informatics

MDPI

Article

Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles

Włodzimierz Lewoniewski *,†,‡, **Krzysztof Węcel** ‡ and **Witold Abramowicz** ‡

Department of Information Systems, Poznań University of Economics and Business, 61-875 Poznań, Poland; krzysztof.wecel@ue.poznan.pl (K.W.); witold.abramowicz@ue.poznan.pl (W.A.)

* Correspondence: wlodzimierz.lewoniewski@ue.poznan.pl; Tel.: +48-(61)-639-27-93

† Current address: al. Niepodległości 10, 61-875 Poznań, Poland

‡ These authors contributed equally to this work.

Academic Editors: Mouzhi Ge and Vlastislav Dohnal

Received: 21 September 2017; Accepted: 2 December 2017; Published: 8 December 2017

Abstract: Despite the fact that Wikipedia is often criticized for its poor quality, it continues to be one of the most popular knowledge bases in the world. Articles in this free encyclopedia on various topics can be created and edited in about 300 different language versions independently. Our research has showed that in language sensitive topics, the quality of information can be relatively better in the relevant language versions. However, in most cases, it is difficult for the Wikipedia readers to determine the language affiliation of the described subject. Additionally, each language edition of Wikipedia can have own rules in the manual assessing of the content's quality. There are also differences in grading schemes between language versions: some use a 6–8 grade system to assess articles, and some are limited to 2–3. This makes automatic quality comparison of articles between various languages a challenging task, particularly if we take into account a large number of unassessed articles; some of the Wikipedia language editions have over 99% of articles without a quality grade. The paper presents the results of a relative quality and popularity assessment of over 28 million articles in 44 selected language versions. Comparative analysis of the quality and the popularity of articles in popular topics was also conducted. Additionally, the correlation between quality and popularity of Wikipedia articles of selected topics in various languages was investigated. The proposed method allows us to find articles with information of better quality that can be used to automatically enrich other language editions of Wikipedia.

Keywords: Wikipedia; information quality; WikiRank; DBpedia

1. Introduction

Sustaining accurate, complete, reliable, and up-to-date information on the Web is very important, particularly during the development of collaborative platforms and the growth of their popularity. These platforms allow Internet users to create content without special technical skills. Despite the fact that even anonymous users can participate in content addition, information in these knowledge bases can be not only abundant but also trustworthy [1].

Wikipedia is one of the best examples of such collaborative platforms. This encyclopedia became a popular source of information on different topics. Nowadays, it is the fifth most visited page in the world (https://www.alexa.com/siteinfo/wikipedia.org). The pages of this online knowledge base often appear among the first in search results using Google, Bing, Yandex, and other search engines. There are about 300 language editions on Wikipedia with over 46 million articles, which cover all subjects of human activity (https://meta.wikimedia.org/wiki/List_of_Wikipedias). The English edition is the largest and consists of over 5.4 million articles.

Despite the popularity and large volume of information in free access, Wikipedia is often criticized for unreliable content (more information about criticism of Wikipedia can be found on the page https://en.wikipedia.org/wiki/Criticism_of_Wikipedia). This is due primarily to the fact that everyone can participate in the creation and editing of the articles without proving competence or education. Changes made by users (even anonymous users) are immediately available for a wide range of Wikipedia readers. There is no professional editorial control. Articles about the same subject can be edited independently in each language version. Therefore, we can observe a difference in quality between languages depending on the described topic.

In order to help readers quickly determine the quality of the content, the Wikipedia community has defined a grading system for assessing the quality of the articles. Each language version of Wikipedia can have its own rules and standards for writing. In many language versions, there are special awards for articles of the highest quality. In English Wikipedia, these articles are labeled as "featured articles" (FAs)—they must be well-written with appropriate structure, comprehensive, well-researched with reliable sources, and present views fairly and without bias (https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria). Another distinction—"good article" (GA)—can be awarded to an article that has not met the criteria for a FA but was close enough. These awards used in English Wikipedia often have equivalents in other language editions of Wikipedia. For example, for FA and GA awards in German Wikipedia are "exzellente artikel" and "lesenswerte artikel", respectively. However, the share of the best articles in each Wikipedia language is relatively small—on average, around 0.3% in each language.

In some language editions of Wikipedia, there are also other quality grades, which can reflect the maturity of an article. In English Wikipedia, in addition to the highest FA and GA grades, there are also A-class, B-class, C-class, start, and stub. In Russian Wikipedia, there are additionally "solid article", "I level", "II level", "III level", and "IV level" grades. Polish Wikipedia has three additional grades: "four", "start", and "stub". Equivalent classes between language versions can have differences in assessing standards. For example, in some language versions, for high grades, there is a limitation on the article's length. Therefore, each language version can have its own quality model, even if those languages have the same number of grades.

The observation is then that there are differences between Wikipedia languages in grading schemes, and not all language versions have a developed system of quality grades for articles. For example, one of the largest versions of Wikipedia is the German edition, and it has only two highest grades—equivalent to FA and GA. Differences in quality grades do not allow us to directly compare the quality of the articles between the various language versions. An additional challenge is a large number of articles without grades. For example, in German and Polish Wikipedia, over 99% of articles are unassessed (i.e., over 2 million and 1.2 million articles, respectively).

The goal of this paper is to research the relation between the quality of Wikipedia articles and their popularity. Our hypothesis is that relative popularity is positively correlated with the relative quality of an article. We introduce a method of quality assessment of Wikipedia articles as a synthetic measure, on a scale between 0 and 100. This approach is used to evaluate more than 28 million articles in 44 language versions of Wikipedia. In addition, a comparison of quality between the articles in different languages on selected topics is conducted. The paper also presents results of the estimation of relative popularity of these articles. This makes it possible to study the association between quality and popularity in each language–topic pair.

The paper is structured as follows. Section 2 describes related work concerning both the quality and popularity of Wikipedia articles. Section 3 introduces a synthetic measure used by us to assess the quality of articles, and we present various statistics. Section 4 explains how popularity is measured. In Section 5, we study the association between quality and popularity. Section 6 presents the results of the quality and popularity assessment of Wikipedia articles in 44 languages on different topics. Section 7 concludes the paper.

2. Related Work

2.1. Quality Assessment

Automatic quality assessment of Wikipedia articles is a relatively developed topic in scientific works. Using different methods, it is possible to estimate the quality of articles on the basis of content, edit history, the article's discussion page, the article's links, users' reputations, and other sources. Related studies have proposed different sets of metrics, which can be divided into two groups: content-based and user-based methods.

First works concerning content-based methods have concluded that longer articles in Wikipedia often had a higher quality [2]. Other papers have showed that high-quality articles tend to have more images, sections, and references [3–5]. Some scientific works have analyzed language features, which can characterize the writing style of articles. High-quality articles cover more concepts, objects and facts than lower-quality articles [6,7]. According to these studies, the number of facts in a document can indicate its informativeness. The writing style of Wikipedia articles can be also estimated by analyzing character trigram metrics [8]. Basic lexical metrics based on word usages in Wikipedia articles are used in another study as the factors that can reflect the articles' quality—high-quality articles often used more nouns and verbs and less adjectives [9]. Finally, a quality evaluation of Wikipedia articles can also be based on special quality flaw templates [10].

The second group of studies—user-based—is related to editors' behavior. These aim to analyze how the user skills, experience, and coordination of their activities affect the quality of Wikipedia articles. These methods use different metrics related to the user's reputation and changes that they have made in pages [11,12]. If an article has a relatively large number of editors and edits, then often this article will be of high quality [13]. Cooperation among authors and edited articles can be visualized as a network. Using graph theory, it is possible to determine structural features associated with an article's quality [14]. Artificial intelligence methods can be applied to score the article quality by discovering damaging edits [15]. However, described user-based approaches often require complex calculations, and they cannot indicate what needs to be corrected in the article to improve its quality.

Among other suggested methods, it is also necessary to note the Objective Revision Evaluation Service (ORES) [15], which can classify an article to one of the quality grades and also can automate the vandalism detection. In this case, the article quality can be evaluated on an interval scale (between 0 and 1). However, currently, automatic quality assessment of an article by the ORES is only possible in three Wikipedia language versions (https://www.mediawiki.org/wiki/ORES): English, French, and Russian. This may be due to the fact that the approach works well on large language editions of Wikipedia (with over 1 million articles), for which it is possible to obtain a sufficient amount of data for a training set. Another limitation is the specifics of grading schemes—a relatively well-developed grading scheme is necessary, with six or more quality grades. Our previous works have showed that each Wikipedia language version can have its own grading scheme [4,5], and some of these versions use only 2–3 grades. For example, German Wikipedia with over 2 million articles has only two highest-quality grades for articles' assessment. A less developed quality grading scheme is one of the main reasons for the large number of unevaluated articles—more than 99% of articles in German, Polish, and other Wikipedia language versions do not have any quality grade.

