Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia

Luís Couto
Faculty of Engineering of the University of Porto
Porto, Portugal
mieic1204994@fe.up.pt

Carla Teixeira Lopes
INESC TEC, Faculty of Engineering of the University of Porto
Porto, Portugal
ctl@fe.up.pt

ABSTRACT

Wikipedia is a free, multilingual, and collaborative online encyclopedia. Nowadays, it is one of the largest sources of online knowledge, often appearing at the top of the results of the major search engines, being one of the most sought-after resources by the public searching for health information. The collaborative nature of Wikipedia raises security concerns since this information is used for decision-making, especially in the health area. Despite being available in hundreds of idioms, there are asymmetries between idioms, namely regarding their quality. In this work, we compare the quality of health information on Wikipedia between idioms with 100 million native speakers or more, and also in Greek, Italian, Korean, Turkish, Persian, Catalan and Hebrew, for historical tradition. Quality metrics are applied to health and medical articles in English, maintained by WikiProject Medicine, and their versions in the above idioms. With this, we contribute to a clarification of the role of Wikipedia in the access to health information. We demonstrate differences in both the quantity and quality of information available between idioms. English is the idiom with the highest quality in general. Urdu, Greek, Indonesian, and Hindi achieved lower values of quality.

CCS CONCEPTS

• **Applied computing** → **Consumer health**; Health informatics;
• **Information systems** → **Wikis**.

KEYWORDS

Information Quality, Wikipedia, Health information, Multilingual information access

ACM Reference Format:
Luís Couto and Carla Teixeira Lopes. 2021. Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3479986.3480000

1 INTRODUCTION

Online health information-seeking behavior improves the patient-physician relationship and patients' engagement in health decision-making [32]. Three health-related terms are in the top ten Google searches [13] for 2020 - "coronavirus", "coronavirus update" and "coronavirus symptoms". A study conducted by the Health On the Net Foundation [27] shows that when information needs relate to health, 44% of respondents admitted looking for this information more than three times a week, with the main point of access being search engines, which may end up leading them later to Wikipedia [18]. The same study revealed that the quality of information remains the most significant barrier encountered by respondents (80%) when searching for health information online, and the factor most valued regarding the quality of information is the reliability/credibility (96%). Different authors have concluded that Wikipedia is a reliable source of health information in surgical information [6], pediatric otolaryngology [35], pharmacology [17] and cancer [28].

Nowadays, the most popular health article - "COVID-19 pandemic", has, on average, more than 40 thousand daily views [43]. Wikipedia was created in English, and the second idiom after that was German, followed immediately by Catalan, these remaining the only idioms for two months. By the end of the first year, Wikipedia had articles written in 18 different idioms. It is currently available in 321 idioms, with 310 of them active [39].

Given existing differences in access to health information among speakers of different idioms [1], Wikipedia can potentially reduce or accentuate this imbalance. In this work, we will compare the quality of information available to users speaking different idioms, checking if it is similar between the different versions or whether there are disparities, and if so, to be able to quantify them.

Section 2 describes how quality is assured in an open, collaborative resource such as Wikipedia. In Section 3, we describe differences between idioms at Wikipedia. Our methodology to compare the quality of Wikipedia health contents in several idioms is described in Section 4. Results and respective discussion are presented in Sections 5 and 6, respectively. Finally, conclusions are presented in Section 7.

2 WIKIPEDIA INFORMATION QUALITY

Quality has always been a concern for Wikipedia, which has established frameworks to ensure it ever since its creation. Wikipedia currently has more than 400 million articles, so assessing the quality of so much information asks for automation. Several authors have addressed this issue, one of the most prominent being Stvilia *et al.* [30].

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09.
https://doi.org/10.1145/3479986.3480000

2.1 Wikipedia internal quality mechanisms

Despite the large number of articles created initially, they did not have the desired quality, which led Larry Sanger to define rules published on the Wikipedia pages "Wikipedia is not a dictionary"[1] and "What Wikipedia is not"[2], which still exist, with changes over time. In this context, five principles define the rules and recommendations for preparing content [42]. The first principle states that "Wikipedia is an encyclopedia", pointing that it combines features of encyclopedias, almanacs, and gazetteers; the second principle refers that "Wikipedia is written from a neutral point of view", indicating that articles should have an impartial tone, documenting and explaining significant points of view; the following principle declares that "Wikipedia is free content that anyone can use, edit, and distribute", evidencing that authors freely license their work to the public; the fourth principle expresses that "Wikipedia's editors should treat each other with respect and civility", denoting the etiquette all users should use; the last principle states that "Wikipedia has no firm rules", signifying that Wikipedia policies and guidelines are flexible and mutable over time.

These five pillars are common to Wikipedias in different idioms, but policies are defined for each version. These policies are created by the community, by consensus or by vote, with a transversal character to all the articles present and all its users. There are sanctions for those who violate them, such as blocking users for some time [40].

There are control mechanisms to ensure compliance with these principles that can be summarized into nine types. First, there are many users, where the well-meaning vastly outnumber the malicious, with their unique characteristics working together for a typical result. It is the supervision of users. Next, many editors guarantee neutrality and different points of view on the one hand and, on the other hand, ease in repairing errors. This is the collaborative knowledge construction mechanism. The next control mechanism relates to the fact that there is only one page for everyone, pressuring for a consensus among all and the desired neutrality. Also noteworthy is that no superior entities control the content, avoiding manipulations motivated by secondary interests. It is the wiki structure. Another control mechanism relates to the rules, policies, and principles, defined to ensure good conduct on the one hand and ensure on the other the disruptive potential necessary for evolution. It is the respect for policies and principles. For the next control mechanism, we refer to the concerns and opinions of minorities, taken into account in trying to reach a decision that reflects the values of the community. It is the consensus-based *ethos*. There are intrinsic escalation mechanisms, such as that users will more closely watch items that are more prone to vandalism to stop it. There are also extrinsic mechanisms, such as the possibility that anyone can request disputes in progressive stages. This control mechanism is the escalation and dispute resolution processes. The next mechanism refers to the software tools used by the most active editors, such as Huggle[3], to automatically detect vandalism in real-time, among other tools facilitating the identification and correction of quality problems. It is the software facilitating monitoring and editing control mechanism. Tools exist to block problematic publishers and protect pages from low-quality publishers, capable of filtering combinations of accounts or IP addresses. This control mechanism refers to blocking and protection systems. Finally, inline tags can be used in the text to individual flag statements, individual statements, quotes, or articles as a whole, request verification or citation, and indicate to other users that a fact or presentation is not supported as is. It is the categorization of information control mechanism.

The various versions of Wikipedia generally have an article quality assessment system [41] that is not standardized. For example, in the English version, this system is based on letters that indicate how complete an article is, taking into account different factors. WikiProjects[4] members assess quality using tags that can be used to generate statistical data about the articles. These assessments make it possible to determine the quality of the information in specific areas and prioritize articles for improvement according to expectations. It should be noted that this evaluation has no official character. In addition, there may also be a ranking of the priority or importance of an article, reflecting the level of expectation or desire that a particular topic is portrayed. The scale generally ranges from "unimportant" to "extremely important". This importance rating is also relative to each WikiProject.

2.2 Metrics for assessing quality

To assess Wikipedia information quality, authors propose different metrics based on different features. Generically, Wu *et al.* [44] used four groups of metrics, with a total of 28 metrics: lingual - e.g., readability; structural - e.g., links; historical - e.g., article age and reputational - e.g., amount of editors. Li *et al.* [20] and De La Robertie *et al.* [5] proposed metrics based on the relationship between articles and their editors. Marrese-Taylor *et al.* [22], in 2019, based their work on the articles' editions, also considering the description of each edition.

