkov. Controlled analyses of social biases in wikipedia bios. *arXiv preprint arXiv:2101.00078*, 2020.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

R Stuart Geiger. Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 17(3):342–356, 2014.

R Stuart Geiger and Aaron Halfaker. When the levee breaks: without bots, what happens to wikipedia’s quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, pp. 1–6, 2013.

Satdeep Gill. Balinese wikisource (wikipustaka) gives a new life to palm-leaf manuscripts, Nov 2021. URL https://diff.wikimedia.org/2021/11/15/balinese-wikisource-wikipustaka-gives-a-new-life-to-palm-leaf-manuscripts/.

Lennart Gulbransson. Swedish wikipedia surpasses 1 million articles with aid of article creation bot, Jun 2013. URL https://diff.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. Wiki-40b: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2440–2452, 2020.

Aaron Halfaker and R Stuart Geiger. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–37, 2020.

Shiqing He, Allen Yilun Lin, Eytan Adar, and Brent Hecht. The _tower_of_babel.jpg: diversity of visual encyclopedic knowledge across wikipedia language editions. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*, pp. 11–20, 2009.

Molly G Hickman, Viral Pasad, Harsh Kamalesh Sanghavi, Jacob Thebault-Spieker, and Sang Won Lee. Understanding wikipedia practices through hindi, urdu, and english takes on an evolving regional conflict. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31, 2021.
Benjamin Mako Hill and Aaron Shaw. Consider the redirect: A missing dimension of wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, pp. 1–4, 2014.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. Wikiconv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

Isaac Johnson. Analyzing wikidata transclusion on english wikipedia. In *Companion Proceedings of the 20th International Conference on Semantic Web*, 2020.

Isaac Johnson, Martin Gerlach, and Diego Sáez-Trumper. Language-agnostic topic classification for wikipedia. In *Companion Proceedings of the Web Conference 2021*, pp. 594–601, 2021a.

Isaac Johnson, Florian Lemmerich, Diego Sáez-Trumper, Robert West, Markus Strohmaier, and Leila Zia. Global gender differences in wikipedia readership. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pp. 254–265, 2021b.

Isaac L Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*, pp. 13–25, 2016.

Brian Keegan and Casey Fiesler. The evolution and consequences of peer producing wikipedia’s rules. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pp. 112–121, 2017.

Suin Kim, Sungjoon Park, Scott A Hale, Sooyoung Kim, Jeongmin Byun, and Alice H Oh. Understanding editing behaviors in multilingual wikipedia. *PloS one*, 11(5):e0155305, 2016.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*, 2021.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. Why the world reads wikipedia: Beyond english speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 618–626, 2019.

Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. Analysis of references across wikipedia languages. In *International Conference on Information and Software Technologies*, pp. 561–573. Springer, 2017a.

Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. Relative quality and popularity evaluation of multilingual wikipedia articles. volume 4, pp. 43. Multidisciplinary Digital Publishing Institute, 2017b.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3291–3301, 2019.

Katherine Maher. Wikipedia is a mirror of the world’s gender biases, 2018. URL https://wikimediafoundation.org/news/2018/10/18/wikipedia-mirror-world-gender-biases/.

Benjamin Mako Hill, Kaylea Champion, Aaron Shaw, Jeremy Foote, Nora McDonald, Andrea Forte, Sneha Narayan, and Rachel Greenstadt. Value of ip editing, 2019. URL https://meta.wikimedia.org/wiki/Research:Value_of_IP_Editing.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Marc Miquel-Ribé and David Laniado. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pp. 620–629, 2019.
Marc Miquel-Ribé and David Laniado. The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In Proceedings of the 16th International Symposium on Open Collaboration, pp. 1–4, 2020.

Blagoj Mitrevski, Tiziano Piccardi, and Robert West. Wikihist.html: English wikipedia’s full revision history in html format. In Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pp. 878–884, 2020.

Jonathan T Morgan, Michael Gilbert, David W McDonald, and Mark Zachry. Project talk: Coordination work and group membership in wikiprojects. In Proceedings of the 9th International Symposium on Open Collaboration, pp. 1–10, 2013.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsudeen Muhammad, Salomon Kabongo Kabenamulu, Salomey Osei, FRESHIA SACKEY, et al. Participatory research for low-resourced machine translation: A case study in african languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2144–2160, 2020.

Uzoma Ozurumba. Content translation tool helps create one million wikipedia articles, Nov 2021. URL https://diff.wikimedia.org/2021/11/16/content-translation-tool-helps-create-one-million-wikipedia-articles/.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(11):100336, 2021.

Tiziano Piccardi and Robert West. Crosslingual topic modeling with wikidata. In Proceedings of the Web Conference 2021, pp. 3032–3041, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392, 2016.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. A taxonomy of knowledge gaps for wikimedia projects (second draft). arXiv preprint arXiv:2008.12314, 2020.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15, 2021.

Yumiko Sato. Non-english editions of wikipedia have a misinformation problem. Slate, Mar 2021. ISSN 1091-2339. URL https://slate.com/technology/2021/03/japanese-wikipedia-misinformation-non-english.html.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1351–1361, 2021.

Shilad W Sen, Heather Ford, David R Musicant, Mark Graham, OS Keyes, and Brent Hecht. Barriers to the localness of volunteered geographic information. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 197–206, 2015.

Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. The wisdom of polarized crowds. Nature human behaviour, 3(4):329–336, 2019.
Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990, 2022.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2443–2449, 2021.

Nathan TeBlunthuis, Aaron Shaw, and Benjamin Mako Hill. Revisiting "the rise and decline" in a population of peer production projects. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–7, 2018.

Jakob Voss. Measuring wikipedia. 2005.

Jennifer Wang Wang, Sandister Tei, and Niharika Kohli. Impact report of turning off ip editing on portuguese wikipedia, 2021. URL https://meta.wikimedia.org/wiki/IP_Editing:_Privacy_Enhancement_and_Abuse_Mitigation.

Morten Warncke-Wang, Anuradha Uduwage, Zhenhua Dong, and John Riedl. In search of the ur-wikipedia: universality, similarity, and translation in the wikipedia inter-language link network. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, pp. 1–10, 2012.

Trust & Safety Disinformation Team Wikimedia Foundation. The case of croatian wikipedia: Encyclopaedia of knowledge or encyclopaedia for the nation?, 2021a. URL https://meta.wikimedia.org/wiki/Croatian_Wikipedia_Disinformation_Assessment-2021.

Trust & Safety Disinformation Team Wikimedia Foundation. Universal code of conduct/research - wikipedia, 2021b. URL https://meta.wikimedia.org/wiki/Universal_Code_of_Conduct/Research_-_Wikipedia#Policy_links,_description,_and_uniqueness_rating.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? In Proceedings of the 5th Workshop on Representation Learning for NLP, pp. 120–130, 2020.