Although existing works propose various sets of metrics for assessing the quality of Wikipedia articles, there is no universal feature set for this task [16]. An additional challenge is to consider different language versions, which can have different quality models [4,5]. Extraction rules of some metrics (e.g., lexical) can also be language-sensitive [6,7,9]. There are also a few works that aim to combine metrics from articles' content and edition history [16,17].

Concluding, by using different metrics and models, it is possible to estimate the quality of an article. The majority of the approaches are focused only on one (usually the largest—English) or several language versions. Additionally, these methods essentially allow for the evaluation of articles

and the comparing of their quality only within one selected language version of Wikipedia. This is due to the differences that can arise in the quality models between various Wikipedia languages [4,5].

In this paper, for the particular task of comparing the quality using synthetic measures, we decided to take into the account only important content-based metrics. Most of the existing studies evaluate the quality of Wikipedia articles as a binary classification problem, which is limited when comparing articles with similar quality classes. Some of the researchers have aimed to build models by taking into account all (or major) quality grades in developed language versions (such as English), but in this case, the precision decreases significantly. Additionally, previous studies have examined the quality of an article within one selected Wikipedia language instead of comparing different language versions of this article.

2.2. Popularity Measures

The second measure that we analyze in this paper is the popularity of articles. Earlier studies have showed that for some developed language versions of Wikipedia (such as English, German, and Spanish), the popularity of the articles was correlated with its number of edits [18]. Our prior work has showed that popularity can play an important role in the estimation of quality in specific language versions of Wikipedia [5]. Other studies have showed that measuring a topic's popularity in English Wikipedia can help in determining its number of articles of good quality—if the topic is popular, then it has a larger number of high-quality articles [19]. Warncke-Wang et al. showed misalignment between the popularity and quality of the articles in Wikipedia; however, the study was limited to four language versions of Wikipedia [20]. Additionally, none of the studies provided a comparative analysis of the popularity of the same article between language versions and its impact on the quality. Popularity can also show to a certain degree the importance of the article for groups of Wikipedia users that read it in a selected language version. This can also provide motivation to assign a higher-quality grade for an article in a given language version compared to other languages—a greater number of users can check the completeness, timeliness, and reliability of facts described in the article. Therefore, our hypothesis is that popularity can affect the quality dimension of an article.

This study is the continuation of work on building a synthetic measure for the quality assessment of Wikipedia articles in different languages [21]. Preliminary results have shown the high efficiency of this method in assessing articles on language-sensitive topics. Compared with our previous work [21], we decided to increase the number of analyzed languages (from 7 to 44), expand the rules for quality assessment, and analyze the popularity of the articles.

3. Quality Measure

Many of the existing studies solve the problem of automatic quality assessment of articles as a classification task: articles can be marked as complete or incomplete [3–7,9]. This is a large limitation for comparing articles in different languages, as it is not possible to show to what degree one article is better than the other if both are tagged with the same class (e.g., incomplete). Additionally, it is necessary to take into account different standards in the quality assessment met in various language editions of Wikipedia, defined by each community.

In order to build a synthetic measure, we chose five important content-based metrics:

- *len*—article length (in bytes);
- *ref*—number of references;
- *img*—number of images;
- *hdr*—number of first- and second-level headers;
- *ral*—the ratio of the number of references to the article length.

These metrics previously have showed high prediction power in quality assessment of English Wikipedia [3], as well as for other language editions of Wikipedia [4,5]. According to our findings, the above metrics are positively correlated with the quality grades [4,5,21] (see Figure 1).

Considering over 4 million articles with assigned quality classes in English Wikipedia, we have calculated the values of proposed metrics by quality classes. We can observe that the values of the metrics increase with an increase in the quality (stub—the lowest; FA—the highest). Table 1 presents medians of each metric of all articles in a particular quality class. As a side note, we do not take into account the A-class, because this class is usually assigned to articles that already have a FA or GA grade. We also excluded 111,412 articles that had two or more different quality grades assigned by various Wikipedia projects.

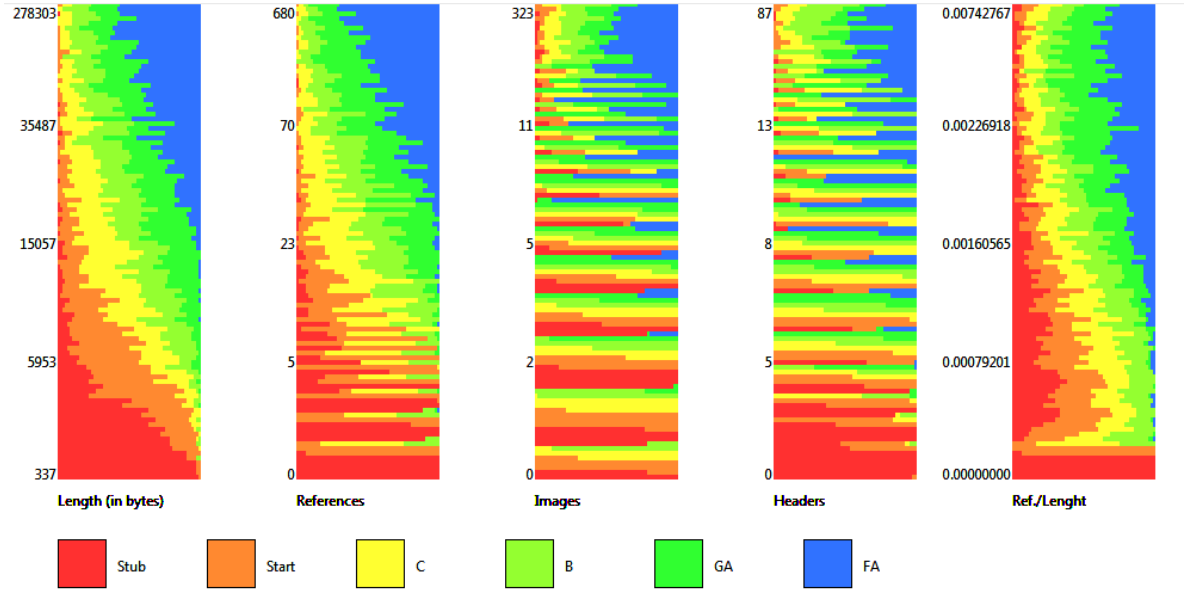

Figure 1. Distribution of metrics in articles of each quality class in English Wikipedia (featured article (FA)—the highest grade; stub—the lowest). Source: own calculation.

Table 1. Median of metrics in each quality class in English Wikipedia. Source: own calculations.

Quality/Metric	Len	Ref	Img	Hdr	Ral	No. of Articles
FA	49,292.5	113.0	13.0	14.0	0.00231	5117
GA	25,862.0	57.0	8.0	10.0	0.00215	26,126
B	21,791.0	32.0	6.0	11.0	0.00157	69,545
C	14,751.0	21.0	4.0	9.0	0.00147	178,902
Start	6526.0	6.0	2.0	5.0	0.00097	1,300,912
Stub	2182.0	1.0	2.0	2.0	0.00073	2,604,331

In addition to the above metrics, which were used in our previous work [21], we also decided to take into account special quality flaw templates, which can indicate some problems as identified by Wikipedia editors in a considered article. There are 12 types of this template in English Wikipedia, for example, verifiability, the style of writing, the structure, and neutrality [10]. We conducted a preliminary analysis of the best articles for finding quality flaw templates. It turned out that articles with a FA grade virtually did not contain important quality flaw templates. Therefore, including this additional metric is important for decreasing the quality score for articles with high values of content-based metrics and some quality problems at the same time.

3.1. Language Versions

We applied the following selection criteria for language editions of Wikipedia: (a) more than 100,000 articles and (b) editing depth value higher than 20. The latter value reflects the depth of

collaborativeness, that is, how frequently articles are updated (https://meta.wikimedia.org/wiki/Wikipedia_article_depth). This descriptor is highly relevant for Wikipedia. These criteria were met by 44 language versions. The list of languages along with a number of extracted articles and redirects is presented in Table 2.

Table 2. Number of articles and redirects in considered language versions of Wikipedia.