In the health area, there have been other approaches by authors such as Thomas [33], in 2013, using: comprehensibility - the ratio of medical codes in articles; trust - number of references in articles and readability. In 2014, Conti *et al.* [3] assessed 2,400 medical articles using metrics from types: lingual - Flesch Reading Ease and Flesch-Kincaid scales; structural - e.g., number of links and citations; historical - e.g., number of editions and number of editors; reputational: e.g., age of editors and duration of editions. Modiri *et al.* [25] assessed articles in the neurosurgery area, using readability indexes, the "Center for Disease Control Clear Communication Index"[5] and DISCERN[6]. Later, in 2019, Suwannakhan *et al.* [31] assessed the information quality in anatomy, with readability indexes and DISCERN in association with Wikimedia X-tools [7]. In the same year, Domingues and Teixeira Lopes [7] compared the quality of the Portuguese version with the Anglophone version of Wikipedia in articles related to medicine. They used metrics defined by Stvilia *et al.* [30] and more specific metrics, such as the number of medicine templates, number of medicine infoboxes, and number of citations. Later, in 2021, Couto and Teixeira Lopes [4]

[1] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary
[2] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not
[3] https://en.wikipedia.org/wiki/Wikipedia:Huggle
[4] https://en.wikipedia.org/wiki/Wikipedia:WikiProject
[5] https://www.cdc.gov/ccindex/index.html
[6] http://www.discern.org.uk/discern_instrument.php
[7] https://www.mediawiki.org/wiki/XTools

evaluated the quality of health-related Wikipedia articles, using the same metrics from Stvilia, with the proposal to add health-related features of Wikipedia articles, such as health templates, medical codes, or recommended sections.

Stvilia *et al.* [30] defined seven metrics: authority, completeness, complexity, informativeness, consistency, currency, and volatility to assess the Wikipedia quality. These metrics use 19 features from Wikipedia articles and their history. We consider the metrics and their features, as defined by the authors, complete and comprehensive, as they include several dimensions of the quality of information on Wikipedia. They are also, specific in the form of calculation. They are, therefore, a reference for other authors in different works [2, 7, 15, 19, 44]. Considering this, we use the same metrics and respective features as described in Section 5.

3 DIFFERENCES BETWEEN IDIOMS

There are currently 7,139 living idioms worldwide [9]. Such diversity can naturally raise questions about their presence on the web. In 2009, Pimienta *et al.* [26] described an investigation carried out from 1996 to 2008 by UNESCO through FUNREDES and Union Latine on linguistic diversity on the web that revealed a large discrepancy in the presence of idioms in cyberspace.

Based on language as the primary mean of communication, Wikipedia is an indicator of online multilingualism across the range of idioms present. The Wikimedia Foundation has defined policies for proposing new idioms, created by its “Language committee” [37], responsible for processing the proposals and associated projects. Proposed idioms must not yet exist on Wikimedia, must have a valid ISO 639-1 code, and must have a sufficient number of fluent users to form a viable community of contributors and audiences interested in its content. Regional dialects and different forms of it are excluded. For approval, it is also a requirement that a test project exists on Wikimedia and that there is an ongoing effort to translate the Wikimedia interface into that idiom. There are currently 273 requests for new idioms to be added to Wikipedia [38].

3.1 Wikipedia articles and users

Using the statistical data provided by Wikipedia [39] as of May 2021, Table 1 was adapted, showing the number of articles, edits to articles, administrators, and active users for each version of Wikipedia in idioms with more than 1 million articles.

We can conclude that there are currently 18 idioms on Wikipedia with more than 1 million articles. The superiority of the English idiom in terms of the number of articles available is quickly confirmed. Cebuano, an Austronesian idiom spoken in the Philippines by 27.5 million people in 2020, follows. Cebuano’s popularity happens because a bot, Lsjbot, created more than 17 million articles, accounting for most of the articles written in Cebuano, Swedish, and Waray, which also explains why Swedish comes third, followed only then by German. This bot activity also explains why the number of articles does not keep up with the other metrics in the table, where English also stands out, followed by German and French.

In 2009, Dijk [34] addressed Wikipedia edits in minority idioms and ways to measure them for comparison. He mentions the obsession with the number of articles in each Wikipedia and the comparison with other idioms. He concluded that it is difficult to attribute the factors that contribute to the growth of each version of Wikipedia but emphasizes the number of speakers, as they represent the potential article editors of that idiom. This, however, not always corresponds to reality [17]. In 2017, Matei [24], using data from edits from the first decade of Wikipedia’s existence, concluded that only 1% of the editors created 77% of the articles, which raises problems about its collaborative spirit. Dijk mentions the importance of people’s attitude towards projects such as Wikipedia, pointing this as the main factor for the growth of Latin idioms in Wikipedia. He concludes with the importance of the collaboration of institutions related to idiom issues in content development, especially in minority idioms. Later in 2011, Hale [14] studied the role of multilingual editors as enablers of the development of the various idioms within Wikipedia.

Table 1: Wikipedia statistics for idioms with more than 1 million articles

	Articles	Edits	Admins	Active users
English	6,296,349	1,018,157,853	1,096	138,226
Cebuano	5,729,196	31,370,481	6	170
Swedish	3,187,113	49,166,960	63	2,617
German	2,575,270	210,442,253	187	20,382
French	2,327,445	182,377,650	156	21,964
Dutch	2,054,789	58,767,964	35	4,343
Russian	1,723,112	113,865,981	79	11,723
Italian	1,692,357	120,298,053	114	9,911
Spanish	1,682,915	135,034,634	67	17,133
Polish	1,473,158	63,082,287	102	4,802
Egyptian Arabic	1,283,253	5,590,058	6	210
Japanese	1,267,954	83,259,311	41	15,260
Waray	1,265,315	6,233,530	3	76
Vietnamese	1,263,818	64,842,686	20	2,300
Chinese	1,196,344	65,280,127	79	8,365
Arabic	1,115,708	53,691,546	27	5,422
Ukrainian	1,091,668	31,568,224	45	3,442
Portuguese	1,066,210	60,993,351	71	10,358

Source: adapted from https://en.wikipedia.org/wiki/List_of_Wikipedias

The Wikidata[8] inter-language system is a system launched in 2012 by Wikimedia Foundation, which together with the interlinguistic links[9] provides a centralized solution based on a collaborative database. It allows connecting the same concept across multiple versions of Wikipedia and even between other Wikimedia projects. Essentially, items are stored, each with a label, a description, and a list of alternative names, linking the items and their data together. Hale found that most editors are active in only one idiom, with 15% doing so in different idioms, and they are usually more active than others.

3.2 Content quality

In 2009, Filatova [10] described the multilingualism of Wikipedia through a framework created for this purpose and using only the text of the articles. The author mentions that articles about the same thing differ a lot between versions, especially in terms of the amount of information covered in each version and the aspects that authors choose to cover about the general topic of the article, directly affecting its quality. Domingues and Teixeira Lopes [7] conducted a comparative study on the quality of medicine-related

[8] https://www.wikidata.org/wiki/Wikidata:Wikidata_Concepts_Monitor
[9] https://en.wikipedia.org/wiki/Help:Interlanguage_links

articles in the Portuguese and English versions of Wikipedia in 2019. The authors found significant differences between the two versions in the vast majority of the metrics evaluated. The results suggest that English articles demonstrate more significant effort in content organization, information reuse, and citation usage. The overall conclusion is that Wikipedia's English health contents are substantially better in terms of quality.

Despite the scarce research available on the differences in content quality between different Wikipedia versions according to idioms, there seems to be a direct relationship between the quantitative and qualitative aspects of the information available. Assuming that idioms with lower quantitative expression in Wikipedia translate lower quality information, and given the importance of Wikipedia as a source of information, this is an inequality problem that has received the attention of UNESCO, which recognizes that the information present in cyberspace is a significant factor for the development of humanity, as it is a primary way of sharing information and knowledge.

4 METHODOLOGY

Our approach has five major steps, schematized in Figure 1. Numbers identify the execution sequence, and arrows identify information flow.

