

(Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories

Anna Samoilenko^{1,2}, Florian Lemmerich^{1,3}, Maria Zens¹, Mohsen Jadidi^{1,2},
Mathieu Génois¹, Markus Strohmaier^{1,3}

¹GESIS – Leibniz-Institute for the Social Sciences, ²University of Koblenz-Landau, ³RWTH Aachen University
firstname.lastname@gesis.org,firstname.lastname@humtec.rwth-aachen.de

ABSTRACT

In this paper we present a large-scale quantitative comparison between expert- and crowdsourced writing of history by analysing articles from the English Wikipedia and Britannica. In order to quantify attention to particular periods, we extract mentioned year numbers and utilise them to study historical timelines of nations stretched over the last thousand years. By combining this temporal analysis with lexical analysis of both encyclopedic corpora we can identify distinctive historiographic points of view in each encyclopedia. We find that Britannica focuses on social and cultural phenomena, e.g. religion, as well as the geographical characteristics of states, while Wikipedia puts emphasis on political aspects, concentrating on wars and violent conflicts, and events of high popularity. Finally, both encyclopedias exhibit characteristics of English Academic prose, with Britannica being slightly less readable compared to Wikipedia, according to several readability scores.

KEYWORDS

Computational history, Collective memory, Wikipedia, Britannica, Null Model, Focal points, Readability, Natural language processing

ACM Reference Format:

Anna Samoilenko^{1,2}, Florian Lemmerich^{1,3}, Maria Zens¹, Mohsen Jadidi^{1,2}, Mathieu Génois¹, Markus Strohmaier^{1,3}. 2018. (Don't) Mention the War: A Comparison of Wikipedia and Britannica Articles on National Histories. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3178876.3186132>

1 INTRODUCTION

The Encyclopedia Britannica is an important authoritative reference on a multitude of topics and subjects. Written by *experts*, it also provides extensive information on the history of countries. With the advent of the World Wide Web and collaborative technologies, Wikipedia has emerged as a *crowdsourced alternative* to traditional encyclopedias, such as Britannica. As of 2017, Wikipedia is among the top five accessed websites globally, while Britannica has a popularity rank of 2,153¹. Over the years, Wikipedia has also accumulated a rich body of collaboratively written articles on history which are among its top accessed [51] subjects. Just

¹<http://www.alexa.com/siteinfo/britannica.com> and <http://www.alexa.com/siteinfo/wikipedia.org> (accessed 16, October 2017)

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW 2018, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04.

<https://doi.org/10.1145/3178876.3186132>

as awareness about history is crucial for developing a sense of national, cultural, and personal identity, understanding the differences offered by various history-related sources is important. In this paper, we investigate the ways in which Wikipedia articles about national histories differ from their equivalents in Britannica. Thus, we take an important first step towards comparing the views of the past offered by expert- and crowdsourced sources.

Research question: We ask, *How do the descriptions of national histories in English Wikipedia compare to the corresponding articles in Britannica?* Particularly, we examine the temporal and topical aspects of coverage, and linguistic presentation of the material.

Approach: We aim to offer a first large-scale quantitative investigation of how history articles written by Britannica experts compare to those collaboratively produced by Wikipedians. We take a reader perspective and investigate how the national histories of all UN member states are presented in these encyclopedias. Precisely, we quantify the temporal, topical, and linguistic differences across the articles. We concentrate on *year mentions* as accessible representations of temporal coverage. We retrieve from article texts all date mentions (in the form of 4-digit numbers between 1000–1999), and use them as a unit of comparison [45] across the datasets. To assess temporal coverage differences, we apply a randomisation-based filtering method [46] and subsequently, statistical inference. Our empirical results are validated by history experts. To compare linguistic features, we compute text statistics, apply a range of well-established readability tests, and run a Part of Speech analysis.

Findings: We find that Britannica and Wikipedia exhibit different approaches to historiography, where Britannica leans to a more *spatial and territorial concept of the history of states*, and Wikipedia – to presenting their history as *a sequence of political events*. Precisely, Wikipedia puts a disproportional emphasis on periods of conflict and war, with a preference for events well-known to the general public. In comparison, Britannica articles emphasise conflicts with underlying cultural and religious tensions. Semantically, Britannica relies on vocabulary with religious connotations and on geographical terms, while Wikipedia is heavy on political and military words. Finally, both show characteristics of English Academic prose, although Wikipedia's writing is slightly easier to comprehend.

Contributions and implications: Our investigation is extensive, and the first to offer large-scale quantitative insights on how the expert-written historiography of Britannica differs from Wikipedia's popular view of the past. We combine computational and linguistic analyses to arrive at a comprehensive account of structure (coverage, timelines, and their focal points), content (historical reference of these focal points, semantic differences), and presentation (readability) of both encyclopedias. Our motivation is that collaborative sources like Wikipedia challenge the authority

of traditional encyclopedias, both in popularity and presentation of content, and have become a global facilitator of knowledge.

We commence by presenting an overview of related work in Section 2 and outlining the details of data collection and pre-processing in Section 3. Our analysis (Section 4) is split into several subsections examining each research question in detail. We continue by discussing the findings (Section 5) and the limitations of the study (Section 6), and finally, present concluding remarks in Section 7.

2 RELATED WORK

Our work draws on several theoretical domains. It directly relates to research on cultural history, collective/public memories [14], and the analysis of nations as imagined communities [2]. The comparison of crowd-sourced and traditional encyclopedias is related to theoretical studies on how the digital turn and the rise of mass media culture challenge the traditional notion of expertise [9, 23, 40].

Wikipedia vs. Britannica comparisons: Comparisons between Britannica and Wikipedia have attracted substantial academic interest in the recent years. Most research has focused on verifying the quality and accuracy of Wikipedia’s content by comparing it to authoritative sources. The scepticism about Wikipedia’s credibility was mainly due to the new crowdsourced, self-emerging expertise that the encyclopedia draws upon, unlike peer-reviewed, expert-produced content of traditional encyclopedias [23, 40].

Although even earlier studies showed little difference in quality, breadth, and validity of the content between Britannica and Wikipedia [21], the claims of Wikipedia’s credibility were met with criticism [16, 34], and inspired a range of follow-up studies examining a range of topical domains. For example, Wikipedia articles on mental disorders [41], military history [27], and Top Fortune companies [36] have been scrutinised by the field experts, and in every case have been found at least as accurate and broad, or even more up-to-date than Britannica or other authoritative peer-reviewed sources. Other studies, however, suggest that the quality of Wikipedia articles might vary depending on the chosen field [12], and even from article to article within one domain [24]. Most of research on Wikipedia’s reliability is unfortunately based on very small samples (several articles), and can not be scaled up due to reliance on qualitative methods and field experts.

Several studies looked into the differences in content presentation between the encyclopedias. Messner [36] reported that Wikipedia uses a more positive/negative language than Britannica when it comes to articles on large corporations. Greenstein et al. [16] computed political slant and bias in 4K Britannica and Wikipedia articles on the US politics, and found that Wikipedia is more biased towards Democratic views. Their results vary depending on the length of the article and the computation method, though. Finally, the encyclopedias have been compared in terms of content readability, but the results are also controversial [17, 26, 31].

