Why the World Reads Wikipedia: Beyond English Speakers

Florian Lemmerich
RWTH Aachen University
florian.lemmerich@cssh.rwth-aachen.de

Diego Sáez-Trumper
Wikimedia Foundation
diego@wikimedia.org

Robert West*
EPFL
robert.west@epfl.ch

Leila Zia
Wikimedia Foundation
leila@wikimedia.org

ABSTRACT

As one of the Web's primary multilingual knowledge sources, Wikipedia is read by millions of people across the globe every day. Despite this global readership, little is known about why users read Wikipedia's various language editions. To bridge this gap, we conduct a comparative study by combining a large-scale survey of Wikipedia readers across 14 language editions with a log-based analysis of user activity. We proceed in three steps. First, we analyze the survey results to compare the prevalence of Wikipedia use cases across languages, discovering commonalities, but also substantial differences, among Wikipedia languages with respect to their usage. Second, we match survey responses to the respondents' traces in Wikipedia's server logs to characterize behavioral patterns associated with specific use cases, finding that distinctive patterns consistently mark certain use cases across language editions. Third, we show that certain Wikipedia use cases are more common in countries with certain socio-economic characteristics; *e.g.*, in-depth reading of Wikipedia articles is substantially more common in countries with a low Human Development Index. These findings advance our understanding of reader motivations and behaviors across Wikipedia languages and have implications for Wikipedia editors and developers of Wikipedia and other Web technologies.

ACM Reference Format:
Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291021

1 INTRODUCTION

Wikipedia is the world's largest encyclopedia and one of the primary knowledge sources on the Web, providing content every day to millions of readers from across the globe in more than 160 actively edited languages. Despite its global reach, very little is known about Wikipedia readers' motivations and information needs across languages. For years, English Wikipedia has been the primary focus of Wikipedia studies, and this has had implications on the way Wikipedia has been developed and supported over the years. In this study, we challenge the focus on English Wikipedia by expanding an earlier study [33] in order to better understand the readers behind different Wikipedia languages. Without understanding similarities and differences between readers across the globe, improving user experience through new content, products, and services will continue to be challenging [3, 8].

Background and objectives. Most research on user motivation and needs has been dedicated to understanding the content producer perspective [1, 24]. Only recently, a study conducted on the English Wikipedia investigated why users read Wikipedia, via a large-scale user survey [33]. However, the focus on the English Wikipedia in that study neglects that, even under similar technical preconditions, the usage of Web contents can significantly differ depending on the cultural background of users [6, 26]. In contrast, the present work aims to understand *why the world reads Wikipedia.*

Materials and methods. We base our analysis on a large-scale multiple-choice survey with questions identical to previous research [33], but with a massively extended scope—engaging readers of 14 Wikipedia languages and receiving more than 210,000 responses. Linking the survey participants to their traces in Wikipedia's server logs and comparing the data with the traces of random samples of readers allows for correcting misrepresentation of user groups and enables us to identify associations between usage patterns in the log data and specific use cases of Wikipedia that hold consistently across languages. Furthermore, we employ country-level datasets to correlate Wikipedia's use cases with socio-economic and cultural indicators.

Contributions and findings. The following are our main contributions: (i) We quantify and compare the prevalence of Wikipedia use cases with respect to motivations, information needs, and prior familiarity across 14 Wikipedia languages via a large-scale survey (Sec. 4.1). (ii) We match survey responses to the respondents' traces in Wikipedia's server logs to characterize usage patterns associated with specific use cases (Sec. 4.2). (iii) We match the survey data with country-level socio-economic and cultural data to allow for a deeper exploration of survey responses (Sec. 4.3).

Based on our analysis, we conclude that Wikipedia is read for a wide variety of use cases in any given language, and the distribution of use cases differs substantially between the languages. English Wikipedia is not fully representative of other Wikipedia languages. Additionally, we conclude that several (but not all) Wikipedia use cases can be associated with similar usage patterns across Wikipedia languages. Finally, we observe that socio-economic characteristics of a reader's country show remarkable correlations with the prevalence of Wikipedia use cases. For example, readers from less developed countries are more likely to be motivated by intrinsic learning and to read articles in depth.

[*] Robert West is a Wikimedia Foundation Research Fellow.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5940-5/19/02.
https://doi.org/10.1145/3289600.3291021

The outcomes of this research can help Wikipedia editors across languages, Wikipedia developers, and the Wikimedia Foundation to create content and build tools and services with a deeper understanding of the needs of Wikipedia readers across the globe.

2 RELATED WORK

To understand Wikipedia readers across languages, our study draws on three different lines of research, described next.

Cultural differences in social media. Notable differences in the usage of social media platforms across countries and cultures have been found in a wide variety of platforms, such as Foursquare [31], Yahoo! Answers [14], Twitter [9, 25], and Google+ [20]. Moreover, previous studies show that, although Chinese sites like Renren and Weibo are technically very similar to Facebook and Twitter, their culture is perceived as more collectivist, suggesting that cultural background could be more important than the technology used in describing the observed usage differences [6, 26]. A comprehensive survey on HCI and cultural differences [16] emphasizes that understanding cultural values is essential for the design of successful user interfaces. Certain aspects of social media usage, *e.g.*, topics discussed, could also be linked to socio-economic factors in Foursquare [28, 36] and Twitter [27].

Wikipedia across countries and languages. Several independent studies have covered specific aspects of cultural differences on Wikipedia. It was found that Wikipedia language editions have a high degree of self-focus, *i.e.*, bias towards the knowledge of the editor community [10, 21] and set different priorities on the information included [5, 13, 17, 30]. Those studies all focus on the editor or content perspective of Wikipedia, while in this paper we investigate the motivations and behaviors of readers.

Wikipedia users' behavior and motivations. The behavior of Wikipedia readers has also been a main topic of interest, but primarily focused on content popularity [18, 29, 34] and navigation patterns [32, 37]. Studies of the motivations of Wikipedia users focused mainly on contributors [1, 24]. By contrast, the motivation of readers has only been picked up recently in the predecessor study of this work, which studied reader motivation in the English Wikipedia only [33]. All these studies neglect the multilingual and cross-cultural perspective that is the focus of this paper.

