Network perspectives on collective memory processes across the Arabic and English Wikipedias

Laurie Jones
Dept. of Information Science
University of Colorado Boulder

Brian C. Keegan
Dept. of Information Science
University of Colorado Boulder

Alexandra Siegel
Dept. of Political Science
University of Colorado Boulder

Abstract

The Arab Spring was a historic event that had a major impact on several countries in the Middle East and North Africa. While there has been extensive research on the role of computer-mediated communication tools on the Arab Spring, there has been limited analysis of the collective memory processes of the event across different languages. We propose an investigation to understand the divergent collective memory processes of the Arab Spring by analyzing the English and Arabic version of the Wikipedia articles on the event. We will use a combination of statistical analysis, natural language processing, and network analysis to understand how the articles have evolved over time and across languages. Our analysis will explore the variance across time, space, and language utilizing language-specific tools. We will attempt to engage current and previous Wikipedia editors of these articles to contextualize changes in article' content and temporal variation.

Introduction

Arabic is a major international language with 5.1% of the world using it as a first language. However, it is poorly represented on Wikipedia with 5,899 active users for 1.2 million articles compared to 123,291 active users for 6.5 million articles on English Wikipedia even though it has a similar 5.1% first language rate. There has been work describing the discrepancies between language editions on Wikipedia [9,19], but very few papers include Arabic in their analysis despite its popularity and impact. Wikipedia is at the forefront of documenting current events and these accounts continue to evolve over time. Even with Wikipedia's strive toward neutrality there are still gaps across language editions, ranging from complete topics missing to asymmetries within the content of articles. There are articles that people have just directly translated from one language to another in an attempt to address this imbalance but machine translation for low-resource languages such as Arabic can be inaccurate. Bridging these gaps and creating a more encompassing understanding of global phenomena by understanding and addressing the cross-lingual divides in under-resourced languages like Arabic should be a high priority.

The Arab Spring was a historic set of conflicts, protests, and movements that changed the histories of governments, countries, and the region. Its use of internet-based communication also redefined collective action for the modern world [22]. Unlike a traditional paper encyclopedia, Wikipedia articles about these events were written contemporaneously with the events themselves. Wikipedia's revision history preserves these earlier encyclopedic descriptions of the events and captures the fine-grained evolution of every change since then. This archive of evolving content, potentially captures shifting narratives and framings from the immediate context of the events to its stabilization years later. Wikipedia's multi-lingual quality also provides a unique ability to address these evolving narratives within linguistic communities and perhaps identify topics or themes that are linguistically siloed. More than a decade after these Arab Spring conflicts, how are these events remembered across languages?

The structure of Arab Spring content on Wikipedia is rooted in the "Arab Spring" articles. These abridged articles provide a summary of the various protests, conflicts, and revolutions within the Arab Spring phenomenon and then links to separate country-based articles for each conflict. Variations between the linguistic versions appear even within the asymmetry of these country-based article titles. For example, the main links within the English "Arab Spring" article for the events within Egypt are the "2011 Egyptian

Revolution”, and “Egyptian Crisis (2011–2014)”. Within the Arabic Arab Spring article, the only link associated with the events in Egypt is, “ثورة 25 يناير”. This directly translates to “The January 25th revolution”. It is the interlingual link for the “2011 Egyptian Crisis” but the interlingual link or counterpart for “Egyptian Crisis (2011–2014)” is not in the Arabic “Arab Spring” article at all or any other related counterpart. This can also be seen with the events in Yemen. The article linked within the broader English “Arab Spring” article cites two articles for Yemen: “Yemeni Revolution” and “Yemeni Crisis (2011–present)”. However, on the Arabic Arab Spring article, the only link present concerning the events in Yemen is “ ثورة الشباب اليمنية” which directly translates to, ‘The Yemeni Young People’s Revolution” and the interlingual link for the “Yemeni Crisis (2011–present)”, "(الأزمة اليمنية 2011-الآن) ” is not linked within the Arabic Wikipedia article. Within these two examples we can already see inconsistencies between how these events are presented and contextualized within the larger perspective of the Arab Spring, each country’s history, and across Wikipedia. Computational methods from natural language processing and network analysis provide tractable strategies for analyzing this data for changing patterns and structures.

