The Influence of Multilingualism and Mutual Intelligibility on Wikipedia Reading Behaviour

A Research Proposal

Meier, Florian Maximilian

Published in:
Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities

DOI (link to publication from Publisher):
10.5283/epub.44937

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

The Influence of Multilingualism and Mutual Intelligibility on Wikipedia Reading Behaviour

A Research Proposal

Florian Meier

Science, Policy and Information Studies
Department of Communication and Psychology
Aalborg University, Copenhagen, Denmark
fmeier@hum.aau.dk

Abstract

Given the important role of Wikipedia in our everyday lives, a better understanding of how language skills affect Wikipedia usage is needed. If content is not available in a reader's native language or a language that she can readily understand, access barriers and knowledge gaps are created, threatening Wikimedia's goal to create knowledge equity among all its projects and their consumers. This article argues for research on the effects of multilingualism and mutual intelligibility on Wikipedia reading behaviour, focusing on the Nordic countries, Denmark, Norway, and Sweden. Initial exploratory analysis shows that while residents of these countries use the native language editions quite frequently, they rely strongly on English Wikipedia, too. Research questions and methods for future work in this area are presented.

Keywords: multilingualism; mutual intelligibility; information behaviour; Wikipedia

1 Introduction and state of knowledge

With an average of around 1.4 billion page views per month,[1] Wikipedia is the most popular source of encyclopedic knowledge on the web. Although

1 https://stats.wikimedia.org/#/all-projects

Wikipedia is used in many countries and native language editions play an important role for local populations (Johnson, 2019), a considerable proportion of Wikipedia studies puts a strong focus on English Wikipedia. This is criticized by researchers, as recent studies found that English Wikipedia is an outlier regarding several aspects and not representative of the many language editions (Lemmerich et al., 2019). For example, why and how users read Wikipedia (i.e., their motivations, information needs prior knowledge and resulting usage behaviour), differs significantly between countries and language editions (Singer et al., 2017; Lemmerich et al., 2019). The anglophonocentric bias is also mirrored in the many ways and strategies that have been developed and used to improve and extend Wikipedia and other Wikimedia sister projects. However, this bias possesses a threat to one of the 2030 Wikimedia Foundations strategic directions[2] which aims to operationalize and quantify knowledge equity in Wikimedia projects thereby building the basis for bridging knowledge gaps, giving people equal access to unbiased knowledge around the globe (Zia et al., 2019; Redi et al., 2020).

Knowledge gaps are defined as "disparities in participation or coverage of a specific group of readers, contributors, or content" (Redi et al., 2020, p. 4). One of the most well-studied knowledge gaps is the gender gap, which evolved over the years because editors are mostly white and male (Lam et al., 2011). This resulted in gender-skewed content with, e.g., less than 20% of biographies being about women (Wagner et al., 2016; Konieczny & Klein, 2018). A recent study documented the gender gap even among Wikipedia readers, finding that two-thirds of all readers are men (Johnson et al., 2020).

A one-size-fits-all approach developed solely on the basis of research on English Wikipedia is likely not enough to bridge knowledge gaps like the aforementioned gender gap. Complete knowledge equity requires a deeper understanding of the differences between Wikipedia language editions and what role language plays in Wikipedia reading behaviour in general. Unlike other user-generated content platforms, where the effects of multilingualism on usage behaviour have been studied before (Eleta & Goldbeck, 2014; Lee & Chau, 2018), little is known about the dynamics that are at play when Wikipedia consumers speak or are able to readily understand multiple languages. Multilingual editing behavior has been studied before (Hale, 2014; Kim et al., 2016). Hale (2014), for example, studied multilingual Wikipedia editors and found that they play a key role in information diffusion. Editors editing the

2 https://meta.wikimedia.org/wiki/Research:2030

same page in multiple languages leads to the reduction of self-focus bias by 25%.

To the best of our knowledge, multilingualism among readers hasn't been studied yet. Anecdotal evidence from a recent study combining survey and log data in 14 countries shows that up to 95% of participants reported being monolingual and relying on native language editions (Johnson, 2019). Exceptions are the English and French Wikipedia, which are essential for non-native speakers in Africa (ibid.). Moreover, this data showed that although the survey was administered in only 14 languages, participants viewed articles in over 100 languages during their reading sessions (Johnson et al., 2020). Finally, preliminary analysis of the same survey data "shows that on average roughly 20% of reader sessions involve the reader switching from one Wikipedia language to another" (Lemmerich et al., 2019, p. 625).