3 DATASETS AND METHODOLOGY

First, we describe the datasets used in this work in detail.

3.1 Survey Data

We selected 14 Wikipedia languages (*cf.* Table 1) with the following considerations: the language family, language-specific Wikipedia article and pageviews counts, and the number and distribution of speakers worldwide. We also took into account the requests by Wikipedia volunteers to include their languages as part of the study.

Table 1: The surveyed Wikipedia languages with the number of articles, the number of pageviews in the survey period, the sampling rate used for selecting survey participants, and the number of responses.

language	lang	# articles	# pageviews	rate	# resp.
Arabic	ar	523,917	38,102,782	1:10	2,158
Bengali	bn	51,015	1,865,887	1:1	1,198
German	de	2,079,460	227,823,185	1:5	28,000
English	en	5,414,505	1,945,323,873	1:40	24,140
Spanish	es	1,292,245	264,464,604	1:5	39,021
Hebrew	he	208,859	14,088,014	1:3	8,848
Hindi	hi	121,867	9,041,447	1:2	3,064
Hungarian	hu	412,483	11,436,690	1:2.5	2,455
Japanese	ja	1,065,498	307,436,312	1:5	19,996
Dutch	nl	1,904,240	43,017,893	1:8	3,277
Romanian	ro	377,090	8,302,363	1:2	3,829
Russian	ru	1,402,293	224,732,227	1:5	67,621
Ukrainian	uk	703,665	12,446,880	1:2.5	8,041
Chinese	zh	946,356	116,703,091	1:20	5,957

The survey was run from June 22 to June 29, 2017. The sampling rates were chosen with the intention to obtain roughly 30,000 responses from high-pageview languages *vs.* 3,000 from the lower-pageview languages, resulting in sampling rates ranging from 1:40 for English Wikipedia to 1:1 for Bengali Wikipedia (Table 1). We sampled from all users with requests to the specific Wikipedia languages' mobile and desktop sites, excluding requests to non-article pages (discussion pages, search pages, *etc.*), those to the main page of Wikipedia and from browsers with *Do Not Track* enabled. Potential survey participants were marked by assigning a token to their browsers. They were then shown a survey widget inviting them to participate in a three-question survey to improve Wikipedia. The reader had the choice to ignore the message, dismiss it, or opt in to participate. This would take the reader to an external site (Google Forms) with a questionnaire titled *"Why are you reading this article today?"* that contained, in random order, the following questions on their *motivation*, *information need*, and *prior knowledge*, respectively:

- *I am reading this article because...*: I have a work- or school-related assignment; I need to make a personal decision based on this topic (*e.g.*, buy a book, choose a travel destination); I want to know more about a current event (*e.g.*, a soccer game, a recent earthquake, somebody's death); the topic was referenced in a piece of media (*e.g.*, TV, radio, article, film, book); the topic came up in a conversation; I am bored or randomly exploring Wikipedia for fun; this topic is important to me, and I want to learn more about it (*e.g.*, to learn about a culture); other. Users could select multiple answers for this question.
- *I am reading this article to...*: look up a specific fact or get a quick answer; get an overview of the topic; get an in-depth understanding of the topic.
- *Prior to visiting this article...*: I was already familiar with the topic; I was not familiar with the topic, and I am learning about it for the first time.

Prior to submitting their survey answers, readers were informed through a privacy statement[1] about the collection, sharing, and usage of the survey data. Translations of the questions, answers, and privacy statement were provided by known Wikipedia editors

[1] https://wikimediafoundation.org/wiki/Survey_Privacy_Statement_for_Schema_Revision_15266417

and in close collaboration with one of the study authors, to preserve the specifics in the translated texts (*cf.* also Sec. 5.2).

We obtained more than 210,000 survey responses after removing empty or incomplete responses as well as responses that could not be mapped to a user trace (*cf.* Sec. 3.2). Table 1 displays the breakdown of responses by language. The survey responses along with the associated article information are made publicly available along with extended results from this paper.[2]

3.2 Auxiliary Data

We are interested in understanding how users' motivation, desired depth of knowledge, and prior knowledge (*i.e.*, their answers to our survey) are reflected in reading behavior across languages and whether they can be explained through the socio-economic and the cultural context the users operate in. For this purpose, we link survey responses to the auxiliary data sources described here.

Webrequest logs and article data. To analyze respondents' reading behavior in context and to apply bias correction to the survey data collected, we connect survey responses to Wikipedia's webrequest logs. For every request, the corresponding webrequest log contains, among others, the referrer URL, timestamp, client IP address, browser version, and rough geo-location derived from the client IP. Due to the absence of unique user IDs in the webrequest logs, we rely on the concatenation of client IP and user agent as a pseudo ID. To obtain additional information on the requested articles, we extract the text of all articles and the Wikipedia link network for the 14 languages from the Wikipedia dumps[3] of July 2017 (the dump following the survey period). We then follow the methodology of Singer *et al.* [33] to construct sessions for each user ID and extract a variety of features for each webrequest log entry. These features include

- *request features* such as the country or continent of the user, local time, requested Wikipedia host (mobile or desktop), and referrer type (internal navigation, external search engine, or other);
- *article features* such as the degree in the link network for this Wikipedia language, PageRank, text length, and topic (derived via Latent Dirichlet Allocation [4] with $n = 20$ topics separately for each language);
- *activity features* such as the number of articles requested, duration of the session in minutes, time between two requests, and number of sessions during the survey period.

In addition to the survey participants' webrequest logs, we also select a fully random sample of 200,000 Wikipedia readers per language and compute the same set of features for them to enable bias correction (Sec. 3.3).

Country-level data. For a more detailed analysis of the survey responses in the context that survey respondents are in, we connect the survey data and webrequest logs with two external datasets: first, the Quality of Government dataset [35], which provides rich information on a large range of socio-economic statistics at the country level; and second, the well-known Hofstede dimensions [11, 12], which describe the culture and values within a society.[4]

3.3 Correcting Survey Bias

Inferring properties of a general population from a research survey is subject to different kinds of biases, including *coverage bias*, *sampling bias*, and *non-response bias*. To correct for non-response bias, we apply a weighting scheme that gives higher weights to survey participants with user features that are underrepresented in the set of survey participants compared to the representative random sample. For this purpose, we use inverse propensity score weighting [2, 19] based on a gradient boosting classifier, as described in detail by Singer *et al.* [33]. We calculate response weights for each language version independently.