This project will use a combination of within- and between-subjects designs to analyze articles about the 2011 Arab Spring primarily across English and Arabic to understand distinct collective memory processes. We will leverage several types of variance to understand these dynamics: variance over time (what has changed between 2011 and 2023?), variance across space (how similar are articles about different countries’ conflict?), and variance across language (how similar are articles about similar topics across language editions?). This project will use quantitative methods from natural language processing (leveraging our teams’ linguistic familiarity with Arabic as well as Arabic-specific computational tools) and network analysis to analyze articles’ revision history data to measure similarity and identify hyperlink and collaborative structures. We will also explore the possibility of employing causal inference methods (difference in differences, matching, etc.) to estimate the effects of articles’ changing content and relationships over time on outcomes like editor retention, contributions, and conflict. While the Arab Spring had impacts on countries within and beyond the Arab world, we will focus on the countries that are considered “revolutions” or “الثورات“ in the Arabic Arab Spring article: Tunisia, Egypt, Libya, Yemen, and Syria. This set of countries will form the “country level” seeds from which will snowball to analyze additional articles and editors impacted by the larger Arab Spring.

Date: This project would start at the beginning of the school year at the University of Colorado Boulder, August 1, 2023 and would conclude by June 30, 2024.

Related work

As a platform of collaborative knowledge production and sharing, Wikipedia has been used by researchers to study collective memory construction, the reliability of the narratives produced in this collaboration, and the knowledge gaps that have been produced across languages [17]. Some of the earliest multilingual research on Wikipedia identified a self-focus bias within community maintained information structures on Wikipedia where articles that were spatially closer to the population of language speakers were more detailed than articles about more distant topics [9]. Massa and Scrinzi 2012 identified varying perspectives within linguistic groups, identifying a Linguistic Point of View (LPOV) [16]. This variation in information is then identified to transcend not only language but culture [3]. When language and culture are at the center of the information being discussed for example in more sensitive topics such as conflict, this knowledge asymmetry becomes clear [10,20]. This asymmetry however might not be unintentional [7].

The protests, movements, revolutions, and wars that unfolded through 2011 demonstrate that the Arab Spring is far from settled history or stable knowledge [6]. There has been work looking at individual event articles but the Arab Spring was a phenomenon where actions within multiple countries were inspired by actions in others. The histories and narratives of this phenomenon are intertwined and a more

comprehensive look comparing multiple articles within the Arab Spring, as well as adjacent people, organizations, and events are important to map out the collective memory processes. Elements of the Arab Spring have discrepancies between languages [1,4,5] but an in-depth comparative analysis of the linguistic and collaborative dynamics across time, space, and language remains missing. Developing a more encompassing analysis using consistent metrics would identify the asymmetries and gaps in content that are consistent between Arabic and English.

We hypothesize that divergent framings of conflict emerge and stabilize relatively early in articles' histories and opportunities for changing this consensus only happen after new events bring in new editors. Leveraging relational perspectives on collaboration, we expect that articles that are embedded within more strongly overlapping links of editor-article coauthorship are likely to have more similar content. These processes are likely to be dynamic where articles with increasing embeddedness are likely to become more similar in content and *vice versa*. Identifying the structures and dynamics of Wikipedia's coverage of events [11,12,13,18], network perspectives will be leveraged to reach a more multi-dimensional understanding of the major historical event and the cooperation of editors to create a Wikipedia page. The changing content of articles can be triangulated with content consumption data (pageviews, clickstream, *etc*.) to illustrate collective memory processes like reappraisal and migration [23]. The diversity of data sources and multiple sources of variation has powerful potential to lend insight into collective memory processes like narrative evolution.

Methods

Networks of editor interaction and collaboration are examples of the collective memory processes by mapping the discussion and cooperation that is negotiated within the discussion pages of Wikipedia [6]. The activity of the editors captures distinct power dynamics and governance processes that could be prevalent in the documentation of the Arab Spring within and across languages. These two linguistic communities employ different practices and framings to describe the events, link to related topics, and cite reliable sources. Based on preliminary analyses, there is a lack of overlapping links between the respective English and Arabic versions of "Arab Spring" Wikipedia articles.