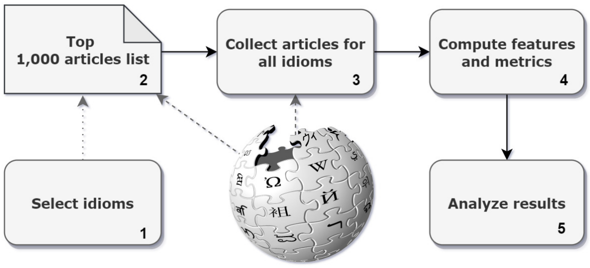

Figure 1: Methodology

We began by selecting the idioms for our dataset. Then, we collected a list of health-related articles. These first steps are described in Subsection 4.1. In the next step, we gathered the articles in that list for all idioms, as described in Subsection 4.2. After that, we assessed the quality of the articles, as described in Section 5 and finally, we discussed the results, as described in Section 6.

4.1 Idioms selection

We have selected idioms available on Wikipedia with at least 100 million speakers as a native or second idiom. We also extended this collection to six other idioms for their cultural or medical importance, namely their neurosurgical tradition since ancient times: Greek, Italian, Korean, Turkish, Persian, and Hebrew.

In Table 2, we can see the number of speakers for each idiom of our dataset, as first or second idiom, sorted by decreasing number of total speakers. English is the idiom with most speakers, mostly as a second idiom. Chinese follows closely, mostly as first idiom. In third place, Hindi comes with a significant difference from the previous two. Hebrew and Catalan are the less spoken idioms in our dataset, with only 9 million total speakers.

Table 2: Number of speakers for each idiom of our dataset

	First idiom	Second idiom	Total
English (en)	369.9 million	978.2 million	1.348 billion
Chinese (zh)	921.2 million	198.7 million	1.120 billion
Hindi (hi)	342.2 million	258.3 million	600 million
Arabic (ar)	-	-	274 million
Bengali (bn)	228.7 million	39.0 million	268 million
French (fr)	79.6 million	187.4 million	267 million
Russian (ru)	153.7 million	104.3 million	258 million
Portuguese (pt)	232.4 million	25.2 million	258 million
Urdu (ur)	69.0 million	161.0 million	230 million
Indonesian (id)	43.6 million	155.4 million	199 million
German (de)	76.6 million	58.5 million	135 million
Japanese (ja)	126.3 million	121,500	126 million
Turkish (tr)	82.2 million	5.9 million	88 million
Persian (fa)	56.3 million	17.9 million	74 million
Korean (ko)	-	-	82 million
Italian (it)	64.8 million	3.1 million	68 million
Greek (el)	-	-	13 million
Hebrew (he)	-	-	9 million
Catalan (ca)	-	-	9 million

4.2 Data collection

Our selection of health-related articles was based on a list maintained by WikiProject Medicine [43]. This list contains the 1,000 most viewed articles for the English Wikipedia.

First, all articles written in English were collected from the mentioned list. Data for articles written in other idioms other than English was obtained by following the idiom link in each of the English articles, and each of them was iteratively collected.

We used the MediaWiki API to collect the article's contents and metadata, revision history, idiom links, internal links, and external links, following the approach of Domingues and Teixeira Lopes [7]. Other data was obtained from the article's markup. We also obtained images through markup because the API does not distinguish content images from others, such as Media Wiki and Wikimedia logos. Templates, infoboxes, and citations were also collected from the article's markup. To compute some measurements, such as readability scores, "InfoNoise", or the article's length, we removed all the markup from the article's content to obtain the required plain text. We faced some challenges when collecting data because there is considerable heterogeneity among the idioms chosen. There is also heterogeneity between the different versions of Wikipedia for each idiom. Moreover, there is also heterogeneity within each Wikipedia version, as edits are made by several users, who do not always comply with the established standards when they exist.

As some articles only have versions in some idioms, the complete dataset consists of 14,456 articles. The distribution of the articles by idiom is visible in Figure 2. This figure also includes the distribution of all Wikipedia articles by idiom. Figure 2 shows that English is the only idiom with 1,000 articles, meaning that no other idiom has the corresponding version for all the articles in the top list. It is also evident the prominent differences between idioms, where some only have about half, or even less, of the total number of articles, such as the Urdu idiom. Analyzing the relation of the dataset number of

Table 3: Idioms quality assessment for authority metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	2033.05	1196.7	de	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn	18
German	de	1315.5	416.7	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn		17
Russian	ru	1250.8	265.6	zh	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Italian	it	1240.2	254.7	zh*	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Chinese	zh	1230.4	344.5	hi	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn							12
French	fr	1189.8	233.8	hi	pt*	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn						13
Hindi	hi	1159.6	546.1	pt	ur	id	ko*	el	bn													6
Portuguese	pt	1152.6	273.5	ar	ja	ca	ur	fa	id	ko	el	bn										9
Turkish	tr	1148.5	539.6	ur	fa*	id	ko	el	bn													6
Hebrew	he	1139.3	413.5	ur	fa*	id	ko	el	bn													6
Arabic	ar	1130.3	603.0	ur	id	ko	el	bn														5
Japanese	ja	1128.4	186.8	ur	id	ko	el	bn														5
Catalan	ca	1101.6	468.0	ur	id	ko	el	bn														5
Urdu	ur	1096.6	994.0	fa																		1
Persian	fa	1087.2	490.4	id	ko	el	bn															4
Indonesian	id	1034.5	1086.8																			0
Korean	ko	1024.3	956.2	bn*																		1
Greek	el	800.8	1069.9																			0
Bengali	bn	710.1	1070.2																			0
$\chi 2$		3543.3																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the remaining values

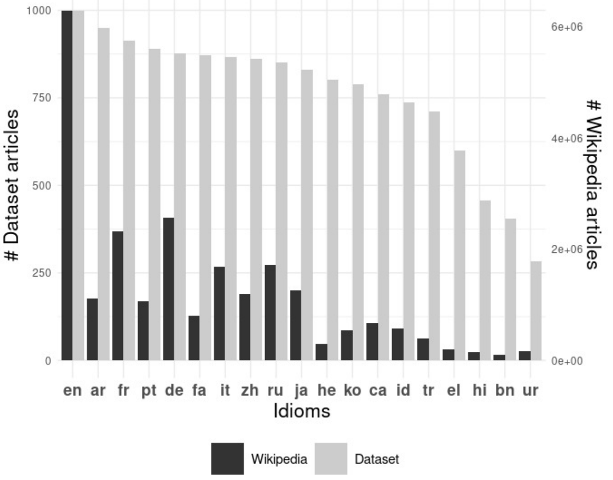

Figure 2: Number of articles by idiom

articles and the total number of articles in each version of Wikipedia, we can observe different distributions. Given that the dataset only contains health-related articles, this data suggests that the size of each version of Wikipedia is not directly related to the number of health-related articles. These data may also point to differences in the importance of the health-related area in each version of Wikipedia.

The datasets and code developed in this work are publicly available in an institutional repository. [10], [11]

4.3 Data analysis

We compared the several idioms in terms of metrics and their features. As most of the metrics and features do not follow a normal distribution in each idiom, we used the median as our measure of central tendency and the interquartile range as our dispersion measure. As the assumptions for the one-way analysis of variance (ANOVA) hypothesis test were not verified, we applied the Kruskal-Wallis to realize if there were significant differences between idioms in each feature and metric. If so, we performed *post-hoc* tests, namely the Dunn pairwise test, with *p*-values adjusted by the Holm method, to identify the significant differences. When reporting our results, we use * to indicate results significant at an alpha=0.05 and ** to indicate results significant at an alpha=0.001.

5 RESULTS

Our results are organized by metric and described in the following sections.