4 RESULTS

This section describes results on why users across the world read Wikipedia articles.

4.1 Survey Results

We start by presenting the distribution of responses to the survey questions across the 14 Wikipedia languages. We compute the weighted percentages of survey respondents with specific motivations, information needs, and prior knowledge, where the weights were computed as described in Sec. 3.3. We visualize the results in Fig. 1.

Motivation. With respect to motivation (Fig. 1a), we observe that Wikipedia is read with a wide range of motivations. In none of the languages does a single motivation clearly dominate. *Intrinsic learning* is the most commonly selected motivation (mean: 37%[5]) across all but three languages: English, Dutch, Japanese. For these three languages, *media* is the top reported motivation instead (mean: 25%), which is also one of the top motivations for all other languages with the exception of Bengali and Hindi. We also observe that considering *intrinsic learning*, there are major differences between language editions: the response shares are generally lower for Western European languages (Dutch: 21%, English: 27%), and substantially higher for Eastern European languages (Romanian: 42%, Russian: 41%, Ukrainian: 41%) as well as Arabic (40%) and Indian languages (Bengali: 55%, Hindi: 48%). Other common motivators are *conversations* (mean: 24%), *work- or school*-related tasks (mean: 18%), *current events* (mean: 17%), and the need for making *personal decisions* (mean: 13%). Finally, we observe that the percentage of respondents being motivated by *work- or school*-related tasks or *being bored* differs significantly across languages. While work- or school-related motivations account for 10% in English Wikipedia, they account for over three times as much (31%) in Spanish Wikipedia. Also, people report being bored as a motivation for visiting Wikipedia in only 10% of responses in Hindi, Romanian, and Ukrainian Wikipedia, and in more than 20% of responses in English, Japanese, Chinese, and Arabic Wikipedia. We further note that the answer *"other"* was selected only rarely (at most 10%), indicating the robustness of the taxonomy defined in earlier research [33].

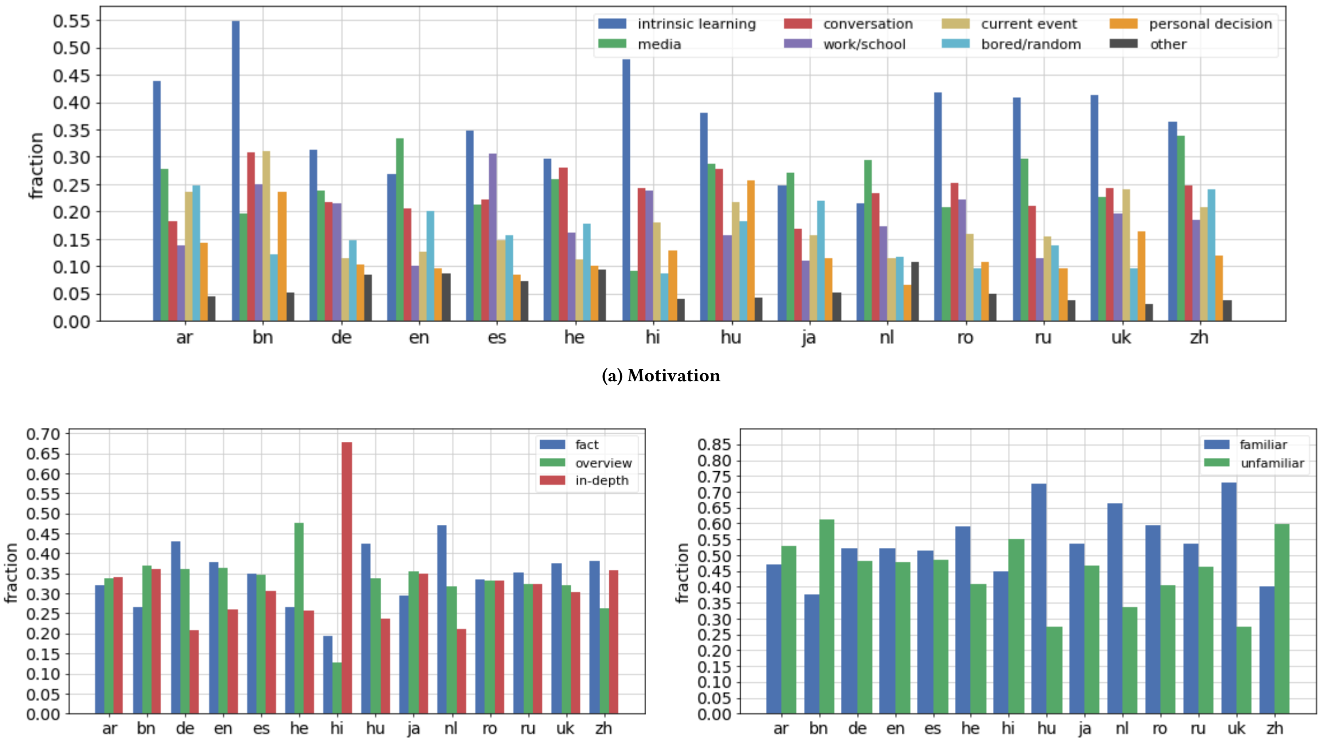
Information need. Considering the information need of readers, we observe that, considering all languages, Wikipedia is visited roughly equally by readers for in-depth understanding (mean: 32%),

[2] https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Reader_Behaviour/Data

[3] https://dumps.wikimedia.org/

[4] https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

[5] Note that this is the unweighted mean of the outcomes for the surveyed language editions. Weighting by language edition size would neglect small editions.



(a) Motivation

(b) Information need

(c) Prior knowledge

Figure 1: Each group of bars reflects the responses in one language edition, each bar represents the share of one response option. For motivation, multiple answer options were allowed for each user.

fact checking (mean: 35%), and obtaining an overview (mean: 33%). We find, however, much diversity between languages. In-depth reading is reported substantially less often for the Western and Central European languages such as English (26%), German (21%), Hungarian (24%), or Dutch (21%). Instead, Wikipedia is more often used for fact checking in these language versions (38%, 43%, 43%, and 47%, respectively). An outlier is the Hindi language, where users report in-depth reading of articles 68% of the time. In Sec. 4.3, we explore this further in the light of socio-economic factors.