To sum up, while there seems to be evidence for within-session language switching, this phenomenon hasn't been studied in detail before. Additionally, there is no study that reports on across-session language switching, i.e., users alternating between various language editions depending on context factors like the type of information need or knowledge coverage. Previous research shows that Wikipedia and search engines form a tightly-knit relationship, with Wikipedia pages being highly ranked on Search Engine Result Pages (SERP) for many common queries. This implies that the way people search and which languages they use etc. also has a significant impact on how they read Wikipedia (Vincent et al., 2019).

This short article argues for the need of research on how multilingualism, i.e., speaking multiple languages, and mutual intelligibility, defined as "the fact that some language pairs are so closely related that the speakers are able to communicate each using their own language without prior language instruction" (Gooskens et al., 2017, p. 170) on Wikipedia usage behaviour. As mutual intelligibility is a phenomenon that is very pronounced in the Nordic countries, i.e., Denmark, Norway and Sweden (Gooskens et al., 2017), initial analysis as well as ideas for future work will focus on these countries and Wikipedia language editions.

2 Initial analysis: Wikipedia usage in the nordic countries

The Nordic countries' Wikipedia editions are good examples of diverse types of Wikipedia projects. While Danish and Norwegian Wikipedia are what one could call medium-sized with around 260 K and 535 K articles, Swedish Wikipedia features around 3.7 M articles and is among the biggest encyclopedic projects.[3] The reason for its size is that most articles are created by a bot called Lsjbot (Guldbrandsson, 2013). To study mutual intelligibility, we, in an initial exploratory analysis, looked at these language projects' page views and their origin meaning which countries the page views come from. We collected data via the Wikimedia REST API[4], which allows retrieving access-data for the last five years.

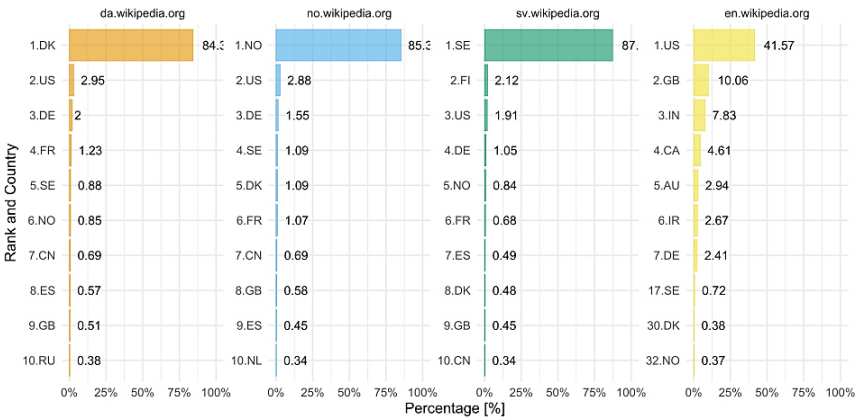

Fig. 1
Percentage of average page views per year coming from different countries for the three Nordic countries and English Wikipedia language editions (Data: 2016–2020)

Figure 1 visualizes the average percentage of page views for the Nordic countries' language projects and English Wikipedia. One can clearly see that most page views (84% or higher) come from within the country. Although, e.g., a Dane could easily read and understand a Norwegian or Swedish

3 https://wdo.wmcloud.org/cultural_context_content/

4 https://wikimedia.org/api/rest_v1/

Wikipedia article, signs that users are taking advantage of mutual intelligibility are very sparse. Only around 2% of access to the Danish Wikipedia comes from Sweden (SE: 0.88%) and Norway (NO: 0.85%). A similar ratio can be observed for the Norwegian and Swedish Wikipedia. A question that arises here is whether this traffic can be explained by, e.g., expatriates or tourists in these countries or whether there are other reasons that give an explanation of the, admittedly low, share of page views across countries and language editions. It is not surprising that English Wikipedia page views come from many different countries whereby most originate from the US, Great Britain and India. The statistic of most page views on en.wikipedia.org lists Sweden, Denmark, and Norway on rank 17, 30, and 32. Next, we tried to estimate the importance of English Wikipedia for the three Nordic countries. Earlier studies suggest that users mostly rely on the native language edition if one exists (Johnson, 2019), but what role does English Wikipedia play for languages/countries with quite large language editions? We estimated the importance of English Wikipedia by dividing the number of page views on en.wikipedia.org by the number of page views on the own language edition plus the number of page views on en.wikipedia.org. Figure 2 visualizes this share in a time-series analysis from January 2016 to October 2020.