5.1 Authority

Authority is "the degree of the reputation of an information object in a given community" [29], and it is computed as: ***Authority = 0,2*** $*$ ***Num. Unique Editors + 0,2*** $*$ ***Num. Edits + 0,1*** $*$ ***Connectivity + 0,3*** $*$ ***Num. Reverts + 0,2*** $*$ ***Num. External Links + 0,1*** $*$ ***Num. Registered User Edits + 0,2*** $*$ ***Num. Anonymous User Edits***. The number of unique editors corresponds to the number of different authors involved in the article's editions and is extracted from its history. Connectivity corresponds to the number of articles connected to a particular article through common editors and is obtained from each article's editors and the articles edited by them. This metric has the drawback of being based solely on articles in the database, requiring a large dataset to be accurate. Reverts correspond to the number of reversions made to editions of the article, and it is based on its editing history. External links correspond to the number of links in the article that points to content outside

[10] https://doi.org/10.25747/ep0v-en19
[11] https://doi.org/10.25747/wfzk-h937

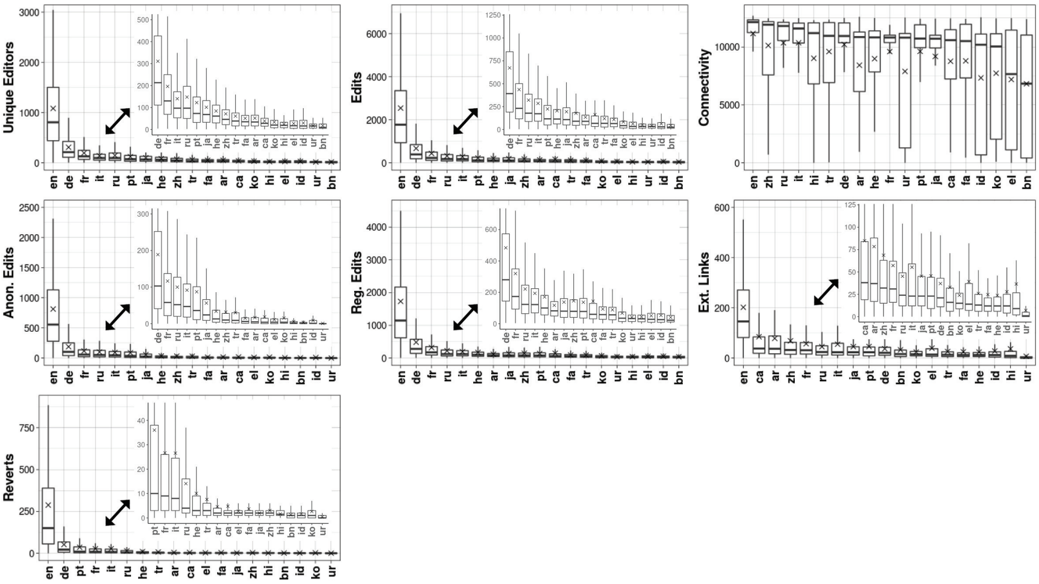

Figure 3: Distributions of authority features

Wikipedia. Registered or anonymous users can make edits, and it is obtained from the article's history.

The Table 3, represents median values for the metric, the interquartile range (IQR), and the idioms with significantly lower quality for each idiom. It is also represented the chi-squared and p-value for the metric. In this table, we observe that English stands out from the other idioms, achieving the top score for the median value. English is also the idiom with higher dispersion in values, achieving the higher IQR. German and Russian comes in second and third places. Bengali occupies the last place. Kruskal-Wallis test reveals significant differences among idioms. Russian is, however, not significantly higher than the fourth idiom - Italian. Although Bengali occupies the last place, three idioms - Urdu, Indonesian and Greek - are not significantly higher than Bengali.

Figure 3 represents boxplots for the features distributions. Outliers are not displayed for a more helpful plot visualization, and a zoom layer has been added on the less discernible areas. From this figure, we can conclude that English is the leader in all of the features. The number of reverts varies considerably between idioms, with a very considerable dominance of English. It should be noted that this difference may be due to the fact that there are significant differences in the number of article reversions, but also because that there are reversions that the authors did not identify as such, something that happens mainly in the less developed versions of Wikipedia. In these versions, there is less care with the structure of the articles in general and the comments in particular. English has the largest IQR for all features but connectivity.

5.2 Completeness

Completeness is defined as "the granularity or precision of an information object's model or content values according to some general-purpose IS-A ontology such as WordNet" [29], and it is computed as ***Completeness = 0,4 ∗ Num. Internal Broken Links + 0,4 ∗ Num. Internal Links + 0,2 ∗ Article Length***. Broken links correspond to those linking to pages that are no longer works. Internal links are those referring to internal pages of Wikipedia. The length corresponds to the number of characters of the article's text.

From Table 4, we can conclude that English emerges as the clear leader among the idioms for completeness metric, followed by German and French. Idioms such as Urdu, Korean, and Chinese stand out negatively, scoring more than ten times less than English. English is, once again, the idiom with the more significant variability for the metric. Kruskal-Wallis test reveals significant differences between idioms for this metric. French is, however, not significantly different from the following idiom - Russian. Korean, in penultimate place, is not significantly different from the last classified - Urdu.

As for the features, from Figure 4, we can observe that English scores the highest quality in all of them, but internal broken links, where it gets the least score and Persian reaches the top score. When we cross the number of internal broken links with the number of internal links, we find that, generally, the idioms with the highest number of internal links have the highest number of broken links. English is, however, an exception because despite being the idiom that has the highest number of internal links, it is the one that has the fewest number of broken links. For article length, English gets more than 150% more median value than the second idiom - German and almost 2,200% than Urdu, the last idiom.

Table 4: Idioms quality assessment for completeness metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	5132.4	5420.3	de	fr	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur	18
German	de	3172.8	4705.0	fr*	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur		17
French	fr	2619.3	4362.4	it*	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur				15
Russian	ru	2154.0	3262.3	ar*	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur					14
Italian	it	2042.0	3176.2	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Arabic	ar	1810.8	2799.0	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Portuguese	pt	1471.4	2244.2	he*	ca	ja	fa	hi	bn	id	tr	zh	ko	ur								11
Greek	el	1203.6	2319.7	ja	fa	hi*	bn	id	tr	zh	ko	ur										9
Hebrew	he	1080.2	1511.2	ja*	fa*	bn	id	tr	zh	ko	ur											8
Catalan	ca	936.0	1853.8	bn	id	tr	zh	ko	ur													6
Japanese	ja	903.2	1346.4	id*	tr*	zh	ko	ur														5
Persian	fa	868.4	1179.3	id*	tr*	zh	ko	ur														5
Hindi	hi	830.9	2350.6	id*	tr*	zh	ko	ur														5
Bengali	bn	676.7	995.9	ko	ur																	2
Indonesian	id	629.8	1312.4	ko	ur																	2
Turkish	tr	625.6	1083.5	ko	ur																	2
Chinese	zh	583.0	846.6	ko	ur																	2
Korean	ko	301.2	534.7																			0
Urdu	ur	271.3	422.6																			0
$\chi 2$		4410.6																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

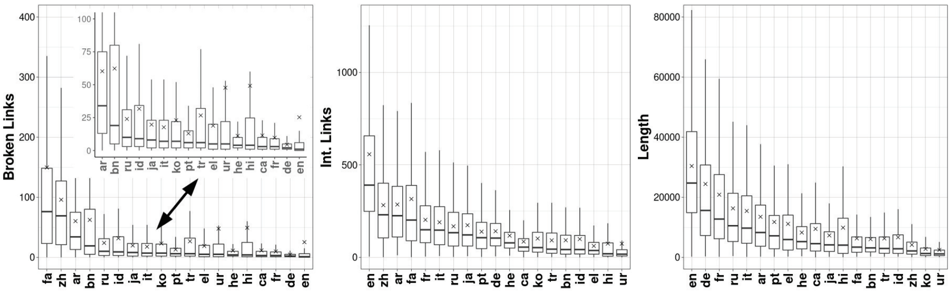

Figure 4: Distributions of completeness features

5.3 Complexity

The definition of complexity is linked to "the degree of cognitive complexity of an information object relative to a particular activity" [29], and it is computed as: ***Complexity = 0,5 ∗ "Flesch Reading Ease" - 0,5 ∗ "Kincaid grade level"***. Both Flesch Reading Ease [11] and Kincaid grade level [16] are tests that assess readability through the number of phrases, words, and syllables of the text. In the Flesch Reading Ease, higher scores mean that the text is easier to read, and lower scores mean that the text is complicated to understand; it is based on a ranking scale of 0-100. The result from Kincaid grade level reflects an American school grade level required to understand the text. They are inversely correlated, as a lower score on the reading ease test corresponds to a higher grade level. These instruments were developed for the English idiom, and so, there are issues when adapting to other idioms. As adaptations of the scales for other idioms are scarce, we decided to consider only the Flesch Reading Ease to calculate this measure. This formula for informativeness was present in an earlier version of Stvilia *et al.* [29]. For the Portuguese, there is an adaptation, consisting of adding 42 to the Flesch Reading Ease [23]; for the Turkish, the Atesman Formula [8] was used; for the English, French, German, and Italian idioms, Flesch Reading Ease was obtained using the Textstat [12] library for Python coding language. For the rest of the idioms, no satisfactory implementation was found.