Prior knowledge. There are nearly the same numbers of people reporting to be familiar *vs.* unfamiliar with the topic they read on Wikipedia across languages (55% *vs.* 45%). This being said, there are substantial differences between the languages: Eastern European languages report familiarity with the content at much higher rates (Ukrainian: 73%, Hungarian: 73%), while Asian languages with the exception of Japanese report to be unfamiliar more often (Bengali: 61%, Chinese: 60%, Hindi: 55%). These differences could potentially be explained by a tradition and social desirability of humility in these Asian societies [22] (*cf.* Sec. 5).

Robustness over time. We examine the reproducibility and stability of the survey results over time by comparing the prevalence of English Wikipedia use cases with the results of the earlier study conducted on English Wikipedia in 2016 [33]. Fig. 2 shows that the survey results are very similar, suggesting that the observed effects are robust. The only noticeable difference between the results is a decrease in work- or school-related motivation (16% in March 2016 *vs.* 10% in June 2017), which may be due to seasonal effects.

4.2 Survey Results and Webrequest Logs

In the previous study on use cases in the English Wikipedia, certain use cases (motivation, information need, and prior knowledge) could be linked to specific usage patterns extracted from Wikipedia's server logs (webrequest logs) [33]. In this section, we investigate whether such patterns are common across languages or whether they are only present in a subset of the languages.

We start by manually extracting binary usage patterns (such as *session_length* ≥ 3) from the log request features based on the previous study results [33], which have been detected using pattern mining techniques [15]. The binarization allows for applying a single framework for categorical and continuous features with vastly different distributions. For each language and each usage pattern, we can then compute the *share S* of users for which the pattern applies (*e.g.*, a share of 20% of the users have a *session_length* ≥ 3). Additionally, we can compute for each language the *effect E* of any survey question answer on any pattern, *i.e.*, the difference between the percentage of users for which the pattern applies among users that gave a specific survey response and the percentage of users for which the pattern applies among users that gave a different

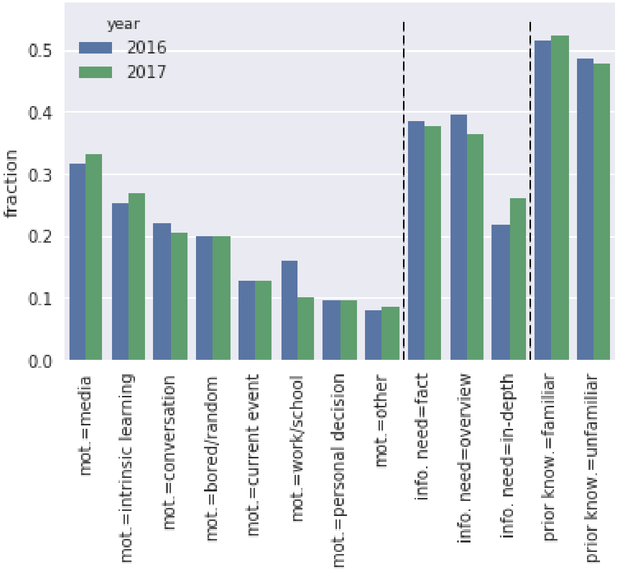


Figure 2: Comparison of English Wikipedia surveys for 2016 (blue) [33] and 2017 (green). We can observe that overall survey responses are stable. Only the share of answers for *motivation = work/school* decreased noticeably.

response to the respective survey question. For example, if for a language the share of users with a *session_length* ≥ 3 is 25% for users with the motivation *work/school,* and 20% for users with another motivation, then the effect of *motivation* = *work*/*school* on *session_length* ≥ 3 is 5%.

To find interesting relationships, we investigate all 247 pairs of binarized usage patterns and survey responses. Given the 14 languages of the survey, we obtain for each pair a distribution of those 14 shares and effects. For summarization, we calculate then the mean share $\mu(S)$ of all languages as well as the relative standard deviation $rs(S)$ (standard deviation divided by the mean share, also known as coefficient of variation) as a measure of variability between the languages. Furthermore, we compute the mean effect $\mu(E)$ of all languages, the relative mean effect (the mean effect divided by the mean share) $\bar{\mu}(E)$, the standard deviation $\sigma(E)$ of the effect, and the normalized standard deviation $\bar{\sigma}(E)$ of the effect (the standard deviation of the effect divided by the mean share). Given these statistics, we are then most interested in pairs with a large (relative) mean effect, since these exhibit a strong dependency between use case and usage pattern across languages. Among those pairs, we can then differentiate between the more general (*i.e.*, consistent between language editions) dependencies, which exhibit a low standard deviation of the effect, and correlations that are more specific to certain languages, which exhibit a high standard deviation of the effect.

We can visualize the relationship between a usage pattern and survey response across languages in plots such as shown in Fig. 3. These plots display one point for each language edition. The coordinates of the point mark the probability of the usage pattern given a specific survey answer on the y-axis and the probability of the usage pattern given a different survey answer for this question on the x-axis. Positive effects of the survey answer on the usage pattern are then indicated by points above the diagonal (*e.g.*, all points in Fig. 3a), negative effects by points below the diagonal. The further the point for a language is from the diagonal, the stronger is the effect of the respective survey response on the usage pattern.

Table 2 shows the top pairs sorted by the relative mean effect. For example, the top pattern (visualized also in Fig. 3a) shows that on average across languages 13.6% of the users arrived on the surveyed page with an internal referrer, *i.e.*, probably by browsing Wikipedia through links [7]. If according to the survey the user is bored or is randomly exploring Wikipedia, then the likelihood of an internal referrer is strongly increased, on average by 9.5% percentage points or by 69.7%. The high relative standard deviation (0.416) indicates strong deviations between the languages.