Lang. Code	Full Name	Number of Articles	Number of Redirects
ar	Arabic	540,604	469,411
az	Azerbaijani	124,758	34,223
be	Belarusian	146,060	187,545
bg	Bulgarian	234,409	111,580
ca	Catalan	555,036	360,622
cs	Czech	389,769	246,868
da	Danish	231,498	140,296
de	German	2,102,498	1,403,049
el	Greek	136,682	67,422
en	English	5,479,834	7,865,769
es	Spanish	1,354,835	1,655,009
et	Estonian	161,221	117,093
fa	Persian	575,876	1,471,443
fi	Finnish	422,047	243,497
fr	French	1,910,815	1,464,984
gl	Galician	141,146	55,341
he	Hebrew	212,814	171,196
hi	Hindi	121,141	45,802
hr	Croatian	177,762	50,454
hu	Hungarian	417,182	187,423
hy	Armenian	230,411	316,974
id	Indonesian	410,170	442,416
it	Italian	1,383,839	660,330
ja	Japanese	1,076,601	641,393
ka	Georgian	117,614	37,333
ko	Korean	397,641	336,249
lt	Lithuanian	182,961	79,476
no	Norwegian	475,291	268,180
pl	Polish	1,241,294	407,200
pt	Portuguese	978,485	748,634
ro	Romanian	379,141	495,065
ru	Russian	1,421,808	1,860,232
sh	Serbo-Croatian	439,889	3,537,980
simple	Simple English	127,963	52,026
sl	Slovenian	158,141	65,893
sr	Serbian	356,250	848,652
ta	Tamil	113,146	36,502
th	Thai	119,425	137,551
tr	Turkish	298,523	239,841
uk	Ukrainian	734,784	416,183
ur	Urdu	123,921	191,456
uz	Uzbek	128,997	315,513
vi	Vietnamese	1,161,311	198,618
zh	Chinese	962,982	760,244

3.2. Metrics Extraction

We used our own parser to extract the six considered metrics. This parser uses some of the files from Wikipedia dumps (a complete copy of all Wikimedia wikis, in the form of Wikitext source, raw database tables in SQL and metadata embedded in XML can be found at https://dumps.wikimedia.org/). Below is list of the files that were used by our parser for metrics extraction:

- {lang}wiki-latest-pages-articles.xml.bz2—Recombined articles, templates, media/file descriptions, and primary meta-pages. Used for calculation of articles' length, number of headers and references.

- {lang}wiki-latest-imagelinks.sql.gz—Wiki media/files usage records. Used in calculation of number of images in articles.
- {lang}wiki-latest-templatelinks.sql.gz—Wiki template inclusion link records. Used in calculation of number of quality flaw templates and for searching of articles with selected infoboxes (topics).
- {lang}wiki-latest-redirect.sql.gz—Redirect list. Used for determining articles' name that redirects to other articles.
- {lang}wiki-latest-langlinks.sql.gz—Wiki interlanguage link records. Used for determining name(s) of the article in other language version(s).

In the above file names, {lang} refers to the language code of the Wikipedia edition (as described in Table 2). Thus, for each language version, we downloaded and then processed these five compressed files.

To obtain the most complete list of language links of each article, it is necessary to follow language links from each language version. For example, if an article in a given language has Wikilinks to relevant articles in other languages, one needs to check if the links are mutual. An additional challenge was to overcome redirections in language links of the articles. Summarizing, we collected about 19.3 million language link sets, and 5.6 million remained after removing duplicates. Further refining, on the basis of the similarity analysis, reduced the number of articles to 4.2 million interlanguage link sets.

In the case of counting quality flaws, we had to take into account various names of templates that pointed to specific English counterparts. For this purpose, we used interlanguage links in important quality flaw templates in English Wikipedia to obtain automatically appropriate names for these templates in other languages.

In this paper, we have used the Wikipedia dumps from September 2017.

3.3. Building Quality Measure

As described in [21], often we can observe a positive correlation between the article quality and the value of each of the five considered quality metrics (article length, number of references, images, headers, and references per length). Figure 1 shows how the distribution of articles varies depending on metrics values considered by the example from the largest English Wikipedia version, which is noticeable if we consider the same number of articles with different quality grades.

As mentioned previously, English Wikipedia is the biggest edition, has an extensive grading system, and has a large number of assessed articles. The less developed languages (e.g., Belarusian, Georgian, Serbian, and Czech) do not always behave similarly to their more developed counterparts. However, taking into account the presence of the highest FA grade in all considered language versions of Wikipedia, we could calculate the median value of these best articles in each language. Medians for each considered metric and language versions are shown in Table 3.

The above values were then used as thresholds in our quality measure. As proposed in [21], on the basis of the medians, we normalized each metric, in particular, the Wikipedia language version, according to the following rule: if the value of the given metric in a given language exceeded the threshold, it was set to 100 points; otherwise its value was linearly scaled to reflect the relation of the value to the median value. For example, if the median for the number of references in Japanese Wikipedia was 118, any article with a larger number of references would score 100 for this metric; an article with 59 references would score proportionally 50 points after normalizing.

Changing the value of any metric in a particular Wikipedia language version would have a different effect on the normalized value. For each language version of Wikipedia, each metric could play an important role in assessing the quality; therefore we first counted the normalized metrics average (*NMA*) by the following formula:

$$NMA = \frac{1}{c} \sum_{i=1}^{c} nm_i \quad (1)$$

where nm_i is a normalized metric m_i and c is the number of metrics.

Next we took into account the number of quality flaw templates QFT in the considered article (if they existed) and our final formula for the quality measure reads as follows:

$$QualityScore = NMA - NMA * 0.05 * QFT \quad (2)$$

In articles with a high quality score value (e.g., 90 points), each quality flaw template reduced the quality score by 5% (for one such template in our example, the article had 85.5 points). This way, if an article had the maximum values of a particular metric but at the same time had quality flaw template(s), this would not allow it to obtain the maximum value of the quality score (100).

Table 3. Median metrics values in the highest quality class in various Wikipedia languages. Source: own calculation.

Lang.	Length	References	Images	Headers	Ref/Len
ar	120,704.5	162.5	41.5	27.0	0.00133
az	76,048.0	124.0	26.0	21.0	0.00162
be	170,430.0	197.0	35.0	27.0	0.00113
bg	76,416.0	60.0	22.0	21.0	0.00081
ca	47,890.0	66.0	18.0	17.0	0.00144
cs	70,012.0	123.0	18.0	21.0	0.00196
da	72,937.5	125.0	22.0	29.5	0.00196
de	56,438.0	55.0	17.0	21.0	0.00095
el	89,168.0	83.5	13.0	18.0	0.00094
en	49,316.0	113.0	13.0	14.0	0.00231
es	76,565.5	99.0	19.0	21.0	0.00133
et	16,834.0	27.0	10.0	12.5	0.00203
fa	10,2343.0	147.5	20.5	22.0	0.00141
fi	49,264.0	113.0	15.0	20.0	0.00224
fr	90,736.0	167.0	29.0	26.0	0.00186
gl	89,990.0	157.0	21.0	22.0	0.00203
he	64,263.0	38.0	17.0	19.0	0.0006
hi	74,027.5	38.5	18.0	16.0	0.00057
hr	36,925.0	25.0	14.0	17.0	0.00073
hu	59,459.5	63.0	22.0	21.0	0.00114
hy	157,587.0	169.0	38.0	33.0	0.00108
id	49,018.0	92.0	14.0	16.0	0.00207
it	82,750.0	141.0	29.0	23.0	0.00177
ja	97,329.0	188.0	22.0	29.0	0.00198
ka	92,822.0	46.0	21.0	20.0	0.00043
ko	72,534.0	131.0	20.0	22.0	0.00186
lt	52,274.0	44.0	27.0	22.0	0.00056
no	62,999.0	77.0	20.0	23.0	0.00108
pl	59,967.0	97.0	16.0	17.0	0.00168
pt	70,432.5	146.0	23.0	17.0	0.00209
ro	83,933.5	154.0	24.0	21.0	0.00197
ru	139,812.0	164.0	24.0	22.0	0.00117
sh	55,668.0	65.0	15.0	17.0	0.00116
simple	22,231.0	51.0	8.0	9.0	0.00227
sl	40,176.0	51.5	12.0	16.0	0.00135
sr	112,775.0	109.0	29.0	24.0	0.00098
ta	96,282.0	24.0	21.0	19.0	0.00017
th	122,833.0	91.0	16.0	22.0	0.00088
tr	65,254.0	98.0	18.0	17.0	0.00177
uk	84,159.0	41.0	25.0	21.0	0.00051
ur	54,045.5	31.5	17.5	21.0	0.00058
uz	55,387.0	27.5	22.0	26.0	0.00081
vi	89,724.0	138.0	21.0	20.0	0.00164
zh	43,215.0	91.0	12.0	12.0	0.00219

After the assessing of more than 28 million articles in 44 considered language editions of Wikipedia, we found that most of the articles obtained scores of between 0 and 30 points. Figure 2 shows the distribution of articles in this scale (a more detailed and interactive chart is found on the Web page: http://data.lewoniewski.info/informatics2017/).