In the Table 5 are only shown results for the idioms where the Flesch Reading Ease was computed. We can see that Italian and English stand out negatively, and Portuguese positively. Italian is significantly lower than all the idioms, and it gets only 17% of the Portuguese score. Regardless of being the top scorer, the Portuguese does not have the higher IQR. The outliers correspond mainly to articles in the different idioms, consisting of lists, resulting in a wrong result when applying the Flesch Reading Ease index, for example, the list of dead people by COVID-19. When analyzing this metric, it is always necessary to consider that Flesch

Reading Ease was initially developed for English and that the application in other idioms is an adaptation of this metric. Kruskal-Wallis test reveals significant differences among idioms for complexity. Portuguese, the top classified, is significantly higher than all idioms.

Table 5: Idioms quality assessment for complexity metric

		Median	IQR	Significantly lower idioms					# idioms
Portuguese	pt	76	18	tr	de*	fr	en	it	5
Turkish	tr	62	34	de	en	it			3
German	de	56	10	fr	en	it			3
French	fr	45	18	en*	it				2
English	en	34	14	it					1
Italian	it	13	26						0
$\chi 2$		11373							
***p*-value**		<2.2e-16							

* significance level 0.001<p≤0.05, significance level p≤0.001 for the rest of the values

The only feature considered in the computation of complexity was the Flesch Reading Ease. The respective distribution is shown in Figure 5.

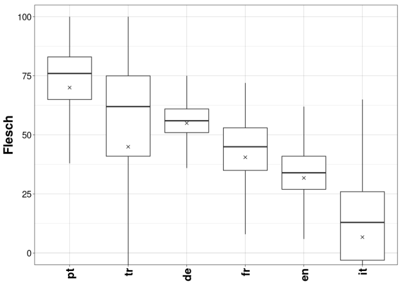

Figure 5: Distribution of complexity feature

5.4 Informativeness

Informativeness is defined as “the amount of information that an information object contains” [29], and it is computed as: ***Informativeness = 0,6 ∗ InfoNoise - 0,6 ∗ Diversity + 0,3 ∗ Num. Images***. InfoNoise [45] refers to the ratio between the size of the information after stemming and stopping and the article size before processing. Diversity corresponds to the ratio between the number of unique editors and the number of edits of an article. The number of images is obtained in the article.

In the Table 6, we can observe that English once again stands out positively, getting more than triple the score of the second idiom - Chinese, in second place. English has the more considerable variability for its median values. Urdu is once more in the last position. Kruskal-Wallis test reveals significant differences among idioms for informativeness. English is significantly higher than all the other idioms, and Urdu, the in the last position, is significantly lower than all the other idioms. Indonesian, Persian, Turkish, and Hindi are all solely significantly higher than Urdu.

Analyzing the features, according to Figure 6, English is just on top for images, getting 20 times the Hindi and Urdu median scores. For diversity, where a lower score means higher quality, Portuguese occupies first place. Japanese get the top score for infoNoise, where Chinese stands out negatively, getting the last place. We can observe similar dispersion for most of the idioms values in all the features.

5.5 Consistency

Consistency is defined as “the extent to which similar attributes or elements of an information object are consistently represented with the same structure, format and precision” [29], and it is computed as: ***Consistency = 0,6 ∗ Administrators Edit Share + 0,5 ∗ Age***. The administrators edit share corresponds to editions made by administrators, and it is obtained in the history. The item’s age corresponds to the time difference, in days, between the collection date and the article’s creation date.

From Table 7, we can observe that English, German and French again come out positively. English is, once again, the idiom with a higher IQR. Bengali gets the last place for this metric. The distribution is more homogeneous in this metric than in those previously analyzed - the top score is less than four times higher than the last idiom. Outliers refer to the various articles, in different idioms, recent and without any editing by administrators - administrator share is zero. Kruskal-Wallis test reveals significant differences among idioms for consistency. There are, however, no significant differences between English and German, the first and second idioms. There are also no significant differences for Bengali, the last, and Greek, the penultimate classified.

Analyzing the features distribution, from Figure 7, we can observe that English is expectedly at the top in the age feature, as it is the eldest version of Wikipedia. German, the second oldest version of Wikipedia, occupies the second place, but the third most old version - Catalan occupies the eleventh place. French occupies third place, and Bengali occupies once again the last place. In the administrator share, the dominance of the English is very significant, followed by Bengali and Arabic. The share of editions made by administrators is generally low, and for 11 idioms, the median value equals zero, although the mean values are higher than zero. There is a high dispersion in the age feature values for the majority of the idioms.

5.6 Volatility

Volatility is defined as “the amount of time the information remains valid” [29]. It corresponds to the length of hours the content remained valid until a later edition reverted it.

Analyzing Table 8 and Figure 8, we can observe that metric distribution does not follow the same pattern as the previous metrics. The top scores belong to Bengali, Catalan, Indonesian, Korean, and Urdu, whose median values equal zero. English comes in fifth place, and Japanese occupies the last place, as smaller scores translate into higher quality, as a lower median revert time means faster recovery from erroneous editions. However, the volatility score has a singularity - when there are no reversions in articles, the score for volatility equals zero, the same score as an article where the median time for its reversions is zero. This situation is also verified in the works of Domingues and Teixeira Lopes [7] - for the Portuguese idiom and Stvilia *et al.* [30] - for the random dataset. Cross-referencing this data with the number of reverts, we can

Table 6: Idioms quality assessment for informativeness metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	12.38	19.48	zh	ar	fr	it	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur	18
Chinese	zh	3.72	7.44	ar	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur				15
Arabic	ar	3.53	2.40	fr*	it	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur			16
French	fr	3.01	3.20	it*	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur				15
Italian	it	2.57	1.83	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur					14
Japanese	ja	2.08	5.14	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur								11
Hebrew	he	2.01	1.03	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur								11
Catalan	ca	1.83	1.78	de*	bn*	ko*	el	id	fa	tr	hi	ur										9
Russian	ru	1.68	1.34	ko	el	id	fa	tr	hi	ur												7
German	de	1.67	1.88	ko	el	id	fa	tr	hi	ur												7
Portuguese	pt	1.65	1.84	ko	el	id	fa	tr	hi	ur												7
Bengali	bn	1.45	2.87	ko*	id	fa	tr	hi	ur													6
Korean	ko	1.38	0.90	fa*	tr	hi*	ur															4
Greek	el	1.25	1.37	id*	fa	tr	hi	ur														5
Indonesian	id	1.12	1.14	ur																		1
Persian	fa	1.10	1.33	ur																		1
Turkish	tr	1.07	0.81	ur*																		1
Hindi	hi	0.82	1.61	ur*																		1
Urdu	ur	0.72	1.03																			0
$\chi 2$		4446.6																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