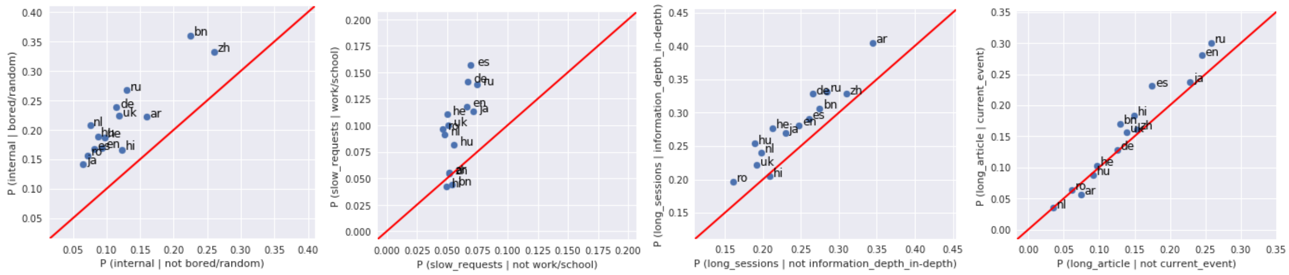
We can find various patterns that are mostly consistent across language editions:

- Users who randomly browse Wikipedia or are bored use internal navigation more often, browse many articles in one session, but

Table 2: Pairs of usage patterns and survey responses with the largest normalized mean effect $\bar{\mu}(E)$ across language editions. This table provides for each pair information on the mean share (likelihood of the pattern) $\mu(S)$, and the relative standard deviation of the share $rs(S)$ across language editions. Furthermore it displays the mean effect (increase of the pattern likelihood in presence of the response) $\mu(E)$, the normalized mean effect $\bar{\mu}(E)$, the standard deviation $\sigma(E)$ and the normalized standard deviation of the effect $\bar{\sigma}(E)$.

Pattern[6]	Response	$\mu(S)$	$rs(S)$	$\mu(E)$	$\bar{\mu}(E)$	$\sigma(E)$	$\bar{\sigma}(E)$
internal	mot.=bored/rand.	.136	.416	.095	.697	.028	.206
slow_requests	mot.=work/school	.065	.220	.038	.594	.030	.457
desktop	mot.=work/school	.342	.303	.187	.547	.122	.358
rapid_requests	mot.=bored/rand.	.102	393	.041	.405	.023	.229
long_sessions	mot.=bored/rand.	.252	.204	.097	.383	.047	.188
time:night	mot.=bored/rand.	.112	.541	.031	.281	.032	.289
long_article	prior knowl.=familiar	.143	.473	.036	.251	.032	.221
time:afternoon	mot.=work/school	.308	.116	.064	.207	.044	.142
time:night	mot.=media	.112	.541	.022	.197	.031	.281
internal	mot.=intrinsic learn.	.136	.416	.022	.163	.018	.131
long_sessions	info. need=in-depth	.252	.204	.040	.158	.019	.075
slow_requests	mot.=other	.065	.220	.009	.140	.021	.324
time:night	mot.=intrinsic learn.	.112	.541	.015	.131	.013	.114
weekday:Friday	mot.=bored/rand.	.113	.238	.015	.131	.018	.155
internal	info. need=in-depth	.136	.416	.017	.127	.015	.112
long_sessions	prior knowl.=familiar	.252	.204	.032	.126	.022	.088
desktop	mot.=other	.342	.303	.042	.124	.058	.169
long_sessions	mot.=intrinsic learn.	.252	.204	.030	.119	.024	.094
time:night	prior knowl.=familiar	.112	.541	.013	.118	.021	.192
long_article	mot.=current_event	.143	.473	.017	.117	.021	.144

[6] Slow and rapid requests are defined based on average amount of time between requests. Slow requests have on average more than 10 minutes in between while rapid request have on average less than 1 minute between requests; Night time describes requests at a local time between midnight and 6 a.m., afternoon between noon and 6 p.m.; long session means at least 3 requests within the survey session; long article denotes articles with at least 40,000 characters.



(a) Effect of motiv. = bored/random on internal referrer

(b) Effect of motiv. = work/school on slow requests

(c) Effect of info. need = in-depth on long sessions

(d) Effect of motiv. = current event on long articles

Figure 3: Visualization of relationships between selected usage patterns and survey answers. Each plot shows one data point for each language, which indicates the likelihood of a usage pattern among users with a given survey response (y-axis) and among user that did not answer that way (x-axis). The diagonal is marked with a red line. This implies that for languages above the read line, answering with this survey response is more likely under this usage pattern.

do not stay long at individual articles, and view Wikipedia at night.

- Users with work- or school-related tasks have longer dwell times on the requested article, use Wikipedia's desktop version more often, and are more likely to visit Wikipedia in the afternoon.
- Similar to bored users, readers motivated by intrinsic learning more often have long sessions and browse at night times using internal navigation.
- By contrast, users who are motivated by a conversation or use Wikipedia for fact checking do so more often using the mobile platform, have shorter dwell times on articles, and use internal navigation less often.
- Users already familiar with a topic have longer sessions and request longer articles.

Most of the above-mentioned dependencies are relatively consistent across language editions, as indicated by a small (normalized) standard deviation of the effect (less than 0.2). This indicates that, independent of the language, certain motivations correlate consistently with certain changes in usage behavior. There are, however, also some specific exceptions. In particular, the effects of work- or school-related motivations appear to differ between language editions (Fig. 3b).

We also notice that, for most of the pairs, the average variability (as measured by the relative standard deviation) between the languages is much higher than the effects of survey responses (*cf.* Fig. 3c and d for typical examples). This observation implies that differences in the use cases alone are insufficient to explain the diverse prevalence of the usage patterns across languages.

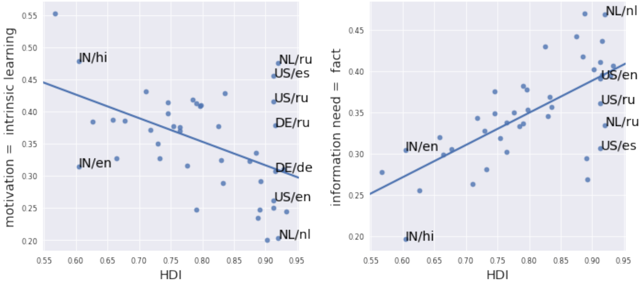
Many noticeable correlations could already be observed in the initial study on the English language. However, not all dependencies found in the English language edition also hold in other languages. For example, it was noted previously that users motivated by current events tend to read longer articles. While this effect is observed in the current survey in the English Wikipedia, it does not hold true considering all language editions (Fig. 3d).