Although the actual differences between Wikipedia and Britannica in terms of content quality and reliability are not great, Wikipedia suffers from perceived credibility and article selection issues [11, 32, 47], especially when contrasted with Britannica [18, 30]. To sum up, most comparative studies focus only on one dimension (usually, content validity), and do not offer a holistic picture of structural differences between the encyclopedias.

Crowd- vs. expert-written history: While Britannica presents a credible, expert-written resource on history, Wikipedia offers an unsupervised, self-emerging, and multifaceted view of the past. In Social Sciences and History literature, Wikipedia is studied in the paradigms of open source history, participatory/amateur history-making [42], collective memories [14, 43], and collaborative re-interpretation of the past [40]. While professional historians do not necessarily share the same understanding of the past as Wikipedians [14], the immense popularity of Wikipedia as a reference source, especially on history [51], makes it an attractive object for studying.

When it comes to the history domain, the possible differences between crowd- and expert-written encyclopedic articles largely remain a terra incognita. To the best of our knowledge, only several studies have juxtaposed the accuracy, breadth, and depth of historical articles in Britannica and Wikipedia. Holman [24], compared the content of nine Wikipedia articles against their equivalents in Britannica, the Dictionary on American History, and American National Biography Online, and found Wikipedia’s accuracy to be less reliable (80% compared to 95% in other sources). Luyt [33] discovered that this weakness is due to many claims in Wikipedia not being verified through citations. A qualitative analysis of the ‘War of 1812’ article in both encyclopedias [27] showed that the Britannica article was briefer and focused more on the causes of the war, while lacking in military and naval aspects. The article also concludes that Wikipedia articles on military history are more detailed and easier to read than their Britannica counterparts.

Apart from qualitative research, several approaches have been used to quantify history on a large scale, including network science [25, 38, 48, 50], mathematical modelling and prediction [28, 52], text mining and topic detection [5, 37], and temporal event extraction [5, 46]. None of them, however, have been applied to compare historical content of online encyclopedias. In this paper, we combine computational methods in order to examine, how collaboratively produced Wikipedia articles on national histories compare to the equivalent Britannica articles, both in terms of temporal and topical coverage of events, as well as the linguistic characteristics.

3 METHODOLOGY

In this section, we describe the process of collecting, pre-processing, and validating the data, as well as outline the methodological details.

3.1 Data collection

We focus on the history of 193 countries² which are the current UN member states³. Although Wikipedia is a multilingual encyclopedia, in this analysis we only focus on its English edition. This is due to the fact that the Encyclopedia Britannica is only available in the English language, and thus, multilingual comparison is not possible.

Wikipedia corpus. For each of the countries we locate an article in the English edition of Wikipedia, titled ‘History of X’, where X is the country name. We retrieve the article’s main text, as well as the text of all Wikipedia articles to which this page outlinks. We focus on the out-links because they provide readers with an

²Throughout the paper we use the terms nation, country, and state as synonyms, being aware of the differences.

³List of the UN member states, <http://www.un.org/en/member-states/index.html> (accessed 16 May 2017)

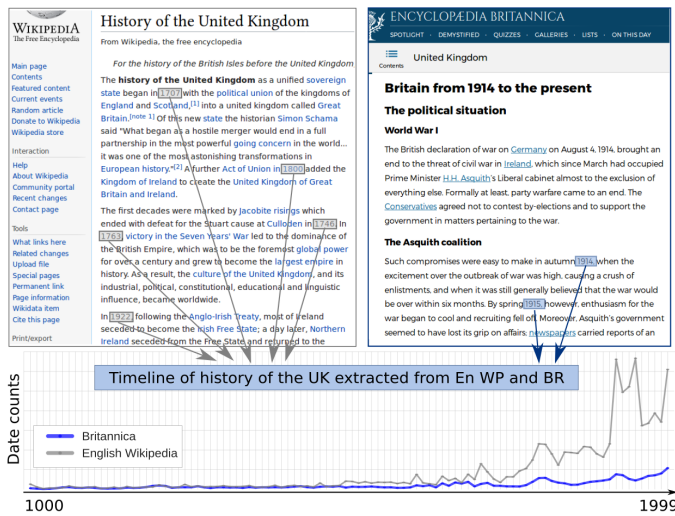


Figure 1: Temporal information extraction. We show parts of the article on the UK history as they appear on Britannica and English Wikipedia websites in 2017. We collect all 4-digit numbers from the main text of each article, as well as from the texts of all outlinked articles, and analyse the resulting distribution (bottom part of the figure). The data provides insights into the temporal focus and attention of encyclopedic articles.

opportunity to follow up and explore topically related material, and thus play a role in shaping user navigation across historical topics.

Britannica corpus. The online Encyclopedia Britannica⁴ has a format similar to Wikipedia: the articles are split into topical sections, some contain infoboxes, and the main text incorporates hyperlinks to other Britannica articles. Unlike in Wikipedia, there are no distinct articles on national histories. Instead, this information is embedded as a separate section in the main article about each nation. Usually, this section has multiple subsections focusing on various important events and periods, including the history of pre-states. For this analysis, we identify Britannica articles on all UN member states titled ‘X’, where X is a country name. For each article, we retrieve the text of the section titled ‘History’, as well as the text of the outlinks⁵. Other sections, such as ‘Economy’, ‘Land’, and ‘Cultural life’ are excluded as irrelevant.

Pre-processing of corpora. For both datasets, we extract data in HTML format, and clean it with the BeautifulSoup parser to exclude text and tags related to, e.g. references, section titles and subtitles, captions, such that both datasets consist only of the main article text. For Wikipedia, we additionally remove (using regular expressions) all instances of citing references (in the format [n], where n is the position of the reference in the article bibliography).

For analysis of language complexity, we prepare several corpora. First, we create a) (main + outlinks) corpus which encompasses all collected text per country, including both seed article and its outlinks. Its reduced version b) (main) consists of the text of the seed articles, excluding the text of the outlinks. In these corpora the

⁴The Encyclopedia Britannica, <https://www.britannica.com/> (accessed 16 May 2017)

⁵One exception is the article on Monaco, which is not split into sections. In this case, we used the entire text of the article and all of its outlinks for the analysis.

length of text about country X in Wikipedia might be of significantly different length compared to Britannica. We create an additional c) (main equalised) corpus based on the text of seed articles, but matched in size between Wikipedia and Britannica articles. To do so, for every country, we compare article length (in words) between the two encyclopedias. We keep the shorter article as it is, and randomly remove sentences from the longer article until the word count is equal or lower than the size of the smaller article. As a result, the word count per country is the same across Wikipedia and Britannica, rounded up to the sentence boundary.

Extracting temporal expressions. In order to assess the coverage of historical periods, we count mentions of year numbers in article texts. Since we are interested in historical events of the last millennium, we retrieve all 4-digit numbers in the range between 1000 and 1999. We use the same procedure (illustrated in Fig.1) for extracting temporal expressions from both datasets. In Wikipedia we encounter examples of paragraphs that consist mostly (more than 50% of words) of hyperlinks. Since there is little narrative in such paragraphs, we record no dates from them.

We ran data collection for both datasets in February 2017, using the access provided by the Wikipedia API⁶, and an HTML scraping script for Britannica. As a result, for Britannica dataset we extracted 326K dates from 27,045 articles including the outlinked articles. In case of Wikipedia, we processed 54,401 pages and retrieved approximately 3M dates. For both datasets, we focus only on the main text of the articles, excluding infoboxes, section titles, and figure captions.