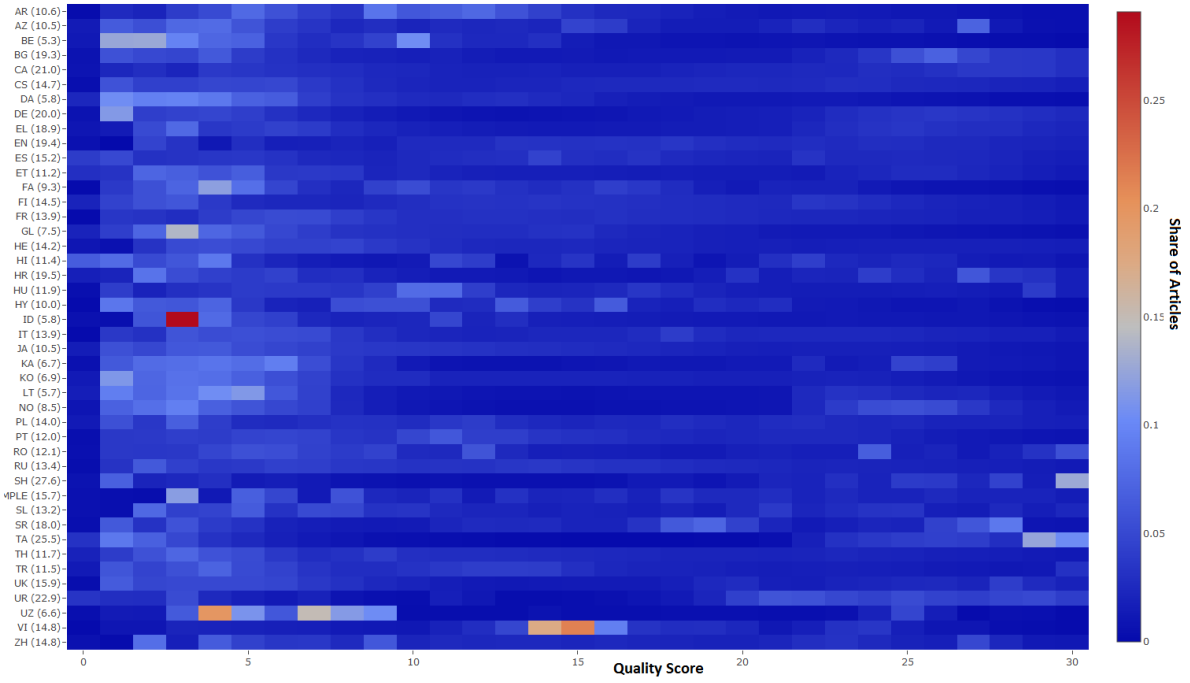

Figure 2. Distribution of articles depending on quality score in each language version of Wikipedia. The medians of the quality scores for each language edition are in parentheses. Source: own calculation.

4. Popularity Measure

The quality of Wikipedia articles can change over time. This is particularly true for articles that contain time-sensitive information. If they are not regularly updated or are updated with delays, their quality will decrease over time. A lower quality will be observed particularly in comparison to equivalent articles in other languages that may be updated regularly. We can expect that more-popular language versions of the article will be verified by authors more often and, if necessary, can be updated faster than less-popular language versions. To some extent, this is reflected in the Wikipedia article depth measure. Concluding, it can be useful to consider popularity metrics of the articles.

Similarly to other studies [19,20], we have used the page view information in order to measure the popularity of articles. Wikipedia records data on users visiting their pages in all language versions every hour to special compressed files (https://dumps.wikimedia.org/other/analytics/). In order to measure the popularity of articles, we downloaded these data files with statistics for the last year (from September 2016 to August 2017)—about 442 GB of compressed raw data.

We define the following popularity metrics:

- *tp*—total popularity: total number of visits during the considered period;
- *sp*—stable popularity: stable number of visits, which is calculated as the median of daily visits during the considered period.

In order to calculate the relative popularity, we normalized both metrics with regard to maximum values of popularity metrics in corresponding articles in other languages. Thus, for the popularity

metric $_p$ of the particular article with v numbers of language versions, the language l_p^* with the maximum value can be found by the formulas:

$$l_{tp}^* = \arg\max_{v=1..n} tp(v), \quad l_{sp}^* = \arg\max_{v=1..n} sp(v) \tag{3}$$

Now, in order to calculate the relative popularity RP (on a scale between 0 and 100) of the selected language version l of the article, we counted using the average of the normalized popularity metrics tp and sp:

$$RP(l) = \frac{tp(l)}{tp(l_{tp}^*)} \times 50 + \frac{sp(l)}{sp(l_{sp}^*)} \times 50 \tag{4}$$

We consider an example. We suppose we have three language versions of the article—*en*, *de*, and *fr*. For each language, we have the following popularity metrics:

- total popularity $tp(en) = 2000$, $tp(de) = 1000$, and $tp(fr) = 500$;
- stable popularity $sp(en) = 30$, $sp(de) = 40$, and $tp(fr) = 20$.

English (1) has the highest value of the tp metric; therefore $l_{tp}^* = en$, and we normalize using the value $tp(l_{tp}^*) = 2000$: $tp^*(en) = \frac{2000}{2000} = 1$, $tp^*(de) = \frac{1000}{2000} = 0.5$, and $tp^*(fr) = \frac{500}{2000} = 0.25$.

German (2) has the highest value of the sp metric; therefore we normalize using the value $sp(l_{sp}^*) = 40$ for $l_{sp}^* = de$: $sp^*(en) = \frac{30}{40} = 0.75$, $sp^*(de) = \frac{40}{40} = 1$, and $sp^*(fr) = \frac{20}{40} = 0.5$.

Now substituting the normalized values into Equation (4), we obtain the following values of the relative popularity measure for each considered language version of the article:

- $RP(en) = 1 \times 50 + 0.75 \times 50 = 87.5$;
- $RP(de) = 0.5 \times 50 + 1 \times 50 = 75$;
- $RP(fr) = 0.25 \times 50 + 0.5 \times 50 = 37.5$.

5. Wikipedia Articles' Assessment

In this section, we present the results of the quality and popularity assessment of Wikipedia articles in 44 languages on different topics: companies, films, persons, universities, and video games.

5.1. Dataset

Wikipedia provides a system of categories, specific to each language, that allows for the grouping of articles. Thus, each language version of Wikipedia usually has its own structure of categories and own practices concerning their assignment. For example, in some languages, it is customary to tag an article with more than 20 categories; in others, the number can be limited to 2–5 categories. The quality of structure of categories also differs among languages. For example, in some language versions, articles about people, events, transport and other topics can be assigned to just one category.

A more reliable approach for classification is based on the infobox system. An infobox is a table, located usually at the top right-hand corner of an article, that concisely presents main facts about the subject. Depending on the topic described, infoboxes have different names. This allows other popular knowledge bases (e.g., DBpedia, https://dbpedia.org) to develop detailed ontology on the basis of these Wikipedia templates [22]. Popular infoboxes usually have their own names in various languages. For the purpose of our research, we have chosen 12 different infobox types on the basis of popularity in English Wikipedia. Using interwiki links, we extracted infobox names in other language versions. Table 4 shows that almost all languages of Wikipedia have equivalents of popular infoboxes in the English version.

In order to define groups of the articles that described the same topic, we extracted lists of articles separately for each infobox in a particular language version. In some languages, the lack of an infobox does not mean the absence of articles on a given topic. For example, German Wikipedia does not use infoboxes for people (office holders, musicians, etc.). Moreover, there is no obligation to add an infobox

at all. However, it is often considered an important element of an article's quality. In such cases, we can use interwiki links from identified articles in some languages to reach articles in other versions. Results of the above procedure are presented in Table 5, which presents the number of articles on a particular topic in the analyzed Wikipedia languages.

Table 4. Number of considered language versions of Wikipedia with particular infobox. Source: own calculation.

Infobox Name	Abbreviation	No. of Lang.
Album	Album	41
Company	Comp.	41
Film	Film	43
Football biography	Footb.	38
Musical artist	Music.	40
Officeholder	Office	35
Person	Person	41
Settlement	Settl.	42
Taxobox	Taxobox	43
Television	Telev.	41
University	Univ.	40
Videogame	Videog.	43