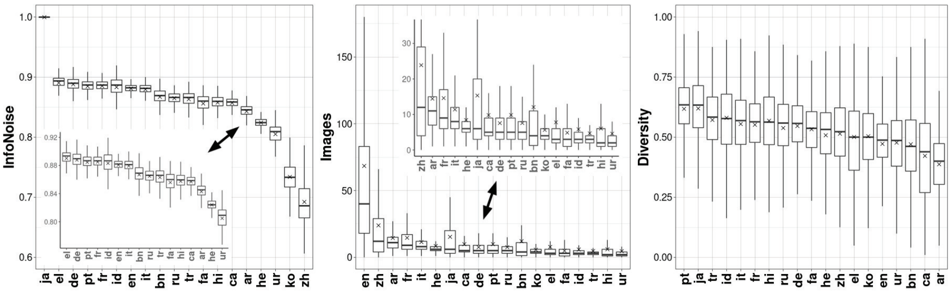

Figure 6: Distributions of informativeness features

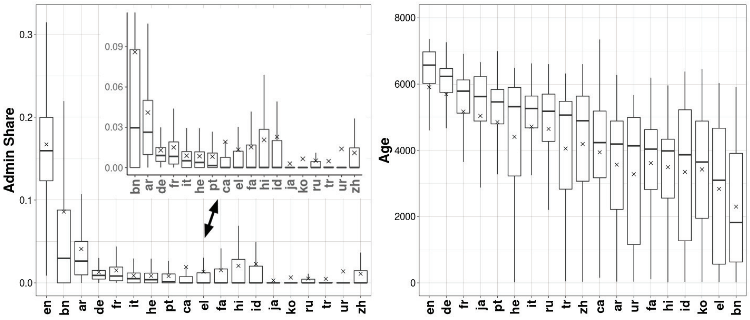

Figure 7: Distributions of consistency features

observe that, for English, we have 0.3% of articles without any revert, while this value rises to 36%, 23%, 36%, 28%, and 74%, for Bengali, Catalan, Indonesian, Korean and Urdu, respectively. Also, we can see that Urdu scores zero for the IQR, as the few values it gets are classified as outliers. According to this analysis, we can consider that these five idioms do not achieve, in fact, more quality, for volatility, than English or French, and so, we consider English as the top score idiom. The Kruskal-Wallis test reveals significant differences among idioms for volatility and median revert time, which is the only feature for this metric. There are, however, no significant differences for the two last classified idioms - Arabic and Japanese.

Table 7: Idioms quality assessment for consistency metric

		Median	IQR	Significantly lower idioms																	# idioms
English	en	3286.4	476.6	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
German	de	3116.5	362.0	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
French	fr	2893.5	501.2	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Japanese	ja	2811.5	669.5	pt*	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Portuguese	pt	2731.0	518.5	he*	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn					13
Hebrew	he	2659.0	1336.5	tr	zh*	ca	ar	ur	fa	hi	id	ko	el	bn							11
Italian	it	2632.0	480.3	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn							11
Russian	ru	2590.0	712.7	tr	zh*	ca	ar	ur	fa	hi	id	ko	el	bn							11
Turkish	tr	2532.5	1324.9	ar	ur	fa	hi	id	ko	el	bn										8
Chinese	zh	2447.0	1284.0	ca	ar	ur	fa	hi	id	ko	el	bn									9
Catalan	ca	2117.0	1007.0	ar*	ur*	fa	hi	ko*	el	bn											7
Arabic	ar	2095.0	1333.5	el	bn																2
Urdu	ur	2067.8	1917.5	bn																	1
Persian	fa	2018.8	904.9	el	bn																2
Hindi	hi	1991.8	889.0	bn																	1
Indonesian	id	1933.0	1977.9	el	bn																2
Korean	ko	1824.3	1467.6	el	bn																2
Greek	el	1549.0	2052.0																		0
Bengali	bn	911.3	1639.8																		0
$\chi 2$		4534.3																			
***p*-value**		<2.2e-16																			

* significance level $0.001<p\leq0.05$, significance level $p\leq0.001$ for the rest of the values

Table 8: Idioms quality assessment for volatility metric

		Volatility	IQR	Significantly lower idioms																# idioms
Bengali	bn	0	63.0	ur	en*	pt*	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Catalan	ca	0	26.0	id*	ko	ur	en*	pt	de*	it	he	hi	ru	fa	tr	el	zh	ar	ja	16
Indonesian	id	0	50.3	ur	en	fr	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Korean	ko	0	32.0	ur	en	fr	it*	he	hi	ru	fa	tr	el	zh	ar	ja				13
Urdu	ur	0	0.0	fr	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
English	en	2	3.0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
French	fr	4	6.0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
Portuguese	pt	5	18.0	he	hi	ru	fa	tr	el	zh	ar	ja								9
German	de	5.5	12.0	it	he	hi	ru	fa	tr	el	zh	ar	ja							10
Italian	it	6.5	30.0	ru	fa	tr	el	zh	ar	ja										7
Hebrew	he	13	31.0	ru	fa	tr	el	zh	ar	ja										7
Hindi	hi	19	31.8	ru	fa	tr	el	zh	ar	ja										7
Russian	ru	26	77.5	ar*	ja															2
Persian	fa	32.5	87.5	ar	ja															2
Turkish	tr	33	99.0	ar*	ja															2
Greek	el	36	63.0	ar*	ja															2
Chinese	zh	40	58.0	ja																1
Arabic	ar	56	83.0																	0
Japanese	ja	58	90.5																	0
$\chi 2$		2775.9																		
***p*-value**		<2.2e-16																		

* significance level $0.001<p\leq0.05$, significance level $p\leq0.001$ for the rest of the values

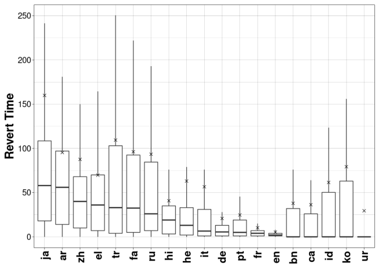

Figure 8: Distribution of volatility feature

5.7 Currency

Currency is "the age of an information object" [29]. It corresponds to the time between the collection date and the date of the article's last update, in days.

In currency, a lower score means higher quality, as lower currency means more up-to-date articles. Analyzing Figure 9 and Table 9, English stands out positively, followed by German and Japanese. English also gets the lowest IQR. Hindi and Urdu stand out negatively. Urdu score is notably lower than the top score (1,785%). The outliers, removed, correspond to various articles in different idioms, where no edition has been made for a long time. Kruskal-Wallis test reveals significant differences among idioms for currency. While English is significantly higher than the rest of the idioms, German is not significantly higher than Japanese. Urdu is significantly lower

Table 9: Idioms quality assessment for currency metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	20	17	de	ja	ca	ru	fr	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur	18
German	de	42	72	ca	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Japanese	ja	42	39	ca*	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Catalan	ca	50	99	ar	zh	bn	ko	el	id	pt	hi	ur										9
Russian	ru	50	90	fa	tr	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
French	fr	54	82	fa	tr*	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
Persian	fa	57	114	bn*	ko	el	id	pt	hi	ur												7
Turkish	tr	63	66	bn*	ko	el	id	pt	hi	ur												7
Italian	it	65	101	bn	ko	el	id	pt	hi	ur												7
Arabic	ar	71	144	ko	el	id	pt	hi	ur													6
Chinese	zh	75	124	he	ko	el	id	pt	hi	ur												7
Hebrew	he	79	105	bn*	ko	el	id	pt	hi	ur												7
Bengali	bn	87	205	el	id	pt	hi	ur														5
Korean	ko	126	169	el*	id	pt	hi	ur														5
Greek	el	130	256	hi*	ur																	2
Indonesian	id	140	349	hi*	ur																	2
Portuguese	pt	157	283	hi*	ur																	2
Hindi	hi	210	218																			0
Urdu	ur	357	370																			0
χ^2		2802.9																				
***p*-value**		<2.2e-16																				

* significance level $0.001<p\leq 0.05$, significance level $p\leq 0.001$ for the rest of the values

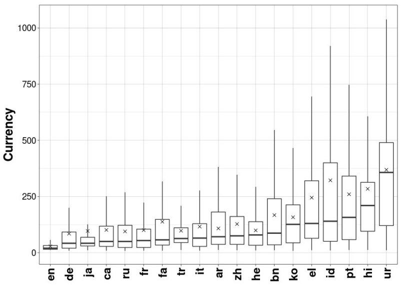

Figure 9: Distribution of currency feature

than the rest of the idioms. The only feature of currency is the currency itself.