3.2 Validation of extracted time expressions

In order to ensure internal reliability of our extraction method, we check whether the extracted numbers are years rather than numerals indicating, for example, height. For each dataset, we create a random sample of 1,000 extracted 4-digit numbers evenly split across 10 centuries, and ask 3 independent human coders to evaluate each number as a date or a false positive. For each century there are 100 evaluation tasks, which consist of the potential date (4 digits), and the text surrounding it (40 characters before and after the number). If the coder is unsure about a number, we treat it as a false positive. Each case is settled by the majority vote. We compute the expected error rates for centuries as

$$\langle E_{corp} \rangle = \frac{1}{D_{corp}} \sum_c \left(\frac{n_{err,c}}{100} D_{corp,c} \right), \quad (1)$$

where D_{corp} and $D_{corp,c}$ are the total counts of collected (potential) dates per corpus $corp$ and century c , and $n_{err,c}$ is the count of false positives in our random sample for century c .

The inter-rater agreement is substantial (Fleiss’ kappa = .79). Both datasets show very low expected error rates (0.01 per dataset). For Wikipedia, we estimate the highest error rate in the 11th century (.24), since a large number of extracted digits turned out to be numerals relating to heights, population counts, etc. Other false-positives, both for Britannica and Wikipedia, are mostly dates from the Before Christ era. In the more recent centuries our extraction method is very exact (expected error for the 20th century is < .001).

⁶Wikipedia API for Python, <https://pypi.python.org/pypi/wikipedia/> (accessed 16 May 2017)

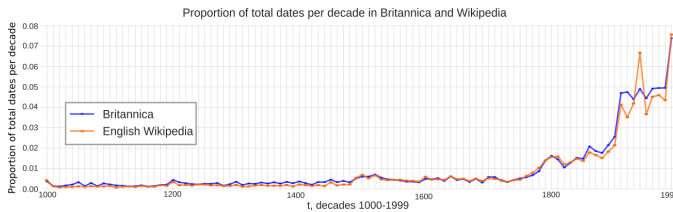


Figure 2: Normalised distribution of collected dates. All collected years are binned into decades, and normalised by the total number of collected dates per dataset. Both Wikipedia and Britannica show a strong bias towards covering the last 100 years. Wikipedia demonstrates a visibly higher peak of coverage in the decade corresponding to the WWII.

4 ANALYSIS AND RESULTS

We present our results in several parts. First, we compare Britannica and Wikipedia in terms of the most covered years and historical periods (Section 4.1). In Sections 4.3 and 4.4 we then narrow the analysis down to selected countries, and calculate the decades that are covered most differently across the datasets, as well as extract and compare temporal focal points of nations. Finally, we report on the linguistic presentation of the articles. In Section 4.5 we compare the most distinctive topics characterising each dataset. We conclude the analysis by presenting an overall comparison of readability and language complexity of the encyclopedias in Section 4.6.

4.1 General patterns of coverage

Before diving into the computational analysis, we compare the datasets in terms of the number of collected dates and their distribution across the national timelines. We observe a startling difference in the number of dates collected from both encyclopedias: while Britannica has a total of 326,021 year numbers between 1000 and 1999, Wikipedia is a tenfold as large with 3,325,946 dates. Some of the most covered countries in both encyclopedias are large European economies (e.g. the UK, Germany, France) accompanied by Australia and the US. The least covered tail of Wikipedia is dominated by the African countries and island states of Oceania. This trend is visible in Britannica too, although it also includes some Asian nations. Overall, there are only 98 countries for which we extract more than 1,000 dates from Britannica articles. In the Wikipedia dataset, even the least covered country has about 1,500 dates.

In order to compare the distribution of dates across the corpora, we bin all collected dates into decades, and normalise them by the total number of dates collected per dataset. Both Wikipedia and Britannica show an uneven distribution of temporal coverage (Fig. 2) with small peaks around 1500 (possibly related to the Age of Discovery) and 1800 (Napoleonic war). A particularly strong peak falls on the 20th century, where the periods of First and Second world wars are most visible. Overall, for both encyclopedias we observe a strong bias towards covering the last century. Additionally, Wikipedia demonstrates a visibly higher peak of coverage in the decade corresponding to the WWII.

4.2 National temporal distributions

We first explore the overall similarity between Wikipedia and Britannica timelines for each country. For that, we present each country as a vector of 100 values (equal to the number of examined decades), each value being the normalised date count. We then compute cosine similarity between the Wikipedia and Britannica country vectors. Overall, the similarity values range between .59 (San Marino) and .98 (Botswana, Rwanda, Australia), with an average of .88. Thus, the timelines are on average very similar.

To continue, we explore how focused the national timelines are on covering particular periods, as opposed to covering every decade to a similar extent. We take an information theory approach: we treat each decade bin of a national timeline as a separate information channel, and compute the entropy across all channels. Thus, the country with an equal number of dates in each decade will have the maximum entropy. Evidently, the minimum entropy corresponds to the case when all country dates are concentrated in just one decade. We compute country entropy as $S_c = -\sum p_i \ln(p_i)$, where p_i is the normalised frequency of dates in decade i .

Fig. 3 demonstrates the distribution of the entropy scores. Based on the location of centroids, we conclude that in both encyclopedias the timelines of European states are presented in the most equalised manner, while for the countries of Africa and Oceania, Wikipedia and Britannica articles are more biased towards covering a limited number of decades. This bias towards covering certain decades is more typical in Britannica (all centroids are above the diagonal). A few exceptions from this rule are large European states (the UK, Germany, Italy, Spain, France), whose historical timelines are presented much more equally on Britannica, compared to Wikipedia. In the next subsection we continue to examine the cases where temporal coverage differs substantially across the encyclopedias.

4.3 Most differently covered periods

As demonstrated in the previous section and in the Fig. 2, the shapes of national timelines in Britannica and Wikipedia are on average very similar. However, discrepancies are also present. In this section we automatically extract and highlight the decades which are covered differently by the encyclopedias. In particular, we explore in which decades the number of dates in one dataset is noticeably higher (or lower) than a fixed expected baseline. It makes intuitive sense to define the baseline R_c as the ratio of Britannica to Wikipedia total dates for a given country c . We assume that this ratio remains constant for each decade in a timeline of a country. Thus, we test the assumption that regardless of decade, Britannica will always have R_c times fewer dates compared to Wikipedia.

