The conundrum of using Wikipedia as training set data

Kimberli Mäkäräinen [1, 2, 3], Jack Rueter [4], Trond Trosterud [5]

[1] Wikimedia Finland (WMFI), [2] Wikimedia Language Diversity Hub, [3] Wikimedia Norway (WMNO), [4] University of Helsinki, [5] UiT The Arctic University of Norway

1 INTRODUCTION

"Harness the power of Wikipedia to train an LLM, create tools, and save an indigenous language from extinction!"*

While Wikipedia might seem like a good place to harvest text from in indigenous or minoritized languages, not all of the text in Wikipedia articles is of good quality, particularly those in indigenous and minority languages (Wilson 2019, Trosterud 2021, Wiechetek et al 2024).

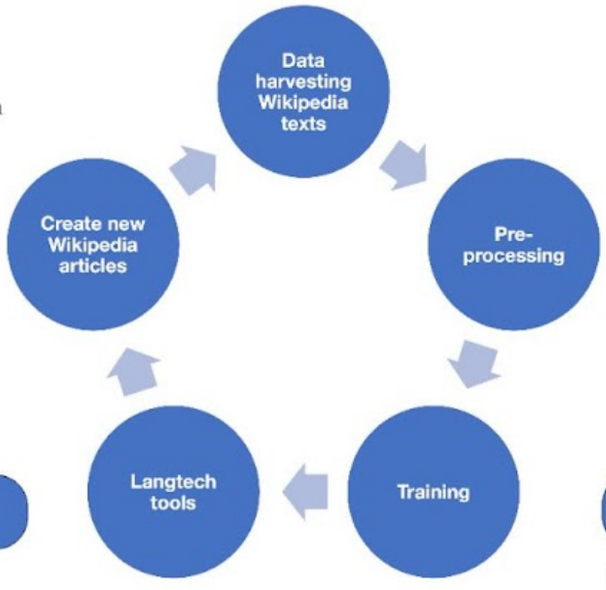
Another issue is that the people harvesting these texts most likely do not know these languages at all (Scannell 2024, Wiechetek et al 2024). This can lead to accuracy issues in text input as well as an inability to recognize potential data-quality problems and bad performance.

When low-quality text is then used to train language models and to create tools, it can create an unethical, vicious circle where these new tools are then used to create even more Wikipedia articles in these languages that are then reused as training set data to improve now existing tools (Wilson 2019, Scannell 2024, Wiechetek et al 2024).

2 QUALITY ISSUES

There are a number of reasons for the quality issues in Wikipedia articles, including the following:

1. Orthographical issues
2. Multiple languages/dialects in one Wikipedia
3. Editors from language communities without a written tradition
4. Editors who use various tools to create articles in languages they do not know at all



Each step is a potential point of entry for more low- to no-quality text that will then be (re)cycled through the process creating data of even worse quality

Circumpolar Wikipedias (Trosterud 2021)

- Mostly small projects
- The majority of these are "created and written by people who do not know the language and who do not belong to the language community."

Greenlandic Wikipedia

- Under threat of being closed down due to a lack of contributors, particularly those who know the language
- Contributors using Google Translate exacerbating the issue of poor quality
- "From my point of view, this fake Translation is even worse than 'vandalism', as nobody from outside can detect it, and it might have [sic] bad influence on the Greenlandic language." (klwiki)

3 PHYSICAL ISSUES

1. Orthographical issues can include the following:
   - Characters indistinguishable to most people
     - Ö ö (U+04E6 U+04E7)
       Ö ö (U+00D6 U+00F6)
     - Combining and precomposed characters
   - Unicode issues
   - Writing norms (Komi -иа vs. -ия)
   - Regurgitated misspellings: Če′vetjäu′rr (correct), Če'vetjäu'rr (the Finnish NLS Geographic Names Register!), Če´vetjäu´rr

2. Multiple languages/dialects in one Wikipedia
   - Akan Wikipedia (akwiki)
     - Mainly Twi, which also had its own Wikipedia at the same time
     - Prevented the other language communities from having Wikipedias
   - Norman Wikipedia (nrmwiki)
     - Has articles coexisting in 8 variants of the Norman language

3. Contributors who do not necessarily know how to write their own language for whatever reason
   - Not taught in schools
   - Often the consequence of colonialism and/or assimilation policies

4 WIKIHIJACKERS

The actions of *Wikihijackers* (editors who do not actual know the language they are creating articles in) can and will contaminate datasets to the point where the datasets can be effectively useless. AI and MT have now started to play a role in this as people are using them to create articles in the languages these tools are available in.

Irish Wikipedia (Scannell 2024)

- Stub articles mainly created by 2 contributors with an unsatisfactory knowledge of Irish
- Articles considered to be of such poor quality that virtually none of them are included in the Fiontar corpus

Scots Wikipedia

- > 20,000 articles created by 1 contributor who did not know the language and was not part of the language community
- The language community was forced to delete or clean up this contributor's and other similar contributors' mess

5 NOW WHAT

Not knowing the language will inevitably end up with researchers and big actors creating low- to no-quality datasets and tools. This is unethical (Wiechetek et al 2024) and can undermine the community's own work, including language revitalization. Poor quality tools can be worse than having no tools at all.

Researchers and big actors must work with and listen to the language communities as well as respect their wishes.

Indigenous data sovereignty and governance

- Te Hiku (Māori)
- CARE principles for indigenous data governance (Moshagen et al 2024)

REFERENCES

akwiki = Sandiooses and Robertjamal12 (2023). "Closing Akan Wikipedia: not a goodbye, but a hello". Accessed at https://diff.wikimedia.org/2023/04/20/closing-akan-wikipedia-not-a-goodbye-but-a-hello/

klwiki = Proposals for closing projects/Closure of Greenlandic Wikipedia. Meta-Wiki

Moshagen, S. et al (2024). "Indigenous language technology in the age of machine learning." *Acta Borealia*, 41(2), pp. 102–116.

Scannell, K (2024). *Minority Language Wikipedias in an AI-Dominated World*. Celtic Knot Conference 2024.

Trosterud, T. (2021). *The Circumpolar Wikipedia Editions*. Arctic Knot Conference 2021.

Wiechetek, L. et al (2024). "The Ethical Question–Use of Indigenous Corpora for Large Language Models." *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* pp. 15922–15931.

Wilson, K (2019). "Wikipedia has a Google Translate problem". The Verge.

LT4All 2025
Language Technologies for All