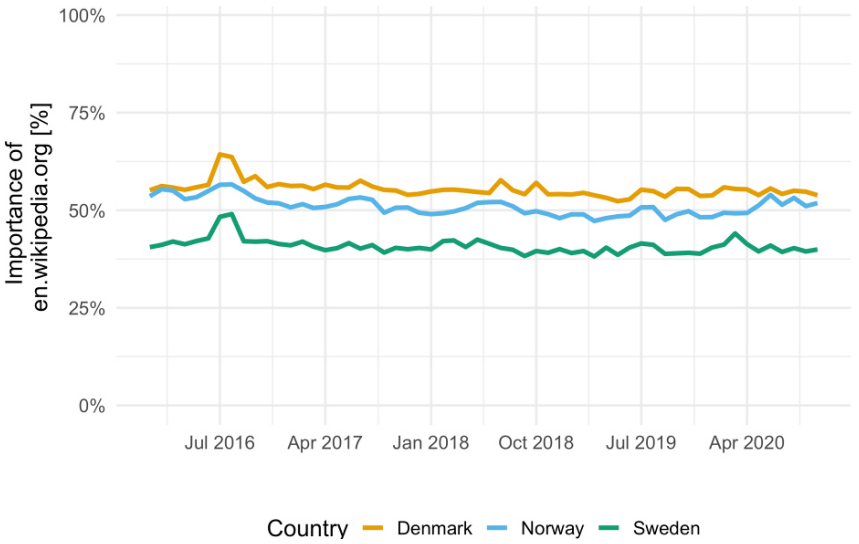

Fig. 2 Time series visualization of the importance of en.wikipedia.org for the three Nordic countries measured in percentage of page views (Data: 2016–2020)

The overall pattern is quite revealing. While the share of English Wikipedia accesses in Sweden lies constantly below 50% which means that the Swedes use sv.wikipedia.org more frequently than English Wikipedia, the Danes rely more strongly on the English Wikipedia. They use it more than their local language edition as is indicated by the line continuously lying above 50%. A possible explanation for this observation could be the smaller size of dk.wikipedia.org, and thus missing content that motivates users to look for information in other language projects. Observations like these are important as they help with prioritizing strategies for closing knowledge gaps.

3 Future work

Research on how language proficiency in multiple languages affects Wikipedia reading behaviour is anecdotal and hasn't been studied in depth. In the future, we want to investigate this phenomenon in detail, specifically focusing on the Nordic countries. The following questions will guide our research:

- How frequently do Wikipedia users make use of their language skills and switch between language editions?
- How proficient does a user need to be in a language so she would consider switching to it?
- What is the extent of within and across-session language switching?
- Which factors motivate this behaviour, i.e., what role do different types of information needs, prior knowledge or sociodemographic aspects play?
- How do the characteristics of Wikipedia language editions, i.e., size, quality of articles, the share of cultural context content, influence user choice (Miquel-Ribé & Laniado, 2020)?

To find answers to these questions, we will follow a mixed-methods approach combining a large-scale quantitative survey with a data science approach. While the survey will gain an understanding of the extent and characteristics of within and across-session language switching when reading Wikipedia, the data science approach will use the Wikipedia Cultural Diversity dataset (Miquel-Ribé & Laniado, 2019) and the Wikipedia Diversity Observatory[5] to investigate the importance of local content for Wikipedia readers

5 https://wdo.wmcloud.org/

in the Nordic countries to help develop strategies for closing language proficiency-related knowledge gaps.