Results: We visualise the outcome of this simulation for a selection of countries in Fig. 4. It is visible that the encyclopedias have data sparsity issues, especially pronounced in earlier decades and in non-European countries. This issue affects Britannica to a much greater extent than Wikipedia. Nevertheless, based on the decades where both encyclopedias have enough dates, Britannica pays proportionally more attention to the earlier periods. Precisely, in most of the decades before the 20th century the ratio of Britannica to Wikipedia date counts exceeds the expected R_c ratio for that country. Wikipedia, on the other hand, has a strong bias towards more recent events. We also notice an overproportional emphasis

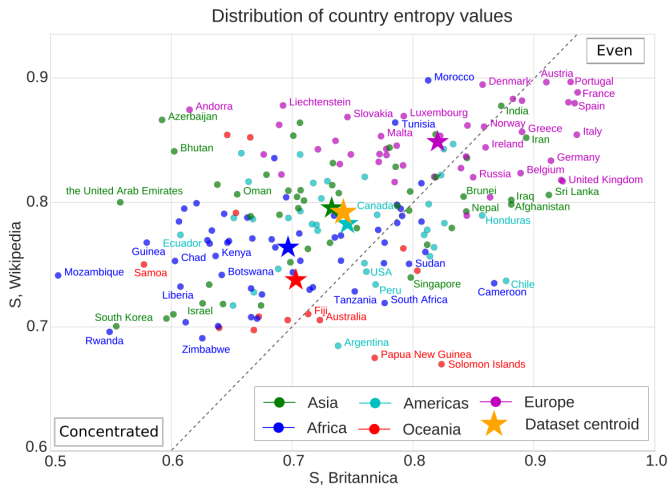


Figure 3: Distribution of country entropy values. The scores are normalised to range between 0 (all country dates in one decade) to 1 (all decades have equal number of dates). Based on the location of centroids (stars), in both encyclopedias the timelines of European states are presented in the most equalised manner, while for the countries of Africa and Oceania, Wikipedia and Britannica articles are more biased towards covering a limited number of decades. On average, this bias towards covering certain decades is more pronounced in Britannica (all centroids are above the diagonal).

of Wikipedia on the times of conflict and war, which is true not only for the 20th century’s First and Second world wars, but presumably also adds up to the red Wikipedia-cells in earlier periods (as shown in Fig.4). Some examples we find include: the Franco-Italian wars (1490s to 1550s), the Franco-Dutch war (1670s), the French War of Devolution (1667-68), the War of the Spanish Succession (1701-14), the history of Canada between its invasion (1775) and the war of 1812, the insurrection of Otto of Greece (1843), and the Crimean war (1853-56). Another focus of Wikipedian writing seems to fall on what might be called popular periods: the times that are well known not only to history experts but also to a wider audience, e.g. the reign of Louis XIV or the French Revolution in France, Reformation or the Age of Enlightenment, and the period of Weimar classicism in Germany. Britannica, in comparison, highlights times of conflict to a much smaller extent: the Wars of Religion (1560s, settled by the Edict of Nantes in 1598) in France, Restoration wars in Portugal (1640-48), or the Greek war of independence (1820s). It also shows a noticeable focus on the periods of African (de-)colonisation.

4.4 Historical focal points

To continue our investigation of temporal coverage patterns in Britannica and Wikipedia, we extract and compare *the focal points of national timelines*, i.e. the decades which are mentioned significantly more (or less) compared to what is expected by a Null Model.

Method: A Null Model of focal points. In order to extract the focal points, we adapt to our dataset the randomisation technique introduced in [46]. We first create a pool M of all collected dates. Then, we randomly draw for each country N_i dates from the pool,

where N_i is the number of dates collected for country i . This builds a randomised national timeline. We repeat the process 1,000 times. For each decade we can then build a distribution of the expected dates within the Null hypothesis of events randomly distributed in time. This allows us to compare the mean of this distribution $E[w_i^d]$ with the empirical date count for the country in the same decade, w_i^d , and convert the difference into a z-score. The z-score of country i in decade d is thus given by:

$$z_i^d = \frac{w_i^d - E[w_i^d]}{\sigma_i^d}, \quad (2)$$

where σ_i^d is the standard deviation of the simulated date counts for decade d . With this procedure, we can identify for each country in which decades the number of observed dates w_i^d differs significantly from the expected number of dates given the Null hypothesis.

Comparison of extracted focal points. As a result of the simulation, we obtain two timelines of focal points (Wikipedia and Britannica versions) for each country. We summarise the differences between them by computing cosine similarity. The values of cosine similarity range between .92 for Argentina (both encyclopedias offer practically identical timelines) and -.55 for Morocco (focal points in one timeline are of low interest in the other), and are centred at .45. Thus, in terms of focal points, the encyclopedias offer rather diverging versions of national histories. Evidently, low average similarity is partially related to the missing data in Britannica. (For example, Morocco timeline has less than 20 decades with at least 30 dates.) However we also find dissimilarities between the decades for which data sparsity is not an issue. To illustrate them, we plot the distribution of focal points obtained from each encyclopedia, one under another for 10 top covered countries (Fig. 5).

Two types of signal are evident. Even though we applied the method independently on each dataset, the *agreement in some focal points* is obvious. For Mexico, both encyclopedias focus on the Mexican War of Independence (1820s). In the US timeline, the focal events are the American Revolution (1760-90s), and the American Civil War (1860s). Articles on Canadian history highlight the decades associated with the struggle between France and Britain for dominance in the North America (Seven Year’s War, 1756-63). The history of South Africa in both encyclopedias mostly highlights the colonisation period (Scramble for Africa in late 19th century). For the Netherlands, the specific period of interest between 1560s and 1670s is likely related to the Eighty Years’ War, or as it is also called, the Dutch War of Independence against the Spanish hegemony. The history of Portugal focuses on the dynastic crises: Portuguese interregnum (1380s) and the succession crisis of 1580s. A similar trend shows up in the articles on the history of China, where both encyclopedias highlight the formation of the Jin (1130s), Yuan (1270s), Ming (1360s), and Qing (1640s) royal dynasties.

Perhaps even more interestingly, another signal that we see in the data is the *disagreements between the encyclopedias*. This is pronounced most strongly in the articles on history of Germany. While Wikipedia narratives strongly focus on the WWII, Britannica is disinterested in the 1930-40s. Similarly surprising, the French Revolution (1780s) is pronounced on the Wikipedia’s timeline, but it does not show up on the Britannica’s timeline of history of France. Instead, Britannica focuses on the French Wars of Religion (Huguenot

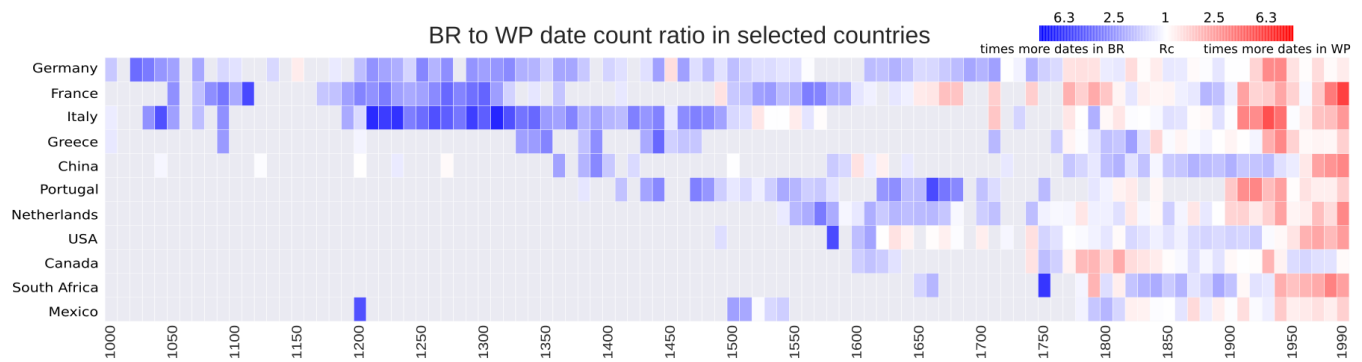


Figure 4: Comparison of Wikipedia and Britannica timelines. For each row we compute R_c , a ratio BR/WP based on the number of collected dates for that country. We then compare this country ratio to each BR to WP datecount ratio in each decade. Cell colour shows by how many times the decade ratio is different from the country ratio. The cells where Britannica has more decade dates than predicted by the country ratio are coloured in blue, otherwise – red. Cells with fewer than 30 dates in either BR or WP are masked out (grey). The cells are white when country and decade ratios are equal. The plot shows that for the decades where there is enough data, Britannica pays proportionally more attention to earlier decades, and Wikipedia focuses on the recent periods related to political instabilities, e.g. WWII.