Table 5. Number of articles on particular topic in various Wikipedia languages. Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Office	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
az	246	540	4692	1759	2042	3773	14,818	9886	7755	204	956	218
be	168	589	251	2346	1157	5524	12,944	10,155	3870	60	436	111
bg	2644	1133	4919	4866	4395	3416	33,340	27,183	30,660	1240	439	203
ca	1396	1483	8729	8082	3488	4130	128,844	25,239	27,938	766	536	1214
cs	6919	3189	5471	10,449	12,007	3748	58,212	18,288	12,716	1467	547	965
da	2859	2515	12,849	6917	6745	3586	24,469	6478	6077	1082	611	770
de	8699	23,052	33,079	37,653	10,977	6998	97,836	33,749	45,935	6183	3643	3037
el	2020	781	2372	2578	2123	4064	28,655	5641	2019	491	239	377
en	161,207	67,416	123,962	149,140	105,658	142,209	559,453	513,861	337,211	43,421	22,934	22,666
es	37,487	9504	23,071	28,571	35,908	36,945	235,382	168,487	160,470	12,323	3927	6920
et	1437	1053	1625	2428	3016	5609	20,821	17,959	5803	397	507	178
fa	6680	5099	20,037	16,516	8979	12,981	83,842	150,348	25,017	3044	1462	1719
fi	22,230	6432	10,382	9950	15,094	10,948	79,126	22,885	19,758	3700	950	3666
fr	42,030	21,845	51,157	43,026	39,090	41,593	278,194	217,022	111,845	11,041	5201	12,364
gl	3883	1146	2458	2172	4080	4229	21,557	11,170	5087	611	222	498
he	4928	2552	4532	5310	6389	9351	41,876	10,703	7155	2421	603	654
hi	908	872	4307	62	626	5140	7829	7100	1333	563	623	69
hr	4875	1022	1991	3232	3593	4306	21,691	25,491	5127	531	200	345
hu	10,453	2353	5980	16,524	7723	10,396	56,162	101,132	21,410	2998	253	1076
hy	2874	855	3196	2473	3970	3286	22,987	76,528	3216	630	440	149
id	8567	4600	10,519	13,226	5360	12,009	39,419	93,622	96,843	4518	1561	673
it	71,368	13,114	60,999	50,138	31,082	34,395	331,480	183,633	37,408	10,590	1705	8790
ja	28,375	31,715	19,029	16,874	26,501	16,449	100,936	43,253	15,758	4832	2917	8696
ka	4634	602	1690	1655	2111	5172	14,554	30,792	10,582	384	248	204
ko	7234	7510	10,446	11,209	11,703	9015	54,498	24,350	14,142	7389	1721	2646
lt	2273	1387	2129	2644	3400	3974	14,870	21,297	9309	249	453	507
no	11,565	5460	6822	10,836	11,341	14,458	136,405	36,224	28,405	1484	1742	1237
pl	30,606	8185	19,506	39,589	20,363	30,018	172,777	230,483	42,047	5972	2605	3372
pt	36,065	9453	24,044	26,859	25,360	18,047	143,961	153,436	100,580	10,123	2232	5501
ro	4452	2593	4390	4743	5321	8085	36,072	157,473	32,008	1417	484	763
ru	22,059	12,940	28,386	32,145	30,248	48,437	199,057	244,567	40,238	5515	3229	6251
sh	1735	657	8898	2261	2445	6521	25,881	119,863	2578	1268	623	101
simple	3605	1488	2689	5713	5281	6061	28,633	25,203	4350	1258	836	968
sl	1536	965	730	2252	2194	3363	31,796	27,712	2441	217	274	140
sr	2571	1068	5621	2508	3105	5633	27,707	102,211	9605	1170	335	315
ta	1254	831	4960	166	652	2972	7537	7371	2501	239	724	34
th	2714	1439	2663	1739	3662	5126	13,911	5535	5918	2687	687	796
tr	9641	3689	7294	17,988	8510	10,058	53,068	57,582	6127	2714	993	1318
uk	9880	6031	13,967	15,391	9170	18,095	82,276	176,111	24,649	1626	1817	1656
ur	143	467	322	69	384	1826	7717	64,090	611	99	533	42
uz	90	184	132	177	567	1034	3427	71,794	1024	43	177	15
vi	4231	1915	2706	2694	3297	6014	19,693	201,490	796,749	2278	648	1038
zh	11,059	11,075	10,129	11,571	7663	19,167	68,975	148,416	97,553	11,040	4669	4477

Table 6 presents the results from another perspective. Here we can find out, for each topic, the number of articles that were translated to a given number of languages. As data is best interpreted using visual cues, we also present the phenomenon in Figure 3 (logarithmic scale on vertical axis).

Table 6. Number of articles that have a certain number of language versions (NoL) in particular topics. Source: own calculations.

NoL	Album	Comp.	Film	Footb.	Music.	Office	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
1	263,745	129,613	226,052	187,770	164,297	198,470	912,559	842,467	113,9504	74,482	31,834	31,437
2	90,299	39,045	92,473	103,279	74,016	84,295	527,270	574,990	425,054	26,976	12,160	17,700
3	54,343	22,640	53,076	66,452	47,037	52,638	368,955	386,807	180,521	15,538	6996	12,385
4	38,451	15,753	36,475	48,100	34,145	38,797	278,537	296,388	100,373	10,656	4512	9059
5	28,929	11,951	27,227	37,441	26,617	30,934	219,460	235,161	70,777	8009	3380	6962
6	22,404	9498	21,456	30,388	21,390	25,496	176,480	196,752	53,416	6199	2670	5514
7	17,602	7800	17,572	25,042	17,693	21,374	143,829	167,865	42,560	4930	2167	4496
8	14,296	6482	14,636	20,365	14,876	18,262	118,437	149,570	34,636	4021	1806	3687
9	11,774	5504	12,344	16,987	12,610	15,663	98,543	131,890	28,594	3324	1526	3063
10	9652	4696	10,617	14,315	10,793	13,625	82,757	118,162	23,935	2808	1310	2536
11	7957	4017	9228	12,065	9337	11,884	70,551	106,233	20,298	2379	1129	2094
12	6647	3481	8060	10,178	8121	10,491	60,805	92,093	17,250	1999	994	1774
13	5579	3014	7056	8846	7116	9277	52,603	76,816	14,855	1664	876	1513
14	4725	2635	6177	7746	6225	8303	46,105	61,660	12,745	1428	779	1309
15	3973	2310	5411	6862	5525	7467	40,671	48,624	11,085	1232	673	1112
16	3357	1994	4729	6154	4925	6725	36,019	37,011	9704	1051	599	935
17	2840	1753	4133	5507	4366	6103	32,020	30,276	8514	893	523	797
18	2398	1567	3642	4956	3861	5536	28,548	25,539	7543	763	470	669
19	2040	1393	3184	4486	3427	4996	25,570	22,281	6694	646	430	575
20	1734	1256	2787	4057	3027	4554	22,873	19,520	5951	543	381	492
21	1449	1130	2445	3656	2658	4181	20,572	16,850	5293	463	340	421
22	1250	995	2164	3276	2385	3841	18,622	13,717	4740	406	300	362
23	1049	879	1907	2971	2126	3503	16,878	10,561	4248	361	261	298
24	899	775	1662	2641	1895	3236	15,245	8476	3801	309	240	249
25	744	690	1462	2371	1676	2955	13,780	7031	3405	268	218	206
26	617	618	1282	2145	1500	2728	12,553	6103	3030	237	191	171
27	504	559	1116	1942	1340	2498	11,406	5407	2681	199	166	134
28	393	483	970	1745	1193	2289	10,371	4845	2360	172	147	113
29	315	440	840	1567	1066	2079	9381	4321	2096	140	130	96
30	234	369	737	1240	945	1858	8200	3876	1864	122	116	80
31	177	335	626	1008	844	1684	7285	3400	1635	98	99	59
32	130	294	523	870	763	1505	6515	2946	1433	85	87	49
33	89	261	431	664	660	1315	5632	2538	1245	73	75	43
34	72	222	346	516	573	1153	4877	2193	1089	62	67	35
35	48	190	277	372	494	1013	4150	1856	924	52	60	26
36	36	162	207	284	416	885	3502	1522	782	47	51	20
37	22	137	155	222	354	751	2970	1321	627	42	40	15
38	16	117	111	172	298	619	2477	1149	518	38	33	8
39	13	96	74	130	241	504	2007	974	414	29	24	4
40	6	80	42	91	195	406	1594	812	328	22	20	2
41	4	65	24	55	136	320	1201	683	251	17	16	0
42	3	47	14	29	97	246	876	546	180	11	12	0
43	1	36	6	13	60	170	584	389	118	6	7	0
44	1	17	1	7	21	94	310	228	59	0	4	0

Another possibility to analyze the data on language versions from Table 5 is to show overlaps between a group of three languages using Venn diagrams. These show how many articles specific languages have in common (see Figure 4).

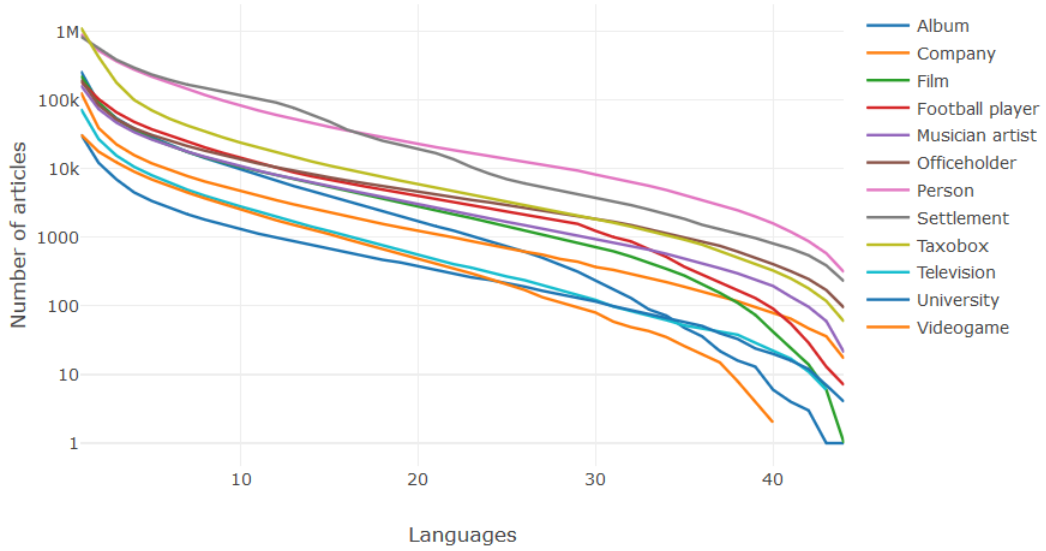

Figure 3. Number of articles that have a certain number of language versions in particular topics. Source: own calculations.