6 DISCUSSION

When we analyze the scores of the idioms in the different metrics and their metrics, we see that some idioms generally occupy the top places and others more often occupy the bottom places. In order to be able to define a ranking of idioms, we computed the mean of the number of significantly lower languages for all metrics. These values are present in Table 10, where the idioms are sorted in descending order of the mean since a higher mean means that the language scored significantly higher than the rest. It is also represented the percentile rank for each idiom. Top scores are highlighted in bold. In currency, it is necessary to take into account the constraints associated with Bengali, Catalan, Indonesian, Korean, and Urdu, and described in Subsection 5.6.

As expected, English is at the top, with a mean of significantly lower idioms of 14.6 and a mean percentile of 85%. German and French scored a mean of 12.1, and mean percentiles of 79% and 76%, respectively. The last place belongs to Urdu, with a mean of significantly lower idioms of 2.5 and a mean percentile of 22%. Greek scored not very far from Urdu, with means of 3.0 and 25%, for significantly lower idioms and percentile, respectively. Given the already discussed idiosyncrasies of volatility and considering English as the top idiom, the only metric that does not rank first is complexity, but this metric may be subject to constraints previously described. Regarding the idioms selected for their historical tradition, we can observe that Greek, Persian, Turkish, and Korean are on the bottom half of the table. On the other hand, Italian was the idiom that got the best mean of significantly lower idioms - 9.4, with a mean percentile of 58%.

These results point in the direction of other works, such as Teixeira Lopes and Ribeiro [21], suggesting that English should be provided to users with higher levels of English proficiency, opening doors for higher-quality content.

To understand if quality has any connection with quantity, we computed the correlation between the quality across the different metrics and the number of speakers and the total number of articles in each Wikipedia version for the selected idioms. Results are shown in Table 11, which presents the Spearman correlation value and *p*-values for the number of speakers and Wikipedia articles, for all idioms, and the metrics quality computed values.

Analyzing the results, we can conclude that there is a significant correlation between quality and the number of total articles in each Wikipedia version, mainly for completeness (0.94), authority (0.9), and informativeness (0.9), with significant *p*-values($\leq$0.001). Informativeness is the metric with more correlation with the number of speakers, with a significant *p*-value ($\leq$0.001), followed by authority (*p*-value$\leq$0.05). There is also a strong correlation (0.63) for the number of speakers and the number of articles in each Wikipedia version, with a significant computed *p*-value($\leq$0.05).

To analyze the effects of the different number of articles in languages, we have conducted a similar analysis, including only the 164 articles having a version in every studied idiom. From this analysis, we could conclude that the significant results are very

Table 10: Idioms ranking summary

	Authority		Completeness		Complexity		Informativeness		Consistency		Volatility		Currency		Mean	
	# SLI	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**SLI**	**%**
English	**18**	**100%**	**18**	**100%**	1	33%	**18**	**100%**	**17**	**95%**	12	68%	**18**	**100%**	**14.6**	**85%**
German	17	95%	17	95%	3	67%	7	47%	**17**	**95%**	10	63%	14	89%	12.1	79%
French	13	79%	15	89%	2	50%	15	84%	15	84%	12	68%	13	79%	12.1	76%
Russian	14	84%	14	84%			7	47%	11	63%	2	21%	13	79%	10.2	63%
Italian*	14	84%	13	74%	0	17%	14	79%	11	63%	7	42%	7	47%	9.4	58%
Catalan*	5	37%	6	53%			9	63%	7	47%	**16**	**100%**	9	74%	8.7	62%
Hebrew*	6	53%	8	58%			11	68%	11	63%	7	42%	7	47%	8.3	55%
Japanese	5	37%	5	37%			11	68%	15	84%	0	5%	14	89%	8.3	54%
Portuguese	9	68%	11	68%	**5**	**100%**	7	47%	13	79%	9	58%	2	16%	8.0	62%
Chinese	12	74%	2	16%			15	84%	9	58%	1	16%	7	47%	7.7	49%
Arabic	5	37%	13	74%			16	95%	2	26%	0	5%	6	42%	7.0	46%
Bengali	0	5%	2	16%			6	42%	0	5%	13	79%	5	32%	4.3	30%
Korean*	1	21%	0	5%			4	32%	2	26%	13	79%	5	32%	4.2	32%
Turkish*	6	53%	2	16%	3	67%	1	11%	8	53%	2	21%	7	47%	4.1	38%
Persian*	4	32%	5	37%			1	11%	2	26%	2	21%	7	47%	3.5	29%
Hindi	6	53%	5	37%			1	11%	1	16%	7	42%	0	5%	3.3	27%
Indonesian	0	5%	2	16%			1	11%	2	26%	13	79%	2	16%	3.3	25%
Greek*	0	5%	9	63%			5	37%	0	5%	2	21%	2	16%	3.0	25%
Urdu	1	21%	0	5%			0	5%	1	16%	13	79%	0	5%	2.5	22%

SLI: Significantly lower idioms, * idioms selected for their historical tradition

Table 11: Correlation between metrics and number of speakers and articles

	Speakers		Wikipedia	
	correlation	*p*-value	correlation	*p*-value
Authority	0.67	0.0015*	0.90	0.0000**
Completeness	0.46	0.0459	0.94	0.0000**
Complexity	0.06	0.8091	0.37	0.1162
Informativeness	0.77	0.0001**	0.90	0.0000**
Consistency	0.29	0.2247	0.69	0.0011*
Volatility	0.01	0.9597	-0.15	0.5361
Currency	-0.08	0.7544	-0.46	0.0451

* significance level $p \leq$7e-3, ** significance level $p \leq$1e-4. (Bonferroni corrected from p=0.05 and p=0.001, 7 tests)

similar to those described above. The top and bottom-ranked languages remain the same, and English is still leading in the same metrics.

7 CONCLUSION

We performed a comparison of health-related articles on Wikipedia across 19 different idioms: English, Arabic, French, Portuguese, German, Persian, Italian, Chinese, Russian, Japanese, Hebrew, Korean, Catalan, Indonesian, Turkish, Greek, Hindi, Bengali, and Urdu. To assess the information quality of the articles, we used a set of seven predefined metrics: authority, completeness, complexity, informativeness, consistency, currency, and volatility.

We faced some challenges due to the heterogeneity of the idioms analyzed and some variation between the different versions of Wikipedia, such as its structure. Given this heterogeneity, we could not use some metrics in some idioms. It is the case of the readability tests for complexity metric.

After analyzing the results, we concluded that there is a significant difference among idioms for quality. English is the idiom that shows the most difference to all other idioms, with the best values for quality metrics, followed by German, French, and Russian. Urdu, Greek, Indonesian, and Hindi are the idioms with worse values of quality in general. We also concluded a correlation between the number of speakers and the number of articles in each Wikipedia version. This correlation is more significant for the number of Wikipedia's articles and for some metrics, such as completeness and authority. With this characterization of the differences between idioms, we hope to raise awareness of this heterogeneity and make the first step towards more equal versions of Wikipedia. To overcome this heterogeneity between the different idiom versions of Wikipedia, the Wikimedia Foundation has an ongoing project - Abstract Wikipedia [36]. This project aims to create a language-independent version of Wikipedia by modeling data from Wikidata. This will allow people to create language-independent content that will be later translated through code. This project also contains Wikifunctions, which allows anyone to create and maintain code and includes code that converts the language-independent article from Abstract Wikipedia to Wikipedia's native language.

REFERENCES

[1] Rita Baeten, Slavina Spasova, Bart Vanhercke, and Stéphanie Coster. 2018. *Inequalities in access to healthcare.* Number November.