Wars of 16th century), and the extension of the Crown Lands of France (1180s to early 14th century), which coincided with the crusades by the Catholic Church against the Cathards. Britannica articles on Italian history focus on the Medieval period between 12th and 13th centuries, which is characterised by the rivalry of the Guelphs and Ghibellines, supporting the Pope and the Holy Roman Emperor. Wikipedia, on the other hand, shows no such emphasis.

4.5 Most distinctive topics and vocabulary.

After looking into some of the temporal coverage characteristics of the datasets, we move to a textual analysis of the articles in order to get a first understanding of the themes covered in the articles. We start by extracting the words that are most distinctly used in one dataset, compared to their usage in the other. For that, we extract the union between the top 1000 most frequent words from Wikipedia and Britannica (main corpus) (1219 words), and compare word frequencies using a χ^2 test of independence of variables in a contingency table. We report the results in Table 1. The words are ranked according to the value of χ^2 statistic, which reflects how significantly biased the usage of the word is towards Britannica (left column) or Wikipedia (right column). Among the analysed words, Britannica relies most distinctly on vocabulary with religious or philosophical connotations, such as *Christ, faith, Jesus, God, spirit, divine; idea, doctrine, systems* and geographical terms (*rivers, plain, basin, mountain, rocks*). Wikipedia, on the other hand, relies heavily on political and military vocabulary, such as *war, killed, colony, soldiers, army, empire, ships, armed, captured*.

4.6 Text complexity and readability

Both encyclopedias aim at a wide range of readership, and thus should be written in a way that is accessible to a diverse audience. In this section, we explore this intuitive hypothesis by computing various language complexity measures. Below we report on how two corpora compare in terms of simple text statistics, article readability, and part of speech usage. Depending on the analysis,

we use either the entire Wikipedia and Britannica corpora (main + outlinks), or their reduced versions (main) and (equalised main). We describe how we construct these corpora in Section 3.1.

Text statistics. We report descriptive text statistics for the (main) corpus. These are computed for each country article separately, averages over each dataset are summarised in Table 2. We use Welsch's t -test to compare the means. On average, Wikipedia articles about history use longer sentences (21.6 words vs. 19.9 in Britannica, $p < .001$), and slightly longer words (5.2 characters vs. 5.1, $p = .005$); the differences are statistically significant. To put the numbers in perspective, note the average sentence length in spoken speech (18 words on average) and academic writing (24 words) [10]. Longer unites of text indicate that Wikipedia uses a slightly more formal writing register. Based on the average word length, both encyclopedias score higher than Academic prose (4.8 characters [6]), and thus belong to the most formal text genre.

Finally, we report the average article length, measured in the number of sentences and words per article (see Table 2). The comparison reveals no significant differences. However, we find interesting particularities in the way both datasets reference temporal information. Precisely, Wikipedia texts cite dates (years) significantly more often. The differences are significant both measured as number of dates per 100 words (1.7 dates in Wikipedia vs. 1.3 in Britannica, $p < .001$) and per 100 characters. This might indicate that Wikipedia leans towards factual, rather than descriptive narratives.

Readability. Text readability is usually estimated as the minimal number of education years needed to understand the text at first reading, and is often interpreted using the US grade level system. Readability scores are commonly based on surface characteristics of text, such as the number of its units (syllables, words, and sentences). Some of the tests also include semantic features, such as word difficulty estimated by the word length (in characters [13, 49] or syllables [19, 22, 35]), or by comparison with pre-computed dictionaries of easily understandable words [15]. In order to benefit from various approaches, we compute several established readability scores. We perform the analysis on the (equalised main) corpus

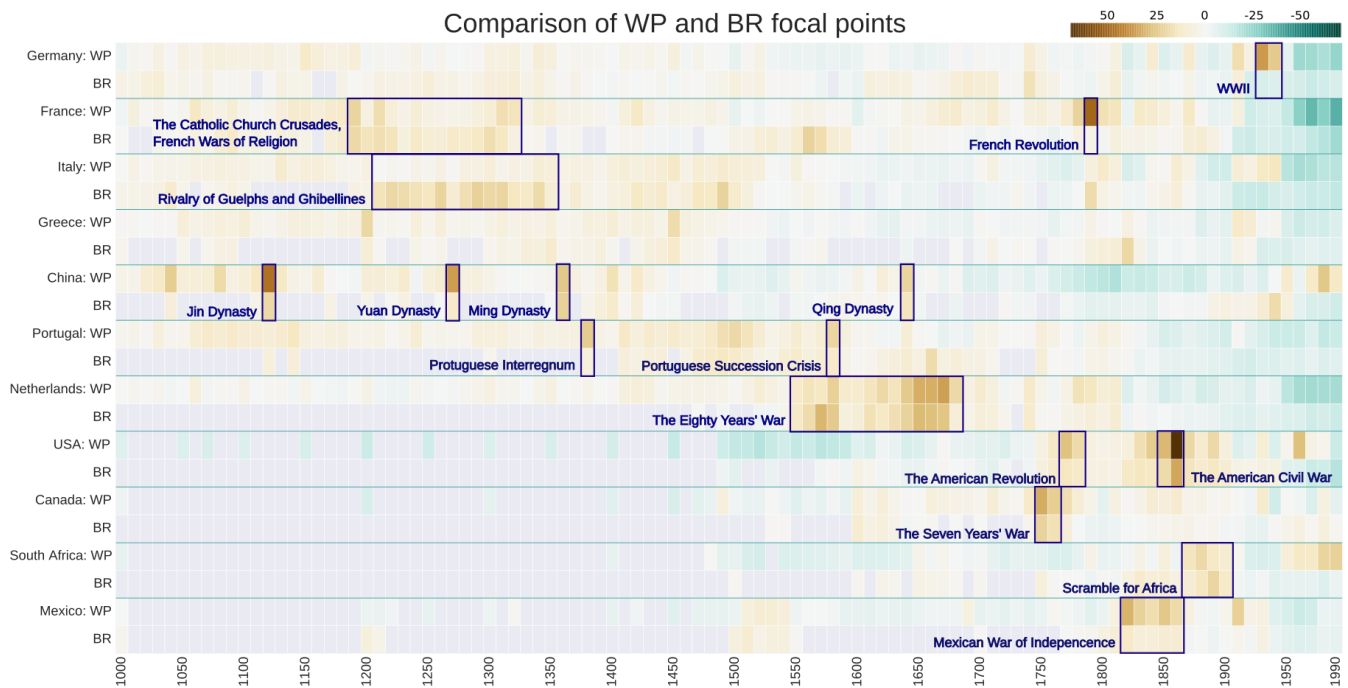


Figure 5: Temporal focal points of selected countries: comparison between Wikipedia and Britannica. z-scores below -4 and above 4 correspond to Bonferroni-corrected p -values < 0.01 , which means the results in all coloured cells are statistically significant. Higher z-scores (orange) correspond to positive differences between the observed and the expected date count per decade, and could be interpreted as focal points of the timelines. Cells with fewer than 30 dates are masked out (grey). z-scores of Britannica vary between $[-50; 50]$, and of Wikipedia – between $[-70; 70]$. The annotations are produced by history experts. While overall the similarities between the distributions of focal points in Britannica in Wikipedia are evident, the differences are indicative of diverging approaches to historiography.