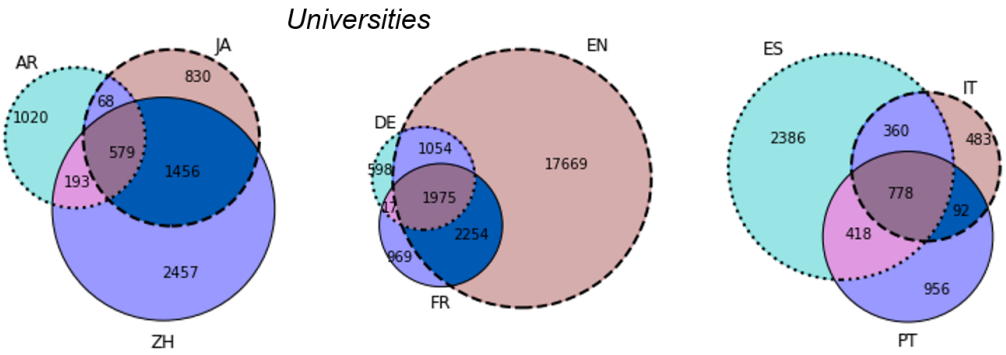

Figure 4. Coverage of articles that describe universities in different languages. Source: own calculation. Other interactive Venn diagrams for this paper with different topics and languages are available on the following Web page: http://data.lewoniewski.info/informatics2017/vn/.

5.2. Quality Assessment

For all articles from our dataset, we calculated a synthetic measure of quality as described in Section 3.3. Table 7 presents the average quality scores of articles for each topic in 44 Wikipedia language editions.

If we consider the distribution of the quality scores of Wikipedia articles, we can also observe differences across language versions and topics. Figure 5 presents the distribution of quality scores for three Wikipedia language versions (English, German, and French) in 12 considered topics (charts for other languages are available from the Web page: http://data.lewoniewski.info/informatics2017/).

Table 7. Average quality scores of articles for each topic in 44 Wikipedia language editions. Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Office	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
ar	12.5	14.0	10.8	8.9	14.2	16.6	13.7	12.8	14.2	13.5	13.9	13.9
az	13.1	12.4	10.3	15.9	11.1	13.1	11.0	16.4	13.1	15.2	20.7	11.7
be	12.0	10.1	11.9	10.8	9.9	9.1	8.7	7.4	9.8	12.1	8.1	20.4
bg	12.4	17.9	16.3	14.1	17.0	16.0	14.9	23.5	26.2	15.3	21.6	11.8
ca	18.3	23.9	23.0	20.0	21.8	22.5	18.1	23.0	25.0	24.4	23.6	22.6
cs	12.1	19.9	10.2	18.6	14.1	17.7	15.1	19.9	22.2	14.2	15.4	15.6
da	11.9	13.5	11.4	11.6	11.3	11.4	9.3	10.6	10.6	11.2	8.5	12.5
de	29.5	29.5	23.2	19.6	27.3	26.8	21.3	28.6	28.5	24.7	25.2	32.7
el	23.6	23.8	22.9	20.7	26.2	23.9	20.6	25.2	31.7	21.9	22.9	24.7
en	23.8	29.5	21.1	19.1	27.7	24.3	26.0	20.8	18.8	23.8	28.1	31.3
es	18.4	21.0	14.0	18.1	18.4	19.9	18.8	16.1	20.9	20.8	18.8	18.2
et	10.6	21.2	9.8	11.9	14.3	14.1	15.3	15.0	21.6	15.1	12.5	14.1
fa	9.3	14.3	7.6	7.8	8.9	12.2	8.7	12.8	7.5	10.2	15.7	12.9
fi	14.4	18.5	14.0	15.1	13.1	16.0	15.1	21.0	21.1	14.9	17.4	14.4
fr	14.9	19.5	13.7	15.8	17.9	16.4	17.0	19.8	11.8	18.9	18.0	16.8
gl	5.9	13.6	10.9	7.9	10.7	10.2	10.7	11.9	22.4	11.8	11.8	11.8
he	18.0	22.5	20.1	14.6	18.1	20.2	19.5	25.6	18.1	19.3	23.3	17.3
hi	23.0	25.8	17.8	46.3	27.0	19.3	19.5	18.9	19.7	19.9	14.1	30.3
hr	18.8	21.8	23.5	20.6	19.8	18.3	18.4	23.4	21.6	24.0	18.9	21.4
hu	16.3	21.8	19.5	15.3	19.2	17.6	16.7	15.2	18.0	18.6	23.5	22.7
hy	17.6	17.2	12.8	18.1	14.7	12.5	13.1	12.1	10.4	16.4	13.9	16.0
id	15.4	19.3	16.3	10.5	21.0	16.8	16.3	6.8	4.2	17.2	18.0	19.4
it	13.9	18.4	12.8	22.2	18.4	16.9	17.7	15.2	21.8	17.2	17.7	15.6
ja	10.5	15.6	15.1	17.4	16.1	15.0	15.7	19.5	18.0	22.9	19.1	18.1
ka	24.6	15.5	12.9	12.9	16.7	9.9	11.1	13.7	19.2	16.6	15.0	26.7
ko	11.5	13.3	8.7	9.0	12.3	12.5	11.2	7.7	16.3	13.9	13.7	17.7
lt	12.7	16.3	8.4	21.6	11.8	13.7	15.1	11.8	7.9	14.6	17.8	14.8
no	11.0	15.3	16.1	17.1	13.9	15.9	15.6	19.8	12.5	19.1	11.2	20.1
pl	16.0	18.5	12.1	11.6	19.4	15.4	15.4	19.1	20.6	17.9	23.6	19.6
pt	19.6	17.3	15.9	14.7	16.7	15.7	15.1	15.2	10.6	19.3	15.7	18.9
ro	17.2	18.1	15.6	15.6	16.4	12.9	13.5	16.1	23.1	15.1	16.0	16.9
ru	22.5	21.1	14.3	20.1	17.2	16.3	16.8	16.2	20.5	20.3	18.4	24.1
sh	17.8	18.7	12.5	9.7	15.4	13.2	12.8	26.0	20.8	12.5	15.5	16.5
simple	18.1	20.4	15.3	14.9	20.1	22.0	20.9	15.6	21.2	16.8	19.4	17.7
sl	25.9	20.5	13.9	8.6	17.9	19.1	14.2	23.2	21.1	15.6	13.8	23.7
sr	11.0	17.3	7.7	14.3	15.7	14.2	13.8	17.6	24.1	13.3	13.3	16.4
ta	16.5	25.6	11.9	17.1	24.1	24.8	23.8	26.5	26.7	18.6	21.4	27.8
th	17.3	19.5	15.7	14.8	18.2	19.9	17.5	18.4	19.2	15.7	21.1	19.3
tr	13.7	15.8	12.4	9.9	14.6	13.6	12.8	14.4	14.5	12.8	16.4	14.5
uk	19.3	20.7	14.9	20.5	18.0	16.0	16.9	24.2	17.0	20.3	18.0	25.7
ur	16.3	21.7	15.7	18.9	16.7	15.5	19.3	24.6	16.2	16.2	22.5	15.0
uz	13.3	15.2	17.1	13.7	13.1	14.9	13.0	8.0	11.0	12.2	11.7	10.0
vi	26.9	20.4	19.2	17.9	21.9	18.4	18.4	12.1	16.0	17.9	17.2	22.2
zh	22.3	25.0	26.8	21.6	27.9	22.5	21.8	12.1	13.5	27.8	29.3	29.4

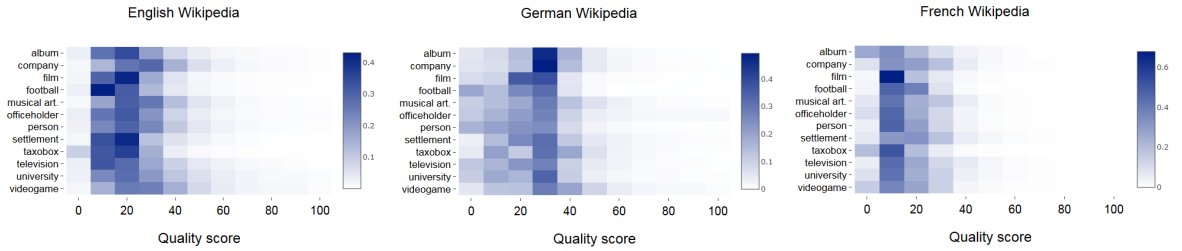

Figure 5. Distribution of quality scores for three Wikipedia language versions (English, German, and French) in 12 considered topics. Source: own calculation.