4.3 Survey Responses and Country Statistics

Finally, we analyze if specific Wikipedia use cases can be associated with the socio-economic or cultural background of users in order to seek potential explanations for the differences between language editions. To link available data for these factors to our survey results, we perform these analyses on a country level.

Correlation of survey responses with socio-economic indicators. We start by correlating survey responses with socio-economic information. In particular, we rely on the *Human Development Index* (HDI), a summary statistic that reflects the development status of a country via the population and its capabilities and not only based on economic growth. It is the geometric mean of normalized indices defined under three dimensions: life expectancy, education, and income. The HDI is published by the United Nations Development Programme (UNDP) and is contained in the Quality of Government dataset (Sec. 3.2). Additionally, we also use Gross Domestic Product (GDP) per capita as well as the percentage of adults with secondary education as two other measures of country development. Since a single Wikipedia language can be viewed in multiple countries and the HDI is reported at the country level, we partition the survey responses for each Wikipedia language by country. For each language/country pair with at least 500 survey responses, we then compute the share of each answer option (*e.g.*, *motivation=media*) for survey participants from this country and add country-specific statistics. Through this step, we obtain 43 language/country pairs, which we use as data points for a correlation analysis.

Table 3 shows the Spearman correlation coefficient and its associated p-value (with Bonferroni correction for $n = 13$ survey responses, but no correction for multiple attributes) when capturing the correlation between survey responses and HDI. We observe that several survey answers show significant correlations with HDI. For example, the more developed the country of a reader, the more likely the reader is to be motivated by *media* when visiting Wikipedia. By contrast, being motivated by *work or school* or by *intrinsic learning* is more likely in developing or newly industrialized countries. Regarding the information need of viewers, we can see that in-depth reading is more often reported in less developed countries,



(a) HDI *vs.* intrinsic learning

(b) HDI *vs.* fact checking

Figure 4: Correlation between the Human Development Index (HDI) and survey responses. Labels mark data points for countries with multiple language editions, capital letters encode the country, lower case letters the Wikipedia edition.

while in industrialized countries fact checking is a more prevalent use case. Finally, reporting familiarity with a topic is also somewhat more common in industrialized countries. We can conclude that there is a clear tendency towards in-depth reading and learning in less developed countries.

By zooming in on individual correlations, *e.g.*, between HDI and *intrinsic learning* (Fig. 4), we obtain additional insights. We observe that in industrialized countries (United States, the Netherlands, Germany), linguistic minorities (such as Spanish speakers in the United States or Russian Speakers in Germany), use Wikipedia more often for intrinsic learning.

Following the interesting correlations with the HDI, we also tried other socio-economic indicators such as the *GDP per capita*, share of the adult population with *secondary education* (also shown in Table 3), the *Gini coefficient* of the income distribution and *Internet availability* (not shown due to space limitations). These indicators correlate strongly with HDI and each other. Thus, they also exhibit very similar correlations with the survey response. Since the correlations are somewhat weaker than for the HDI, we cannot identify a single component of the compound HDI measure that appears most relevant for the correlation with Wikipedia readers' intent.

Human Development Index *vs.* topics. Next, we investigate if the differences in use case prevalences in countries with diverse socio-economic preconditions also manifest themselves in different topics being viewed. For this purpose, we focus on the Spanish Wikipedia edition, since it is viewed from many countries with diverse HDI scores. As this analysis does not directly require survey data, we take all Wikipedia readers of our random sample of users and group them by the country their requests came from. By doing so, we obtain data from 24 countries with more than 500 users each. We then compute, for all these countries, the viewing likelihood of each of the 20 topics computed via Latent Dirichlet Allocation (Sec. 3.2).

We observe that several topics exhibit significant correlations with the HDI of the reader's country. The topics *Math, Physics & Technology* (Spearman's $r_s = -0.75$, Bonferroni-corrected p-value $p < 0.001$), *Research & Education* ($r_s = -0.73$, $p < 0.001$), and *Medicine & Biology* ($r_s = -0.71$, $p \approx 0.002$) show the strongest negative correlations, *i.e.*, these topics are more often viewed in less developed countries. By contrast, topics such as *Media Culture* ($r_s = 0.71$, $p \approx 0.002$) and *Numbers, Lists & Sports* ($r_s = 0.60$, $p \approx 0.03$) show a significant positive correlation, *i.e.*, articles on those topics are more commonly requested by readers in industrialized countries. Overall, we observe the tendency that entertainment-oriented topics are more popular in countries with a high HDI, while science-oriented topics are more prevalent in less developed countries.

For the English Wikipedia, we can also find differences between topics across the countries, but the correlations show a less clear picture, partly because many topics obtained from LDA are focused on articles with a specific regional background. For example, the topic most strongly correlated with the HDI is *Geography/US* ($r_s = 0.57$, $p < 0.001$), while the strongest negative correlation is for the topic *Asia* ($r_s = -0.53, p = 0.004$).

Correlation with cultural dimensions. To investigate cultural influences on reading behavior, we use Hofstede's cultural dimensions, a well-established and comprehensive framework for characterizing national cultures [11, 12]. Hofstede's framework utilizes six dimensions: *Power Distance*, *Individualism*, *Uncertainty Avoidance*, *Masculinity*, *Long-Term Orientation*, and *Indulgence*.

Correlating country-level measures for these dimensions with the survey responses (Table 3, two rightmost columns) shows primarily weak to moderate correlations. An exception is the *Individualism* (IDV) dimension, for which we observe a clear association between Wikipedia visits motivated by media or work/school: in countries with more collectivist societies (low individualism score) people are less likely to be motivated to visit Wikipedia by media,

Table 3: Correlation between survey responses with socio-economic and cultural indicators on a country level, *i.e.*, the Human Development Index (HDI), the GDP per capita, the share of adult population with secondary education, as well as Hofstede's Long-Term Orientation (LTO) and Individualism (IDV) dimensions. The table reports the Spearman correlation coefficient, asterisks indicate the p-value of the coefficient under the null hypothesis of independence of the data points ($^{***}< 0.001$,$^{**}< 0.01$, $^{*}< 0.05$). The table is sorted by correlation with HDI.