to compensate for the article length differences. The results are summarised in Table 3, all differences are statistically significant (Welsch's t -test, $p < .001$). FRE⁷ ranges between 0 (very hard to understand) and 100 (understandable to a 5th grader). For both Wikipedia and Britannica the score is around 40, or appropriate for an average high school graduate. While the practical difference between the scores is not large, Wikipedia appears slightly easier to comprehend. Other measures concur with this result, always mapping Britannica's readability to a higher required US grade level (and thus, lacking readability). While there is variation across the scores as to which graduate level to map each encyclopedia, between the datasets the signal is clear. Wikipedia consistently shows lower readability scores than Britannica, i.e. its articles are written in a language that is accessible to a wider audience. As a note, these grade scores should not be considered as precise values. Depending on a socio-economic and cultural background of the reader and their motivation to read the text, readability formulae are known both to over- and under-estimate comprehension difficulty [29].

Part of speech analysis. We use the (equalised main) corpus to compare the distributions of parts of speech (POS) frequencies. To tokenise the texts, we applied the Penn Treebank POS tokeniser [1]. It erroneously counts multiword proper nouns as separate entities

⁷The acronyms are abbreviated as follows: FRE - Flesch reading ease; FKG - Flesch-Kincaid grade; CLI - Coleman-Liau index; ARI - Automated readability index; DCRS - Dale-Chall readability score; G-FOG - Gunning FOG index; HS - High school.

(e.g. *New York* results is two single proper noun tokens, rather than one multiword proper noun token). Thus, we added a layer of post-processing, merging into one token all instances of adjacent proper nouns which are not separated by punctuation or other POS. The results of the analysis for the most frequent POS⁸ are summarised in Fig. 6. Both encyclopedias show incredible similarity (cosine similarity = .99) in their patterns of POS usage. The most used POS are nouns and adjectives, which is a general property of written Academic English [7]. Since both corpora describe the past, verbs in past tenses are frequent. We discover interesting statistical differences, for example, in usage of proper nouns and numerals. On average, Wikipedia mentions proper nouns and names (e.g. unique entities, people, well-known events) significantly more often than Britannica. It also uses numerals (including dates) with much higher frequency. This hints that Wikipedia might be more focused on famous events, entities, and biographies. Britannica, on the other hand, shows a notably high frequency of nouns, WH-determiners

⁸POS are defined as follows: NN: noun, common, singular or mass; IN: preposition or conjunction; DT: determiner; NNP: noun, proper, singular; JJ: adjective or numeral, ordinal; NNS: noun, common, plural; VBD: verb, past tense; CC: conjunction, coordinating; VBN: verb, past participle; CD: numeral, cardinal; RB: adverb; TO: to; VB: verb, base form; VBG: verb, present participle or gerund; PRP\$: pronoun, possessive; VBZ: verb, present tense, 3rd person singular; PRP: pronoun, personal; VBP: verb, present tense, not 3rd person singular; WDT: WH-determiner; NNPS: noun, proper, plural; WP: WH-pronoun; JJR: adjective, comparative; JJS: adjective, superlative; WRB: Wh-adverb; MD: modal auxiliary; RP: particle; EX: existential there.

(*that, what, which*), and coordinating conjunctions (*therefore, and, but, so*). Thus, it may exhibit a more didactic and impersonal style, as well as an organised and logical flow of narrative with a focus on explaining structural connections between entities.

5 DISCUSSION

Our results indicate that both encyclopedias are biased towards covering the most recent periods, and not the remote past. This recency bias is more pronounced in Wikipedia, which has a strong emphasis on the First and Second world wars. Previous research has shown that this holds across other Wikipedia language editions [46]. The authors attribute it to the general psychological tendency to perceive recent events as more important [44]. This phenomenon is extensively discussed in the literature on collective/social memory [3, 4, 8, 53]. However, it is mainly associated with public, non-professional narratives. It is new to demonstrate the indication of the same bias in the expert-produced Britannica. We observe a better coverage of large economies, including mostly European

Biased towards Britannica			Biased towards Wikipedia		
Word	BR	WP	Word	BR	WP
feet	18474	22	due	53	107590
miles	16615	53	british	22766	377875
metres	15960	17	war	47661	599373
christ	14565	12	government	44254	561361
faith	13855	57	killed	275	84039
jesus	13330	17	japanese	240	79731
god	35350	36828	colony	288	77958
toward	11615	151	soldiers	178	73302
square	9884	62	started	80	65504
spirit	9757	44	anti	186	66960
divine	9369	16	army	17024	260246
rivers	8706	107	campaign	301	65399
football	8410	7	forces	13409	219442
plain	8440	54	empire	25631	342131
idea	8183	97	ships	93	60593
doctrine	7986	48	president	11871	202456
mountain	7728	70	police	171	57468
systems	7708	69	towards	1	54365
beyond	7367	98	portugal	201	57368
complex	7234	93	armed	296	58780
rocks	7029	8	french	23496	310546
basin	7081	78	captured	222	55474
games	6826	12	arrived	145	54157
extensive	6698	154	around	9159	162148
importance	6590	120	post	155	52273

Table 1: Top word usage in the main articles of Wikipedia and Britannica. On the left, top 25 words that appear most distinctly in Britannica (ranked by χ^2 values), compared to Wikipedia, and on the right – most distinct words in Wikipedia. The values correspond to word frequencies in the (main) corpus. While Britannica is distinct in using religious, philosophical and geographical vocabulary, Wikipedia is heavy on political and military terms.

states (the UK, Germany, France), the US, and Australia. On average, the history of the European region is comparably detailed and equalised across their entire timelines, while for African countries and the small island states of Oceania the timelines cover only a limited number of decades. This Eurocentric bias in professional historiography has been criticised by historians [20], but has not yet been discussed in the context of the Encyclopedia Britannica.

When it comes to language, both encyclopedias exhibit general properties of Academic English prose [7]. In terms of text readability, Wikipedia is accessible to a wider audience. Our findings are similar to the results of the 2009 comparative analysis [17], however both should be interpreted with care [29]. To add, our analysis of POS usage suggests that Britannica might offer an overall more didactic, impersonal style with a more organised and logical writing flow.

Juxtaposing the characteristics of temporal coverage across the encyclopedias, we notice that Britannica and Wikipedia might exhibit different approaches to historiography. Our temporal analysis reveals that Wikipedia overproportionally emphasises the periods of conflict and war, with a specific preference to the events well-known to a general public, such as French Revolution, First- and Second World Wars. Britannica articles do not show focal points

Statistic	Wikipedia	Britannica
Av. word length**	5.2±.1	5.1±.1
Av. sentence length (char.)***	156.9±17.1	140.6±12.7
Av. sentence length (words)***	21.6±2.1	19.9±1.5
Av. lexicon count	6,831±8,860	7,040±5,535
Av. dates per 100 chars.***	.33±.11	0.25±.09
Av. dates per 100 words***	1.68±.57	1.28±.46

Table 2: Descriptive text statistics compared for Wikipedia and Britannica datasets. Texts of outlinked articles are excluded. Results are computed per article, averages are reported with standard deviations. Rows with statistically significant differences are starred: *** corresponds to $p < .001$, and ** corresponds to $p = .005$. Comparison suggests that Wikipedia uses a slightly more formal writing style, and on average cites dates more often than Britannica.