Within our project we will initialize our analysis with basic statistical comparison such as length, pageviews, edit history, and editor statistics as well as identifying image discrepancies across language. This grounds our work within the methods of the previous work [5,8,18,19] for a more generalizable addition to the literature. We will analyze variance in articles across time, space and language using natural language processing tools and network analysis methods. Many types of networks can be constructed depending on the needs of our research design: collaboration networks of editors contributing to articles, hyperlink networks of articles linking to articles, citation networks of articles linking to sources, and discussion networks of editors replying to editors. Networks that inspect the evolution of internal Wikipedia links and revision history will give insight into how the main Arab Spring articles in English and Arabic have changed between 2011 and 2023.

Countries that are considered "revolutions" or "الثورات" on the Arabic Arab Spring page: Tunisia, Egypt, Libya, Yemen, and Syria, will support networks that explore how the "Arab Spring" articles have evolved over time and how they are linked within the collection of articles. This contextualizes the Arab spring as a multi-national phenomenon and allows for a more generalizable comparison across these linked events while providing additional insights into event level analysis. Focusing on 'revolutions' opposed to all of the events of the Arab Spring allows us to address topics that would have the most contemporary relevance, largest impact across both linguistic communities, and prompt further research into country based analysis that would account for country based intricacies such as dialect and specific history.

Continuing to compare across English and Arabic, we will use language-based tools such as AraBERTv2 for Arabic language and BERT for English text, to determine asymmetry in topic coverage. Arabic is a language with a lot of nuance within its grammatical

structure making it difficult for NLP models to identify context, but tools like AraBERTv2 are performant for analyzing and categorizing contemporary online Arabic language [2]. This also supports future research on low-resource languages such as Arabic by providing examples and tests for low-resource built tools.

While our research designs are primarily comparative (across time, space, and language) and our methods are primarily descriptive, the exogenous disruptions characterizing many of the events around the Arab Spring both potentially lend themselves to leveraging causal inference research designs. Difference in differences, interrupted time series, and matching designs are all potential methods for evaluating the impact of participating in a collaboration or introducing a type of content, for example.

Acknowledging that collective memory is community based and the construction of narratives is rooted in its editors [1], these quantitative analyses will be contextualized by interviewing previous and current editors within these articles. These semi-structured interviews will use retrospective methods [21] to ask participants to discuss their choices about their revisions, contextualize disputes, and understand their motivations. These interviews will be conducted using video conferencing tools like Zoom or Skype and will be recorded, transcribed, and thematically coded. While analysis of archival data in previous elements of this project will not use human subjects data, these interviews and analysis will be submitted for approval by our Institutional Review Board.

Validating our output, we will partner with other researchers who have Arabic as their primary language and people that have grown up within the region. We will also collaborate with these researchers after key milestones in our analysis to ensure that our conclusions are substantive and thoughtful of all of the variables that should be taken into account. This step is crucial for us to present an output that we believe would be considerate of both linguistic communities we are engaging.

Expected output

This project will produce three outputs. The first output will be a bilingual report intended for a general Wikipedia audience summarizing and contextualizing our findings. This report will combine case studies, descriptive findings and visualizations, and the interview data from contributors. Drafts of the report will be shared via the Village Pump and WikiProject Intertranswiki to solicit feedback before finalizing the report. By engaging the editors in a dialogue for this report's creation, we hope to also bridge offline linguistic communities and participate in the global Wikipedia landscape.

The second output will be at least one paper manuscript analyzing the networked dimensions of cross-language collective memory processes. The manuscript will identify common misalignments across time, country, and language and develop a generalizable framework for classifying these misalignments. Validating theories such as content asymmetry, and temporal, spatial, and language variance, would provide an avenue for future causal inference literature, establishing a link between offline events and online narrative curation and evolution. This work would do so within a global context, investigating the nature of linguistic perspectives and cross-lingual ties or silos. This manuscript will be submitted for peer review and open-licensed publication at a venue like ACM Computer-Supported Cooperative Work and Social Computing (CSCW), AAAI International Conference on Web and Social Media (ICWSM), or the Web Conference (WWW). One salient implication of this research will be to assess the risks of biases, asymmetries, and misalignments being amplified into other algorithmic systems like machine translation, chat assistants, and large language models which we would reflect on in this manuscript.

The third output will be the development and synthesis of open source computational tools for retrieving, parsing, and comparing archival Wikipedia revision history, content, and pageview data. These tools will leverage existing Wikimedia Foundation data services like the MediaWiki API, Quarry, PAWS, and/or Toolforge and will be released

as reproducible computational notebooks that integrate code, figures, and documentation. We expect that these tools will be usable by analysts, researchers, community members, journalists, and the general public. These tools will be shared with the general and research community via announcements to community boards and listservs.