5.3. Popularity Assessment

Our goal is to look for correlation between quality and popularity. Therefore, we also collected data about popularity as described in Section 4. In Table 8, we present the average popularity metric tp for articles in each topic in 44 Wikipedia language editions.

Table 8. Average popularity metric tp in articles for each topic in 44 Wikipedia language editions. Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Office	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
ar	940.6	1842.0	1578.2	328.0	2015.0	2852.0	1294.5	339.6	383.3	1997.4	1102.7	509.7
az	503.5	466.0	130.6	122.3	464.8	511.8	319.9	148.8	152.5	212.0	364.9	126.3
be	111.0	218.2	97.9	40.0	121.0	106.1	124.0	33.4	87.6	150.2	112.8	78.4
bg	162.9	904.3	483.3	376.9	1146.4	1419.1	613.5	247.7	168.3	1306.2	569.0	582.7
ca	94.1	342.7	115.5	66.5	396.0	355.1	130.5	109.1	136.9	545.4	204.2	97.9
cs	275.3	1603.8	850.3	350.9	1542.8	3414.3	1130.7	802.4	1910.9	2246.4	992.1	1352.9
da	208.3	920.0	223.5	326.5	856.2	1337.8	750.7	523.1	613.3	1496.3	312.5	453.5
de	2609.7	5147.7	6075.4	1263.6	11,532.1	12,524.1	6267.5	4579.8	2929.9	15,321.5	2551.3	6210.1
el	276.7	1539.8	1143.6	918.0	1796.0	1563.8	971.2	1114.3	2121.2	3287.0	1129.7	595.0
en	11,111.2	14,451.0	16,943.0	3250.7	18,625.7	9016.0	14,687.2	2491.9	2235.3	26,019.4	7132.1	21,296.7
es	3495.3	7508.5	7622.6	3110.3	7905.3	5143.7	4634.8	1369.0	1014.3	10,001.4	3242.6	5122.7
et	130.2	408.5	214.8	115.5	462.0	343.2	312.6	228.4	474.5	535.4	231.3	243.1
fa	869.7	1154.8	949.2	290.8	845.3	1510.3	801.7	120.0	347.9	1298.4	1297.1	554.7
fi	371.6	964.7	793.9	172.0	1044.6	803.6	561.0	572.3	659.4	1327.2	482.2	609.7
fr	2446.7	3997.3	3541.0	1457.4	4577.6	3759.7	3223.2	1041.3	872.6	9042.3	2020.9	1824.7
gl	33.5	159.0	59.2	48.7	102.5	98.5	96.4	103.1	133.9	105.0	128.9	61.8
he	920.1	1461.4	1438.0	545.9	1198.3	942.4	897.5	1098.1	1089.2	2312.9	861.4	1020.5
hi	265.0	681.2	120.4	614.5	505.8	569.5	961.8	315.6	1255.8	228.2	239.1	174.2
hr	247.7	985.7	623.0	424.4	1087.1	899.8	710.8	329.8	700.2	1186.9	405.3	557.9
hu	522.1	1374.7	1473.9	264.4	1617.8	1326.3	975.0	222.8	622.7	1963.6	1655.3	1112.1
hy	74.7	256.9	119.4	98.2	178.4	286.6	205.9	25.7	353.0	232.5	252.1	151.7
id	489.3	1472.4	621.0	181.7	1382.2	1148.6	718.8	204.4	105.5	920.1	1484.2	833.1
it	1352.8	2849.5	2585.8	1352.4	3431.3	2137.4	1724.3	639.5	1008.2	7068.9	1565.2	1847.2
ja	4217.1	6841.4	10,135.9	2112.0	11,154.6	7079.9	7882.0	2509.4	6141.3	25,687.4	6324.1	8822.1
ka	63.4	555.3	230.2	195.4	404.8	485.1	470.6	108.3	154.8	274.8	395.4	218.3
ko	802.7	1617.6	564.9	334.7	1224.4	1370.5	878.3	385.6	369.2	1762.8	1121.2	862.5
lt	141.0	510.3	210.4	97.8	432.8	593.8	491.1	228.7	460.0	547.5	393.1	307.3
no	190.4	525.6	372.7	241.4	606.4	466.3	270.2	293.4	226.5	931.1	223.3	354.7
pl	922.1	3305.7	1765.5	714.8	3805.7	2328.9	1753.5	485.6	1410.6	3654.2	1151.9	2152.0
pt	1348.1	3011.0	2071.7	1959.6	3637.2	2786.6	2283.5	549.8	412.6	4593.0	1601.3	2611.7
ro	280.3	880.6	499.7	432.1	1209.5	1007.2	781.0	99.8	180.8	996.2	543.9	747.3
ru	7657.9	7968.8	12,011.1	2904.5	8646.7	5561.7	6182.5	1464.4	3507.1	21,073.4	4641.4	17,428.5
sh	170.3	494.3	105.2	144.5	567.0	244.5	264.2	32.2	437.7	268.4	108.8	282.8
simple	139.8	486.4	187.5	79.6	249.7	492.9	321.8	143.4	775.0	187.3	186.0	178.5
sl	133.8	460.6	376.7	131.2	644.9	353.3	234.7	128.1	888.9	717.0	241.5	322.5
sr	449.2	806.9	582.9	562.7	1391.9	1102.1	840.3	114.7	321.4	1358.4	513.6	689.4
ta	111.4	262.0	64.8	67.0	186.5	307.9	302.2	122.7	285.6	130.7	91.5	158.7
th	780.4	2827.1	1651.4	1209.8	3302.6	2371.8	2277.0	2077.7	1624.1	3554.5	4558.2	938.1
tr	846.2	2424.1	1960.9	678.7	2135.5	2785.6	2102.8	464.7	1372.1	3826.7	1714.1	2111.2
uk	271.6	800.8	309.5	141.2	703.7	558.1	420.2	124.3	378.0	897.6	695.6	554.0
ur	70.7	177.2	57.0	69.4	135.7	218.7	146.4	16.8	214.7	81.0	68.3	96.5
uz	105.4	408.9	119.1	112.6	152.5	270.0	197.0	21.9	155.2	167.8	182.1	205.7
vi	794.2	2695.1	1531.6	1004.7	3342.2	2050.4	1686.5	72.9	14.3	2080.0	1798.1	1149.5
zh	8591.2	5689.7	13,477.2	600.2	17,115.2	4524.9	5499.5	361.8	495.2	17,052.6	3535.2	8218.4

6. Association between Quality and Popularity

In this section, we present a comparison of the quality and popularity of Wikipedia articles in different languages.

As there were additional requirements for relations between languages, we have conducted the analysis on a subset of Wikipedia articles. We selected only those articles in each topic that had at least three language versions (cf. Table 6). We further analyzed combinations of a language and a topic—a pair. Table 9 presents the top 25 pairs with a share of articles, which had the highest quality in comparison to other languages (full data is presented in Table A1 in the Appendix). For example, the first row of this table should be interpreted as follows: regarding the topic "videogame", 60.5% of articles according to our quality score were best described in the English version.

Table 9. Top 25 language–topic pairs with share of articles that have the highest quality in comparison to other languages (articles with at least three language versions were considered). Source: own calculations.

Lang.–Topic	Share of Art.
en–Videogame	60.5%
en–Album	55.5%
en–Company	49.7%
en–Musical artist	49.0%
en–Television	47.8%
en–Film	43.7%
en–University	43.5%
en–Officeholder	39.3%
en–Person	38.7%
en–Football	29.1%
en–Taxobox	27.1%
en–Settlement	21.2%
it–Football	18.1%
uk–Settlement	15.1%
de–Film	14.7%
es–Taxobox	13.9%
vi–Taxobox	13.5%
de–Company	13.3%
fr–Settlement	11.5%
zh–University	10.9%
zh–Television	10.5%
it–Person	10.0%
de–Football	9.1%
pl–Settlement	8.8%
de–Taxobox	8.0%

An analogous table was prepared for popularity. Table 10 presents the top 25 language–topic pairs with the share of articles that attracted the greatest popularity in comparison to other languages (full data is presented in Table A2 in the Appendix). Similarly to the previous table, the first row of this table should be interpreted as follows: regarding the topic "album", 85.8% of articles had the English version as the most popular (attracted the greatest number of visits).