Av. read.	Wikipedia	Britannica
FRE	46.67 ± 6.3 [HS]	42.9 ± 5.9 [HS]
FKG	11.92 ± 1.4 [12 th gr.]	12.7 ± 1.4 [13 th gr.]
CLI	13.8 ± 1.1 [14 th gr.]	14.5 ± 1.2 [15 th gr.]
ARI	14.5 ± 1.5 [15 th gr.]	15.6 ± 1.7 [16 th gr.]
DCRS	8.8 ± .8 [12 th gr.]	9.1 ± .8 [13 th gr.]
G-FOG	10.4 ± 1 [10 th gr.]	10.9 ± 1.2 [11 th gr.]
SMOG	8.8 ± 1.6 [9 th gr.]	9.5 ± 1.3 [10 th gr.]

Table 3: Readability scores⁷. Averages are computed on (equalised main) corpus, estimated grade levels are reported in brackets. All differences are statistically significant at $p < .001$. Across several readability scores, the educational requirements for reading articles about national histories on Wikipedia are lower than the corresponding articles on Britannica.

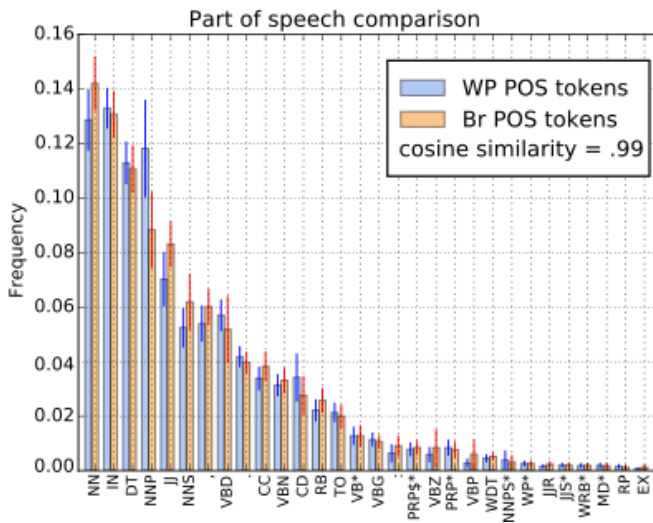


Figure 6: Part of speech⁸ analysis of Britannica and Wikipedia. Mean frequencies and deviations are computed on the (main equalised) corpus. All differences are significant ($p < .001$, Welsh's t -test) except for the starred labels. The encyclopedias demonstrate nearly perfect overall similarity, but Wikipedia tends to mention proper nouns, and numerals (e.g. dates) significantly more often than Britannica. The notably high frequency of WH-determiners, and coordinating conjunctions in Britannica might indicate that its articles have a more structured and logical flow.

associated with these events, but instead emphasise the conflicts with underlying religious tensions, for example, the French Wars of Religion, and the Crusades of the Catholic Church. Comparing the most distinctly used words, we find that Britannica relies on vocabulary with religious connotations and on geographical terms, while Wikipedia is heavy on political and military vocabulary. Moreover, Wikipedia's history articles cite numerals (including dates) significantly more frequently, and have an order of magnitude more dates compared to Britannica. This might indicate that the historiography on Wikipedia is oriented towards outlining facts rather than descriptive narratives. Finally, higher frequency of proper noun usage in Wikipedia supports the earlier observation that Wikipedia is more biased towards covering famous named entities, such as, e.g. well-known events and biographies. Overall, the data seem to suggest that Britannica leans to a more *sociocultural, spatial and territorial concept of history*, whereas Wikipedia – to presenting a *sequence of political events*. Our computational results concur with some of the earlier qualitative observations. For example, a case study of the coverage of the Canadian War of 1812 pointed out Wikipedia's detailed focus on battles, military, and naval affairs, and sparsity regarding social and cultural historical aspects [27]. Britannica, on the other hand, was characterised as focusing on the national border line, and limited in the war thematic.

As a direction for future work, it would be interesting to analyse the accentuation of conflict in the encyclopedias. For instance, whether Wikipedia has a stronger interest in inter-nation conflicts, and Britannica – in sociocultural and intra-nation ones.

6 LIMITATIONS

We would like to note that the results of our study should be interpreted with care due to the limitations listed below.

History of pre-states: Our data might be lacking historical narratives about the history of territories before they reached the current shape. We focus on the history of the current UN member states, however the political map of the world has changed many times throughout history (e.g., post-Soviet bloc). Most Wikipedia and Britannica articles have sections on the history of pre-states in the text of the main article, or outlink to relevant articles. We include the text of outlinked articles to partially solve the issue. Still, some information on pre-history might be lost due to missing links. Also, inclusion of outlinks potentially makes the datasets noisier.

Data validity: Data validation has shown high accuracy of our date extraction method. This is possible because we limited our analysis to the articles evidently related to history. The precision of the method might suffer when analysing texts of broader scope or focusing on the dates from Before Christ era. Already in our sample, we find small numbers of false-positives, e.g. 4-digit numerals expressing heights, lengths, or population counts. Although suitable for the current setup, our dates extraction method might need improvement if applied to a different dataset.

Generalisability: Our findings are valid for the chosen knowledge domain and the English language. It is problematic to generalise how Wikipedia articles on history in other language editions compare to Britannica. Additional research is needed to evaluate if our findings hold for articles with other themes than History.

Temporal coverage: We reduce the complexity of historical writing to a quantifiable unit (date mention). Despite its objectivity, this approach is reductive and might be less precise for earlier periods and countries with few events known/documented by year. The obvious advantage is being able to compare across large datasets.

Focal points: We adapt an established formulation of significant temporal focal points of national timelines, first introduced in [46]. Other formulations of random expectation are possible, which might potentially lead to non-identical outcomes. Our interpretations of the historical events potentially associated with the discovered focal points are a subject to opinion of the experts.

Text analysis: The outcomes of text and readability analyses are sensitive to the tokenisers and text pre-processing [39]. Slightly different results might be expected if applying other methods.

7 CONCLUSION

Knowing history is very important to understand how societies came to be, how whole countries formed, evolved and retained cohesion, and what forces shape their present and future. In this paper, we have analysed the differences between the most popular online reference source, Wikipedia, comparing it to Britannica written by experts. In revealing blank spaces or biases we might contribute to fostering richer and more balanced accounts of history. The undisputed popularity and outreach of Wikipedia make it a worthwhile object of study, since its images of history may distort our view back by focusing on already popular and well-known periods, as well as on violent conflicts and political events.