	Response	HDI	GDP p. cap.	Second. educ.	LTO	IDV
Motivation	media	0.63***	0.58***	0.42	0.39	0.63***
	work/school	-0.55**	-0.54**	-0.40	-0.37	-0.77***
	current event	-0.45*	-0.48*	-0.20	0.13	-0.38
	intrinsic learning	-0.40	-0.43	-0.20	0.00	-0.26
	personal decision	-0.28	-0.32	-0.08	0.31	-0.14
	other	0.26	0.35	-0.08	-0.37	0.04
	bored/random	0.21	0.25	-0.02	-0.17	0.17
	conversation	-0.07	-0.12	-0.02	0.22	0.13
info. need	fact	0.66***	0.62***	0.55**	0.36	0.53*
	in-depth	-0.60***	-0.57*	-0.46*	-0.23	-0.43
	overview	0.25	0.27	0.11	-0.13	0.06
prior knowl.	familiar	0.44*	0.39	0.47*	0.27	0.42
	unfamiliar	-0.44*	-0.39	-0.47*	-0.27	-0.42

while using Wikipedia motivated by work or school related tasks is significantly more likely.

4.4 Summary of Results

Survey responses. We have shown that Wikipedia is read for a variety of use cases across the 14 Wikipedia languages, and no one use case dominates the others. Moreover, we have observed that the prevalence of Wikipedia use cases differs significantly across languages. More specifically, we observe that *intrinsic learning* is the most commonly reported motivation for visiting Wikipedia, except in a minority of languages (including English) where *media* is most common. We also show that the motivations *work/school* and *bored/random* have the highest prevalence discrepancies. We observe that information need and prior knowledge vary to a great extent across languages, with the reported *in-depth* reading ranging from 21% to over 60% and *familiarity* from less than 40% to more than 70%.

Through the analysis of the survey results we show that the English Wikipedia—the sole focus of many Wikipedia studies— is not representative of all Wikipedia languages. Rather, it can be considered an outlier with regard to several aspects. Finally, the survey results for the English Wikipedia line up well with the previous study, providing evidence for the robustness of results over time.

Usage patterns. By connecting survey responses to request logs we have identified several usage patterns in the logs that can be consistently associated with certain Wikipedia use cases across languages. Specifically, motivation *bored/random* can be linked to certain patterns including long sessions with rapid requests and internal browsing, while *work/school* can be linked to slow requests to desktop versions of Wikipedia. We also observe that not all patterns discovered for English Wikipedia use cases [33] hold for the other Wikipedia languages. Furthermore, the different use cases in the languages alone are not sufficient to explain the differences in usage patterns across languages.

Country-level statistics. We find significant correlations between the Human Development Index of a country and the prevalence of Wikipedia use cases reported from there. In particular, less developed countries are more likely to read Wikipedia *in depth* and be motivated by *work/school* or *intrinsic learning*. In industrialized countries, Wikipedia readers are more often checking *facts* and are triggered to visit Wikipedia by *media*. Socio-economic differences also show in different topics being viewed: science-oriented topics are more important in less developed countries, while entertainment-oriented topics are more common in industrialized countries. Cultural factors as measured by Hofstede's cultural dimension, with the exception of *Individualism,* seem to play a lesser role.

5 DISCUSSION

In this section, we present the implications of this study, future directions, and methodological limitations.

5.1 Implications and Future Directions

Beyond English Wikipedia. A tremendous amount of research and development on Wikipedia has been focused on, or informed by, English Wikipedia. This study sheds light on the importance of breaking this cycle and acknowledging that English Wikipedia is not representative of all Wikipedia languages, and in several aspects is an outlier. Wikipedia's endeavor towards knowledge equity[7] requires a deeper and better understanding of the differences between Wikipedia languages. This work should be expanded to enhance our understanding of the access to, and production of, knowledge in Wikipedia. Future studies can investigate the socio-economic factors at a finer granularity than the country level, include demographic information, and attempt to characterize potential Wikipedia readers.

Global *vs.* local solutions. One of the findings of the current study is that, except for a few general patterns, the patterns that describe readers' use cases of Wikipedia differ across languages. This indicates that one-size-fits-all solutions may not work across languages, and a combination of global and local solutions may be needed to satisfy the needs of Wikipedia readers. Future research should focus on scaling locally aware solutions across many languages.

Within-session language switching. Our preliminary analysis shows that on average roughly 20% of reader sessions involve the reader switching from one Wikipedia language to another. Future work can investigate circumstances under which users switch from one language to another, which can in turn inform the prioritization of content creation across Wikipedia languages.

5.2 Methodological Limitations

User identification. Wikipedia does not require users to log in, nor does it use cookies in webrequest logs to maintain a notion of unique clients. Therefore, we rely on a combination of IP addresses and user agents to approximate unique devices (*cf.* Singer *et al.* [33] for an in-depth discussion of this approximation and its limitations).

Survey-response bias. Not all users are equally likely to participate in a voluntary survey, *i.e.*, some groups will be overrepresented in the survey responses. To tackle this issue, we reweighted the responses based on features from the server logs by inverse propensity weighting. However, if other covariates (*e.g.*, age or gender) that are not explicit in the server logs influence the responses, these might skew the results.

Translation bias. Multilingual surveys suffer from differences in the translations of survey questions and answer options. For this study, translations were carefully done by Wikipedia editors who are native speakers of the languages in this study. Translated content was then checked word by word in online meetings between the translator(s) and one of the study authors. Even with a process such as above, we cannot rule out different nuances and connotations between the languages that may not have been captured as part of the translations.