Risks

The primary risks with this project involve the gaps in linguistic and cultural knowledge between the research team and the Arabic Wikipedia community. The graduate student, Laurie Jones, has been studying both Levantine Arabic and MSA for five years and lived in Jordan for several months. Co-PI Alexandra Siegel has strong linguistic and cultural knowledge. Siegel's research focused on the social media dynamics of political communication in the Arab world, has spoken Arabic for over a decade, and lived in Cairo in spring 2011. Co-PI Keegan provides expertise in English Wikipedia practices and governance as a 17-year editor with over 10,000 edits and approximately a dozen peer-reviewed publications about Wikipedia. The planned interviews with editors of these articles will also provide important context that we may miss as Western researchers. We also plan to build off of our team's collective knowledge by collaborating with other researchers who do have Arabic as their first language and people who were raised in the region to validate our analysis. This is important for us as we acknowledge even with our Arabic language comprehension, our team still has a bias due to our western upbringing.

Community impact plan

Our findings comparing collective memory systems have implications for AI systems, low-resource NLP tools, the Wikipedia community, as well as general popular knowledge.

Machine translation, language models such as LLMs and other text-based learning algorithms are often trained on Wikipedia as a neutral source that crosses linguistic communities. Previous research has shown that there is a bias within these linguistic versions and our work would prove this on a digestible scale between two of the most spoken linguistic communities in the world.

Our work prioritizes language based Natural Language Processing tools and since Arabic is considered a low-resource language, this work adds to NLP literature, providing a further evaluation of these tools. We would also be providing an investigation into these Wikipedia pages as possible data sources for future tools to be trained on.

This work also addresses the Wikimedia Movement's 2030 strategic direction. It identifies, measures, and visualizes the multidimensional aspect of knowledge gaps. Our findings will promote a connection between individual editors and multilingual users to build a trusted environment for more sustainable knowledge sharing and collaboration. This would be further addressed through work with Wikimedia affiliates and user groups to identify other knowledge gaps and work with Wikimedia volunteer editor communities to bridge the understanding between these gaps and create a more neutral narrative built through collaborative ties that can lead to more long-term decision making about this dilemma.

The increased use of the internet as well as Wikipedia globally has created a false sense of information homogeneity within broader information consumers. This project would provide an analysis on a well-known topic with a specific scope and generalizable analysis that can be easily translated into a more accessible piece of literature outside of a scientific publication such as a blog post or commentary piece.

Evaluation

The success of this project is due to its thoughtfulness in inclusion of not only a language that is often not included in multi-lingual analysis, but the use of tools based in that language opposed to imposing non-language specific tools on this low-resource language. It would also be due to its inclusion of both Arabic and English speakers, emphasizing bridging these two linguistic communities and promoting a global dialogue about memory curation and

evolution. This project's breadth of events within the Arab Spring utilizing a consistent mode of analysis between these events generates a research pipeline that builds off previous literature and establishes a standard for future work concerning collective memory analysis on larger phenomena.

Evaluation should be done on the multi-modal aspects of analysis that we have included to lend insight into the collective memory narrative and the usefulness of this project for people across computational specialties and language communities.

References

[1] K. Al-Shehari and A. G. Al-Sharafi, "Negotiating Wikipedia narratives about the Yemeni crisis: Who are the alleged supporters of the Houthis?," *Media, War & Conflict*, vol. 15, no. 2, pp. 183–201, Jun. 2022, doi: 10.1177/1750635220938404.

[2] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding." arXiv, Mar. 07, 2021. doi: 10.48550/arXiv.2003.00104.

[3] E. S. Callahan and S. C. Herring, "Cultural bias in Wikipedia content on famous persons," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 10, pp. 1899–1915, 2011, doi: 10.1002/asi.21577.

[4] M. Ferron and P. Massa, "The Arab Spring| WikiRevolutions: Wikipedia as a Lens for Studying the Real-Time Formation of Collective Memories of Revolutions," *International Journal of Communication*, vol. 5, no. 0, Art. no. 0, Sep. 2011.

[5] M. Ferron and P. Massa, "Collective memory building in Wikipedia: the case of North African uprisings," in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, in WikiSym '11. New York, NY, USA: Association for Computing Machinery, Oct. 2011, pp. 114–123. doi: 10.1145/2038558.2038578.