REFERENCES

- [1] 2017. The University of Pennsylvania (Penn) Treebank Tag-set. (2017). <http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html>. Accessed 16 May 2017.
- [2] Benedict Anderson. 2016. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso London.
- [3] Jan Assmann. 2011. Communicative and cultural memory. In *Cultural Memories*. Springer, 15–27.
- [4] Jan Assmann and John Czaplicka. 1995. Collective memory and cultural identity. *New German Critique* 65 (1995), 125–133.
- [5] Ching-man Au Yeung and Adam Jatowt. 2011. Studying how the past is remembered: towards computational history through large scale text mining. In *CIKM'11*. ACM, 1231–1240.
- [6] Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- [7] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of spoken and written English*. (1999).
- [8] Joël Candau. 2005. *Anthropologie de la mémoire*. Armand Colin.
- [9] Manuel Castells. 2011. *The rise of the network society: The information age: Economy, society, and culture*. Vol. 1. John Wiley & Sons.
- [10] Wallace Chafe and Jane Danilewicz. 1987. *Properties of spoken and written language*. Academic Press.
- [11] Thomas Chesney. 2006. An empirical examination of Wikipedia's credibility. *First Monday* 11, 11 (2006).
- [12] Kevin A. Clauson, Hyla H. Polen, Maged N. Kamel Boulos, and Joan H. Dzenowagis. 2008. Scope, completeness, and accuracy of drug information in Wikipedia. *Annals of Pharmacotherapy* 42, 12 (2008), 1814–1821.
- [13] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2 (1975), 283.
- [14] Margaret Conrad. 2007. 2007 Presidential Address of the CHA: Public history and its discontents or history in the age of Wikipedia. *Journal of the Canadian Historical Association/Revue de la Société historique du Canada* 18, 1 (2007), 1–26.
- [15] Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin* (1948), 37–54.
- [16] Editorial. 2006. Britannica attacks... and we respond. *Nature* 440, 582 (2006).
- [17] Antonella Elia. 2009. Quantitative data and graphics on lexical specificity and index readability: the case of Wikipedia. *RAEL: revista electrónica de lingüística aplicada* 8 (2009), 248–271.
- [18] Andrew J. Flanagan and Miriam J. Metzger. 2011. From Encyclopaedia Britannica to Wikipedia: Generational differences in the perceived credibility of online encyclopedia information. *Information, Communication & Society* 14, 3 (2011), 355–374.
- [19] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221.
- [20] Michael Geyer and Charles Bright. 1995. World history in a global age. *The American Historical Review* 100, 4 (1995), 1034–1060.
- [21] Jim Giles. 2005. Internet encyclopaedias go head to head. (2005).
- [22] Robert Gunning. 1952. The technique of clear writing. (1952).
- [23] Elin Johanna Hartelius. 2008. *The rhetoric of expertise*. The University of Texas at Austin.
- [24] Lucy Holman Rector. 2008. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review* 36, 1 (2008), 7–22.
- [25] Cornell Jackson. 2016. Using social network analysis to reveal unseen relationships in Medieval Scotland. *Digital Scholarship in the Humanities* (2016), fqv070.
- [26] Adam Jatowt and Katsumi Tanaka. 2012. Is Wikipedia too difficult?: Comparative analysis of readability of Wikipedia, Simple Wikipedia and Britannica. In *CIKM'12*. ACM, 2607–2610.
- [27] Richard Jensen. 2012. Military history on the electronic frontier: Wikipedia fights the War of 1812. *The Journal of Military History* 76, 4 (2012), 523–556.
- [28] Edgar Kiser and Michael Hechter. 1991. The role of general theory in comparative-historical sociology. *Amer. J. Sociology* (1991), 1–30.
- [29] George R. Klare. 1976. A second look at the validity of readability formulas. *Journal of reading behavior* 8, 2 (1976), 129–152.
- [30] Ida Kubiszewski, Thomas Noordewier, and Robert Costanza. 2011. Perceived credibility of Internet encyclopedias. *Computers & Education* 56, 3 (2011), 659–667.
- [31] Teun Lucassen, Roald Dijkstra, and Jan Maarten Schraagen. 2012. Readability of Wikipedia. *First Monday* (2012).
- [32] Teun Lucassen and Jan Maarten Schraagen. 2010. Trust in Wikipedia: How users trust information from an unknown source. In *Proceedings of the 4th workshop on Information credibility*. ACM, 19–26.
- [33] Brendan Luyt and Daniel Tan. 2010. Improving Wikipedia's credibility: References and citations in a sample of history articles. *J. of the American Society for Information Science and Technology* 61, 4 (2010), 715–722.
- [34] P.D. Magnus. 2006. Epistemology and the Wikipedia. (2006).
- [35] G. Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading* 12, 8 (1969), 639–646.
- [36] Marcus Messner and Marcia W. DiStaso. 2013. Wikipedia versus Encyclopedia Britannica: A longitudinal analysis to identify the impact of social media on the standards of knowledge. *Mass Communication and Society* 16, 4 (2013), 465–486.
- [37] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 6014 (2011), 176–182.
- [38] John F. Padgett and Christopher K. Ansell. 1993. Robust action and the rise of the Medici, 1400-1434. *Amer. J. Sociology* (1993), 1259–1319.
- [39] João Rafael de Moura Palotti, Guido Zuccon, and Allan Hanbury. 2015. The influence of pre-processing on the estimation of readability of Web documents. In *CIKM'15*. ACM, 1763–1766.
- [40] Damien Smith Pfister. 2011. Networked expertise in the era of many-to-many communication: On Wikipedia and invention. *Social Epistemology* 25, 3 (2011), 217–231.
- [41] Nicola J. Reavley, Andrew J. Mackinnon, Amy J. Morgan, Mario Alvarez-Jimenez, Sarah E. Hetrick, E. Killackey, Barnaby Nelson, Rosemary Purcell, Marie B.H. Yap, and Anthony F. Jorm. 2012. Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological medicine* 42, 08 (2012), 1753–1762.
- [42] Roy Rosenzweig. 2006. Can history be open source? Wikipedia and the future of the past. *The Journal of American History* 93, 1 (2006), 117–146.
- [43] Roy Rosenzweig and David Paul Thelen. 1998. *The presence of the past: Popular uses of history in American life*. Vol. 2. Columbia University Press.
- [44] Darío Páez Rovira, Jean-Claude Deschamps, and James W. Pennebaker. 2006. The social psychology of history: Defining the most important events of the last 10, 100, and 1000 years. *Psicología Política* 32 (2006), 15–32.
- [45] Jörn Rüsen. 1996. Some theoretical approaches to intercultural comparative historiography. *History and Theory* (1996), 5–22.
- [46] Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. 2017. Analysing timelines of national histories across Wikipedia editions: A comparative computational approach. In *ICWSM'17*. 210–219.
- [47] Anna Samoilenko and Taha Yasseri. 2014. The distorted mirror of Wikipedia: A quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science* 3, 1 (2014), 1–11.
- [48] Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. 2014. A network framework of cultural history. *Science* 345, 6196 (2014), 558–562.
- [49] R.J. Senter and Edgar A. Smith. 1967. *Automated readability index*. Technical Report. DTIC Document.
- [50] Søren Michael Sindbæk. 2007. The small world of the Vikings: networks in early medieval communication and exchange. *Norwegian Archaeological Review* 40, 1 (2007), 59–74.
- [51] Anselm Spoerri. 2007. What is popular on Wikipedia and why? *First Monday* 12, 4 (2007).
- [52] Peter Turchin. 2011. Toward Cliodynamics – an analytical, predictive science of History. *Cliodynamics* 2, 1 (2011).
- [53] James V. Wertsch. 2002. *Voices of collective remembering: Test*. Cambridge University Press.