[2] Anamika Chhabra, Shubham Srivastava, S Iyengar, and Poonam Saini. 2021. Structural Analysis of Wikigraph to Investigate Quality Grades of Wikipedia Articles. https://doi.org/10.1145/3442442.3452345

[3] Riccardo Conti, Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, and Marinella Petrocchi. 2014. Maturity assessment of Wikipedia medical articles. *Proceedings - IEEE Symposium on Computer-Based Medical Systems* (2014), 281–286. https://doi.org/10.1109/CBMS.2014.69

[4] Luís Couto and Carla Lopes. 2021. Assessing the quality of health-related Wikipedia articles with generic and specific metrics. 640–647. https://doi.org/10.1145/3442442.3452355

[5] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring article quality in Wikipedia using the collaboration network. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015* (2015), 464–471. https://doi.org/10.1145/2808797.2808895

[6] Lara Devgan, Neil Powe, Brittony Blakey, and Martin Makary. 2007. Wiki-Surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons* 205 (09 2007), S76–S77. https://doi.org/10.1016/j.jamcollsurg.2007.06.190

[7] Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. *The Web Conference*

2019 - Companion of the World Wide Web Conference, WWW 2019 (2019), 1203–1207. https://doi.org/10.1145/3308560.3316758

[8] Ateşman E. 1997. Measuring readability in Turkish. *Tömer Language Journal* 58 (1997), 171–174.

[9] Ethnologue. 2021. How many languages are there in the world? Retrieved jun, 2021 from https://www.ethnologue.com/guides/how-many-languages

[10] Elena Filatova. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. (01 2009). https://doi.org/10.3115/1572433.1572438

[11] R FLESCH. 1948. A new readability yardstick. *The Journal of applied psychology* 32, 3 (June 1948), 221—233. https://doi.org/10.1037/h0057532

[12] Python Software Foundation. 2021. textstat 0.7.0. Retrieved jun, 2021 from https://pypi.org/project/textstat/

[13] Google. 2021. Year in Search 2020. Retrieved jun, 2021 from https://trends.google.com/trends/yis/2020/GLOBAL/

[14] Scott Hale. 2013. Multilinguals and Wikipedia Editing. *WebSci 2014 - Proceedings of the 2014 ACM Web Science Conference* (12 2013). https://doi.org/10.1145/2615569.2615684

[15] Imran Khan, Shahid Hussain, Hina Gul, Muhammad Shahid, and Muhammad Jamal. 2019. *An Empirical Study to Predict the Quality of Wikipedia Articles*. 485–492. https://doi.org/10.1007/978-3-030-16187-3_47

[16] J. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

[17] Jona Kräenbring, Tika Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. 2014. Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology. *PloS one* 9 (09 2014), e106930. https://doi.org/10.1371/journal.pone.0106930

[18] Michaël R. Laurent and Tim J. Vickers. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16, 4 (2009), 471–479. https://doi.org/10.1197/jamia.M3059

[19] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2020. Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. *Information* 11 (05 2020), 263. https://doi.org/10.3390/info11050263

[20] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, and Maarten de Rijke. 2015. Automatically assessing wikipedia article quality by exploiting article–editor networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9022 (2015), 574–580. https://doi.org/10.1007/978-3-319-16354-3_64

[21] Carla Teixeira Lopes and Cristina Ribeiro. 2013. Measuring the value of health query translation: An analysis by user language proficiency. *JASIS* 64, 5 (2013), 951–963. https://doi.org/10.1002/asi.22812

[22] Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An Edit-centric Approach for Wikipedia Article Quality Assessment. (2019), 381–386. https://doi.org/10.18653/v1/d19-5550 arXiv:1909.08880

[23] Teresa Martins, Claudete Ghiraldelo, M. Gragas, V. Nunes, and O.N. Oliveira Jr. 1996. Readability Formulas Applied to Textboooks in Brazilian Por- tuguese. (06 1996).

[24] Sorin Matei and Brian Britt. 2017. *Structural Differentiation in Social Media: Adhocracy, Entropy, and the "1% Effect"*. https://doi.org/10.1007/978-3-319-64425-7

[25] Omeed Modiri, Daipayan Guha, Naif M. Alotaibi, George M. Ibrahim, Nir Lipsman, and Aria Fallah. 2018. Readability and quality of wikipedia pages on neurosurgical topics. *Clinical Neurology and Neurosurgery* 166, January (2018), 66–70. https://doi.org/10.1016/j.clineuro.2018.01.021

[26] Daniel Pimienta, D. Prado, and Á Blanco. 2009. Twelve years of measuring linguistic diversity in the Internet: balance and perspectives. *Paris: UNESCO. Retrieved March* 7, September (2009), 2010.

[27] Natalia Pletneva, Sarah Cruchet, Maria Ana Simonet, Maki Kajiwara, and Célia Boyer. 2011. Results of the 10th HON survey on health and medical internet use. *Studies in Health Technology and Informatics* 169, 2008 (2011), 73–77. https://doi.org/10.3233/978-1-60750-806-9-73

[28] Malolan Rajagopalan, Vineet Khanna, Yaacov Leiter, Meghan Stott, Timothy Showalter, Adam Dicker, and Yaacov Lawrence. 2011. Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of oncology practice / American Society of Clinical Oncology* 7 (09 2011), 319–23. https://doi.org/10.1200/JOP.2010.000209

[29] Besiki Stvilia, Michael Twidale, Linda Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-Based Encyclopedia. *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005* (01 2005).

[30] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. 2005. Information quality in a community-based encyclopedia. *Knowledge Management: Nurturing Culture, Innovation, and Technology-Proceedings of the 2005 International Conference on Knowledge Management* (2005), 101–113.

[31] Athikhun Suwannakhan, Daniel Casanova-Martínez, Laphatrada Yurasakpong, Punchalee Montriwat, Krai Meemon, and Taweetham Limpanuparb. 2019. The Quality and Readability of English Wikipedia Anatomy Articles. *Anatomical Sciences Education* 13 (2019), 1–13. https://doi.org/10.1002/ase.1910

[32] Sharon Tan and Nadee Goonawardene. 2017. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *Journal of Medical Internet Research* 19 (01 2017), e9. https://doi.org/10.2196/jmir.5729

[33] Garry R. Thomas, Lawson Eng, Jacob F. de Wolff, and Samir C. Grover. 2013. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in Dialysis* 26, 2 (2013), 159–163. https://doi.org/10.1111/sdi.12059

[34] Ziko VanDijk. 2009. Wikipedia and lesser-resourced languages. *Language Problems and Language Planning* 33, 3 (2009), 234–250. https://doi.org/10.1075/lplp.33.3.03van

[35] Peter Volsky, Cristina Baldassari, Sirisha Mushti, and Craig Derkay. 2012. Quality of Internet information in pediatric otolaryngology: A comparison of three most referenced websites. *International journal of pediatric otorhinolaryngology* 76 (07 2012), 1312–6. https://doi.org/10.1016/j.ijporl.2012.05.026

[36] Wikimedia. 2021. Abstract Wikipedia. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Abstract_Wikipedia

[37] Wikimedia. 2021. Language proposal policy. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Language_proposal_policy

[38] Wikimedia. 2021. Requests for new languages. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Requests_for_new_languages

[39] Wikipedia. 2021. List of Wikipedias. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/List_of_Wikipedias

[40] Wikipedia. 2021. Wikipedia:Blocking policy. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

[41] Wikipedia. 2021. Wikipedia:Content assessment. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

[42] Wikipedia. 2021. Wikipedia:Five pillars. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

[43] Wikipedia. 2021. Wikipedia:WikiProject Medicine/Popular pages. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages

[44] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, and Hua Zheng. 2010. Mining the factors affecting the quality of Wikipedia articles. *Proceedings - 2010 International Conference of Information Science and Management Engineering, ISME 2010* 1, 1 (2010), 343–346. https://doi.org/10.1109/ISME.2010.114

[45] Xiaolan Zhu and Susan Gauch. 2000. Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 288–295. https://doi.org/10.1145/345508.345602