[6] H. Ford, "Writing the Revolution," *MIT Press*.

[7] J. Grabowski and S. Klein, "Wikipedia's Intentional Distortion of the History of the Holocaust," *The Journal of Holocaust Research*, vol. 0, no. 0, pp. 1–58, Feb. 2023, doi: 10.1080/25785648.2023.2168939.

[8] S. He, A. Y. Lin, E. Adar, and B. Hecht, "The tower of babel.jpg: 12th International AAAI Conference on Web and Social Media, ICWSM 2018," *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, pp. 102–111, 2018.

[9] B. Hecht and D. Gergle, "Measuring self-focus bias in community-maintained knowledge repositories," in *Proceedings of the fourth international conference on Communities and technologies*, in C&T '09. New York, NY, USA: Association for Computing Machinery, Jun. 2009, pp. 11–20. doi: 10.1145/1556460.1556463.

[10] M. G. Hickman, V. Pasad, H. K. Sanghavi, J. Thebault-Spieker, and S. W. Lee, "Understanding Wikipedia Practices Through Hindi, Urdu, and English Takes on an Evolving Regional Conflict," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, p. 34:1-34:31, Apr. 2021, doi: 10.1145/3449108.

[11] B. Keegan, "Emergent Social Roles in Wikipedia's Breaking News Collaborations," in *Roles, Trust, and Reputation in Social Media Knowledge Markets: Theory and Methods*, E. Bertino and S. A. Matei, Eds., in Computational Social Sciences. Cham: Springer International Publishing, 2015, pp. 57–79. doi: 10.1007/978-3-319-05467-4_4.

[12] B. Keegan, D. Gergle, and N. Contractor, "Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes," in *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, in WikiSym '11. New York, NY, USA: Association for Computing Machinery, Oct. 2011, pp. 105–113. doi: 10.1145/2038558.2038577.

[13] B. Keegan, D. Gergle, and N. Contractor, "Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events,"

American Behavioral Scientist, vol. 57, no. 5, pp. 595–622, May 2013, doi: 10.1177/0002764212469367.

[14] J. Kubś, "Historical Narratives in Different Language Versions of Wikipedia," *Academic Journal of Modern Philology*, no. 12, pp. 83–94, 2021.

[15] A. Li, R. Farzan, and C. López, "Let's Work Together! Wikipedia Language Communities' Attempts to Represent Events Worldwide," *Interacting with Computers*, p. iwac033, Dec. 2022, doi: 10.1093/iwc/iwac033.

[16] P. Massa and F. Scrinzi, "Manypedia: comparing language points of view of Wikipedia communities," in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, in WikiSym '12. New York, NY, USA: Association for Computing Machinery, Aug. 2012, pp. 1–9. doi: 10.1145/2462932.2462960.

[17] C. Pentzold, "Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place," *Memory Studies*, vol. 2, no. 2, pp. 255–272, May 2009, doi: 10.1177/1750698008102055.

[18] E. Porter, P. M. Krafft, and B. Keegan, "Visual Narratives and Collective Memory across Peer-Produced Accounts of Contested Sociopolitical Events," *Trans. Soc. Comput.*, vol. 3, no. 1, p. 4:1-4:20, Feb. 2020, doi: 10.1145/3373147.

[19] D. Roy, S. Bhatia, and P. Jain, "A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in Low Resource Languages," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 2373–2380. Accessed: Mar. 28, 2023. [Online]. Available: https://aclanthology.org/2020.lrec-1.289

[20] D. Roy, S. Bhatia, and P. Jain, "Information asymmetry in Wikipedia across different languages: A statistical analysis," *Journal of the Association for Information Science and Technology*, vol. 73, no. 3, pp. 347–361, 2022, doi: 10.1002/asi.24553.

[21] D. Russell and E. Chi, "Looking Back: Retrospective Study Methods for HCI," 2014, pp. 373–393. doi: 10.1007/978-1-4939-0378-8_15.

[22] Z. Tufekci, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*, Reprint edition. Yale University Press, 2018.

[23] M. Twyman, B. C. Keegan, and A. Shaw, "Black lives matter in Wikipedia: 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017," *CSCW 2017 - Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1400–1412, Feb. 2017, doi: 10.1145/2998181.2998232.