Social desirability. Survey responses are commonly subject to *social desirability bias* [23], *i.e.*, participants are more likely to reply

[7] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction#Our_strategic_direction:_Service_and_Equity

with options that are viewed in a positive light in their society. Even though the questions in our survey are of non-sensitive nature and the survey is done anonymously, social desirability could still influence our results; *e.g.*, browsing Wikipedia out of boredom could be seen as negative in some cultures and therefore might be picked less often by survey participants. As the effect of such a bias could be different in different societies, this might skew comparisons between languages.

6 CONCLUSIONS

In this work, we study why users from around the world read Wikipedia. Through a large-scale survey with more than 210,000 responses across 14 Wikipedia languages we highlight key commonalities and differences in Wikipedia use cases across these languages. Combining the survey responses with webrequest logs as well as country level socio-economic statistics allows us to characterize Wikipedia use cases across languages with behavioral patterns and socio-economics indicators. The outcomes of this study provide a deeper understanding of Wikipedia readership in a wide range of languages, which is important for Wikipedia editors, developers, and the reusers of Wikipedia content.

ACKNOWLEDGEMENTS

We thank the following Wikipedia users for translating the survey to their languages and acting as our points of contact in their language communities: عباد _ديرانية, Amire80, Antanana, AWossink, Hasive, Kaganer, Lyzzy, Rasco, Satdeep_Gill, Shangkuanlc, Strainu, Tgr, Whym. We thank Philipp Singer and Ellery Wulczyn for contributing their code of the predecessor study, Markus Strohmaier for helpful discussions, and Bahodir Mansurov for running the surveys on Wikipedia.

REFERENCES

[1] Ofer Arazy, Hila Lifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. 2017. On the "How" and "Why" of Emergent Role Behaviors in Wikipedia. In *Conference on Computer-Supported Cooperative Work and Social Computing*.

[2] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46, 3 (2011), 399–424.

[3] Amit Basu. 2003. Context-driven assessment of commercial web sites. In *International Conference On System Sciences*.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[5] Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62, 10 (2011), 1899–1915.

[6] Shaoyong Chen, Huanming Zhang, Min Lin, and Shuanghuan Lv. 2011. Comparision of microblogging service between Sina Weibo and Twitter. In *International Conference on Computer Science and Network Technology*.

[7] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. 2017. What Makes a Link Successful on Wikipedia?. In *International Conference on World Wide Web*.

[8] Henry A Feild, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *International Conference on Research and Development in Information Retrieval*.

[9] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in twitter: Time, individualism and power. In *International Conference on Web and Social Media*.

[10] Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *International Conference on Communities and Technologies*.

[11] Geert Hofstede and Michael H Bond. 1984. Hofstede's culture dimensions: An independent validation using Rokeach's value survey. *Journal of cross-cultural psychology* 15, 4 (1984), 417–433.

[12] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind* (3rd ed.). McGraw-Hill, USA.

[13] Yuncheng Jiang, Wen Bai, Xiaopei Zhang, and Jiaojiao Hu. 2017. Wikipedia-based information content and semantic similarity computation. *Information Processing & Management* 53, 1 (2017), 248–265.

[14] Imrul Kayes, Nicolas Kourtellis, Daniele Quercia, Adriana Iamnitchi, and Francesco Bonchi. 2015. Cultures in community question answering. In *Conference on Hypertext & Social Media*.

[15] Willi Klösgen. 1996. Explora: A Multipattern and Multistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 249–271.

[16] Leantros Kyriakoullis and Panayiotis Zaphiris. 2016. Culture and HCI: a review of recent cultural studies in HCI and social networks. *Universal Access in the Information Society* 15, 4 (2016), 629–642.

[17] Paul Laufer, Claudia Wagner, Fabian Flöck, and Markus Strohmaier. 2015. Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. In *Web Science Conference*.

[18] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. 2014. Reader preferences and behavior on Wikipedia. In *Conference on Hypertext and Social Media*.

[19] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23, 19 (2004), 2937–2960.

[20] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. 2012. New kid on the block: Exploring the google+ social graph. In *Internet Measurement Conference*. 159–170.

[21] Marc Miquel-Ribé and David Laniado. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Digital Humanities* 5 (2018), 12.

[22] Lien Le Monkhouse, Bradley R Barnes, and Thi Song Hanh Pham. 2013. Measuring Confucian values among East Asian consumers: a four country study. *Asia Pacific Business Review* 19, 3 (2013), 320–336.

[23] Anton J Nederhof. 1985. Methods of coping with social desirability bias: A review. *European journal of social psychology* 15, 3 (1985), 263–280.

[24] Oded Nov. 2007. What motivates Wikipedians? *Communications of the ACM* 50, 11 (2007), 60–64.

[25] Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same? Characterizing twitter around the world. In *International Conference on Information and Knowledge Management*.

[26] Lin Qiu, Han Lin, and Angela K-y Leung. 2013. Cultural differences and switching of in-group sharing behavior between an American (Facebook) and a Chinese (Renren) social networking site. *Journal of Cross-Cultural Psychology* 44, 1 (2013), 106–121.

[27] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking gross community happiness from tweets. In *Conference on Computer Supported Cooperative Work*.

[28] Daniele Quercia and Diego Saez. 2014. Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing* 13, 2 (2014), 30–36.

[29] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. 2010. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters* 105, 15 (2010), 158701.

[30] 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 *International Conference on Web an Social Media*.

[31] Thiago Silva, Pedro Vaz de Melo, Jussara Almeida, Mirco Musolesi, and Antonio Loureiro. 2014. You Are What You Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food and Drink Habits in Foursquare. In *International Conference on Web an Social Media*.

[32] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. 2014. Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PloS One* 9, 7 (2014), e102070.

[33] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In *International Conference on World Wide Web*.

[34] Anselm Spoerri. 2007. What is popular on Wikipedia and why? *First Monday* 12, 4 (2007).

[35] Jan Teorell, Stefan Dahlberg, SÃűren Holmberg, Bo Rothstein, Anna Khomenko, and Richard. Svensson. 2017. *The Quality of Government Standard Dataset, version Jan17*. University of Gothenburg: The Quality of Government Institute.

[36] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring urban deprivation from user generated content. In *Conference on Computer Supported Cooperative Work & Social Computing*.

[37] Robert West and Jure Leskovec. 2012. Human wayfinding in information networks. In *International Conference on World Wide Web*. 619–628.