4 Summary and conclusion

This short paper presented a research proposal to bridge knowledge gaps in the Wikipedia, specifically to expand on the availability of qualitative and quantitative evidence on how the nature, scope and impact of reader's language proficiency affect Wikipedia usage. Which and how many languages Wikipedia consumers can understand impacts considerably what knowledge is available to them. However, hardly anything is known about how multilingualism and mutual intelligibility affect Wikipedia information behaviour. This is a threat to one of Wikimedia foundations 2030 strategic goals which aim to create knowledge equity, i.e., to support "the knowledge and communities that have been left out by structures of power and privilege" (Zia, 2019, p. 1). As multilingualism and mutual intelligibility are phenomena that are widespread in the Nordic countries (Denmark, Norway and Sweden), and their Wikipedia language editions are interesting projects regarding their different sizes. At first, this project will focus on Wikipedia reading behaviour in these countries. Initial exploratory analysis showed that although people could make use of their language skills and look up information on either the Danish, Swedish or Norwegian Wikipedia, most page views (~85%) of the three language editions in question come from within the own country. However, all three rely quite strongly on en.wikipedia.org. For Danes and Norwegians, English Wikipedia is even more important than their native language edition. In the future, a mixed-methods approach combining a large quantitative survey and data analysis from the Wikipedia Diversity dataset will study the role of language in Wikipedia usage in greater detail and help with bridging possible knowledge gaps.

References

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior, 41*, 424–432, https://doi.org/10.1016/j.chb.2014.05.005

Gooskens, C., van Heuven, V. J., Golubović, J., Schüppert, A., Swarte F., & Voigt, S. (2018). Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism, 15*(2), 169–193, DOI: 10.1080/14790718.2017.1350185

Guldbrandsson, Lennart (2013). Swedish Wikipedia surpasses 1 million articles with aid of article creation bot. https://diff.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/

Hale, S. A. (2014). Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science (WebSci '14)*. New York, NY: ACM, 99–108. https://doi.org/10.1145/2615569.2615684

Johnson, I. (2019). Research: Characterizing Wikipedia Reader Behaviour/Demographics and Wikipedia use cases. https://meta.wikimedia.org/?curid=10664863

Johnson, I., Lemmerich, F., Sáez-Trumper, D., West, R., Strohmaier, M., & Zia, L. (2020). Global gender differences in Wikipedia readership. *ArXiv*. https://arxiv.org/pdf/2007.10403.pdf

Kim, S., Park, S., Hale, S., Kim, S., Byun J., & Oh, A. (2016) Understanding Editing Behaviors in Multilingual Wikipedia. *PLOS ONE, 11*(5): https://doi.org/10.1371/journal.pone.0155305

Konieczny, P., & Klein, M. (2018). Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society, 20*(12), 4608–4633. https://doi.org/10.1177/1461444818779080

Lam, S. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. New York, NY: ACM, pp. 1–10. https://doi.org/10.1145/2038558.2038560

Lee, C., & Chau, D. (2018). Language as pride, love, and hate: Archiving emotions through multilingual Instagram hashtags. *Discourse, Context & Media, 22*, 21–29, https://doi.org/10.1016/j.dcm.2017.06.002

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2019). Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. New York, NY: ACM, pp. 618–626. https://doi.org/10.1145/3289600.3291021

Miquel-Ribé, M., & Laniado D. (2019). Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. In *Proceedings of the 13th International AAAI Conference on Web and Social Media. ICWSM 2019*. [Palo Alto, CA]: Association for the Advancement of Artificial Intelligence.

Miquel-Ribé, M., & Laniado D. (2020). The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. In *Proceedings of the International Symposium on Open Collaboration (OpenSym 2020)*. New York, NY: ACM. 2 pages. https://doi.org/10.1145/3412569.3412866

Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2020). A Taxonomy of Knowledge Gaps for Wikimedia Projects (First Draft). *ArXiv*. https://arxiv.org/pdf/2008.12314v1.pdf

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J. (2017). Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. Geneva: International World Wide Web Conferences Steering Committee, pp. 1591–1600. https://doi.org/10.1145/3038912.3052716

Vincent, N., Johnson, I., Sheehan, P., & Hecht, B. (2019). Measuring the Importance of User-Generated Content to Search Engines. In *Proceedings of the International AAAI Conference on Web and Social Media, 13*(01), 505–516. https://ojs.aaai.org/index.php/ICWSM/article/view/3248

Wagner, C., Graells-Garrido, E., Garcia D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science, 5*(5). https://doi.org/10.1140/epjds/s13688-016-0066-4

Zia, L., Johnson, I., Mansurov, B., Morgan, J., Redi, M., Saez-Trumper, D., & Taraborelli, D. (2019). Knowledge Gaps – Wikimedia Research 2030. doi.org/10.6084/m9.figshare.7698245

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 64–72. DOI: doi.org/10.5283/epub.44937.