The goal of our research is to analyze the association between quality and popularity. We have done this on two levels, using appropriate statistics, both parametric and non-parametric.

We first present results of a parametric test using a phi coefficient, calculated for each language–topic pair. This is a measure of association for two binary variables. Our variables were coded as follows: if an article about a specific topic in a given language was of the highest quality among all languages, then it was assigned a value of 1 (high score); otherwise, it was assigned 0 (low score). Popularity was coded similarly: if an article about a specific topic in a given language was the most popular among all languages, then it was assigned a value of 1 (high score); otherwise it was assigned 0 (low score).

Then, the phi coefficient was calculated by the following formula:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_1 n_0 n_0 n_1}} \quad (5)$$

where n_{11} is the number of articles of high quality and popularity scores, n_{10} is the number of articles that have a high quality and low popularity score, n_{01} is the number of articles that have a low quality and high popularity score, and n_{00} is the number of articles that have low quality and popularity scores.

Table 10. Top 25 language versions and topics with share of articles that have the highest popularity in comparison to other languages (articles with at least three language versions were considered). Source: own calculations.

Lang.—Topic	Share of Art.
en–Album	85.8%
en–Videogame	85.7%
en–Taxobox	73.6%
en–Film	73.3%
en–Musical artist	67.6%
en–Company	66.3%
en–Television	64.4%
en–Person	62.4%
en–Football	55.2%
en–Officeholder	54.4%
en–University	50.8%
en–Settlement	39.1%
ru–Settlement	16.9%
ru–Officeholder	14.2%
fr–Settlement	11.3%
ru–Football	9.6%
es–Television	9.5%
ja–University	9.3%
es–Football	8.9%
ru–Person	8.4%
ja–Television	8.0%
ja–Company	7.7%
ru–University	7.6%
ja–Videogame	7.5%
es–Officeholder	6.9%

Depending on the language and topic, the correlation may differ significantly. Table 11 shows the top 25 language–topic pairs with the highest correlation coefficients (full data is presented in Table A3 in the Appendix).

The problem with the phi coefficient, a special case of Pearson's correlation coefficient, is that the results have a high granularity and that it cannot be easily generalized. Therefore, we also set up another experiment, in which we estimated the association between the quality and popularity within a topic. For every topic, we prepared two lists of languages: one ordered by the share of articles that were of highest quality (see Table A1), and the other ordered by the share of articles that were the most popular (see Table A2). These lists were effectively ranks. We wished to know whether the order of the languages was similar, which would support the hypothesis that quality and popularity are associated. For this purpose, we used Spearman's rank correlation coefficient between shares of articles (also used by [19] in similar tasks). The results are presented in Table 12.

Spearman's rank correlation assesses the strength of a link between two sets of considered data, which in our case reached 0.87 (for the topic "company"). The results show that depending on the topic, we could find a different correlation between quality and popularity, but a coefficient of no less then 0.61 (for the topic "settlement"). All associations were statistically significant (as shown by p-values). Overall, the results of our calculations supported the hypothesis that there is an association between high quality of articles and their popularity. However, the association strength depends on the topic and the language version of Wikipedia.

Table 11. Top 25 language versions and topics with the highest phi coefficients between articles with the highest quality and popularity (articles with at least three language versions were considered). Source: own calculations.

Lang.–Topic	Correlation Coeff.
th–University	0.838
th–Officeholder	0.762
vi–University	0.719
pl–University	0.717
pt–University	0.707
id–University	0.705
th–Musical artist	0.684
es–University	0.683
tr–University	0.677
en–University	0.676
id–Settlement	0.656
fa–Television	0.655
et–Television	0.65
sl–Film	0.642
cs–University	0.64
ja–Company	0.637
fr–University	0.636
pt–Television	0.632
en–Television	0.63
en–Company	0.625
zh–Officeholder	0.615
ja–University	0.614
vi–Musical artist	0.607
bg–University	0.603
bg–Television	0.602

Table 12. Spearman's rank correlation coefficients for shares of the articles of the highest quality and popularity, on various topics. Source: own calculation using [23].

Topic	Spearman's Rank Cor. Coef.	Two-Sided p-Value
Album	0.7227	3.05×10^{-8}
Company	0.8749	8.29×10^{-15}
Film	0.6408	2.80×10^{-6}
Football biography	0.7872	2.33×10^{-10}
Musical artist	0.8453	5.27×10^{-13}
Officeholder	0.7665	1.32×10^{-9}
Person	0.8370	1.45×10^{-12}
Settlement	0.6146	9.09×10^{-6}
Taxobox	0.6997	1.26×10^{-7}
Television	0.7950	1.15×10^{-10}
University	0.8362	1.60×10^{-12}
Videogame	0.7436	7.35×10^{-9}

7. Conclusions and Future Work

In this paper, we have described how the quality and popularity of Wikipedia articles can be measured across different languages. Depending on the topic and language, different correlations can be observed between the quality and popularity of Wikipedia articles. This can be due to several reasons.

First, there are differences between Wikipedia language communities in terms of the number of experts in each area. In less-developed language versions of Wikipedia, there are a small number of experts (or even an absence of) in some topics. This can be observed particularly in domains not

specific for a given language community. Therefore, for any created article, there is a greater chance to award the highest grade through an assessment procedure than would be possible in a more-developed language version. More-developed versions, having a larger user base, are more demanding and hence more critical. It is more difficult to obtain a high-quality grade when there are more eyes watching.

Second, quality can evolve over time. We suppose that a non-popular article once received a high grade from the community in a less-developed language version. Because it is not popular, there is no incentive to update this article. The same subject in another language can be much more popular, and therefore facts can be updated regularly. In the less-developed language versions, we observe a discrepancy between the "graded" quality and the real quality. The factor that can help in this distinction is popularity.

Third, a large number of unassessed articles make it difficult to build accurate quality models on the basis of awards provided by users. Except for English and French Wikipedia, most of the language versions have a large number of unrated articles. In such models, important metrics are often related to the volume of information (e.g., articles' length, number of references, and images). Unfortunately, these metrics cannot measure other quality dimensions of the article content, such as, for example, timeliness.

The approach for quality assessment presented in this paper takes into account the specifics of the best articles of each language version of Wikipedia. By considering the popularity measure, we can improve the process of identification of language versions with the highest quality.

The proposed quality and popularity measures can be helpful particularly in automated knowledge extraction from Wikipedia articles. One of these such solutions is DBpedia. The problem that is often encountered is a conflict resolution, which is necessary when various language versions concerning the same subject have conflicting information [24]. Our quality metrics can help in building more-effective conflict-resolution strategies for data fusing. An example of such conflict in DBpedia is presented in Figure 6.

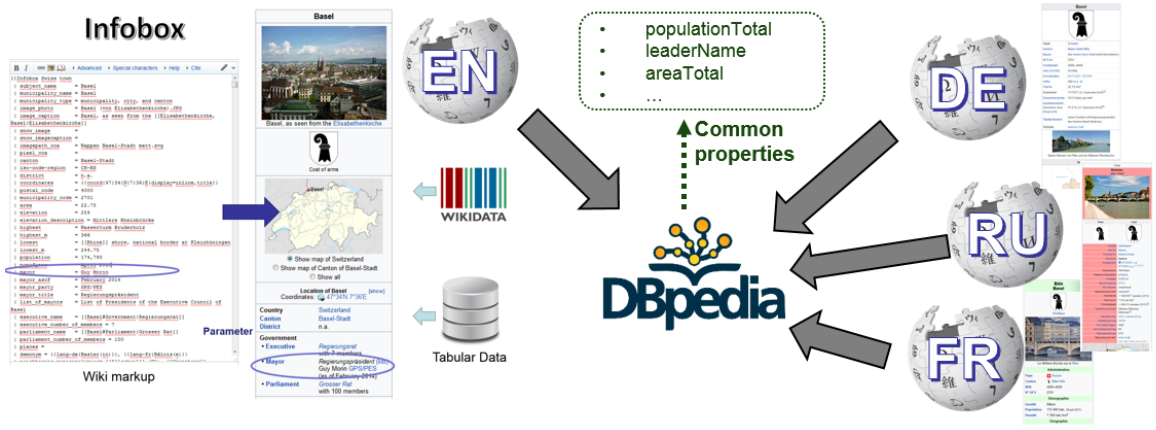

Figure 6. Infobox about Basel with its data sources and its extraction to DBpedia from different Wikipedia language versions.

Conflict resolution is a first step towards the overall objective of enriching less-developed Wikipedia language versions, for which the appropriate information is of poor quality or is even absent. Figure 7 shows the general scheme of enrichment of information by transferring values from infoboxes of language versions with the highest quality and popularity scores. Before transferring values of particular parameters of an infobox, the information is compared to other language versions, and versions with higher quality and popularity scores will have a higher influence (weight) on selecting the proper value.

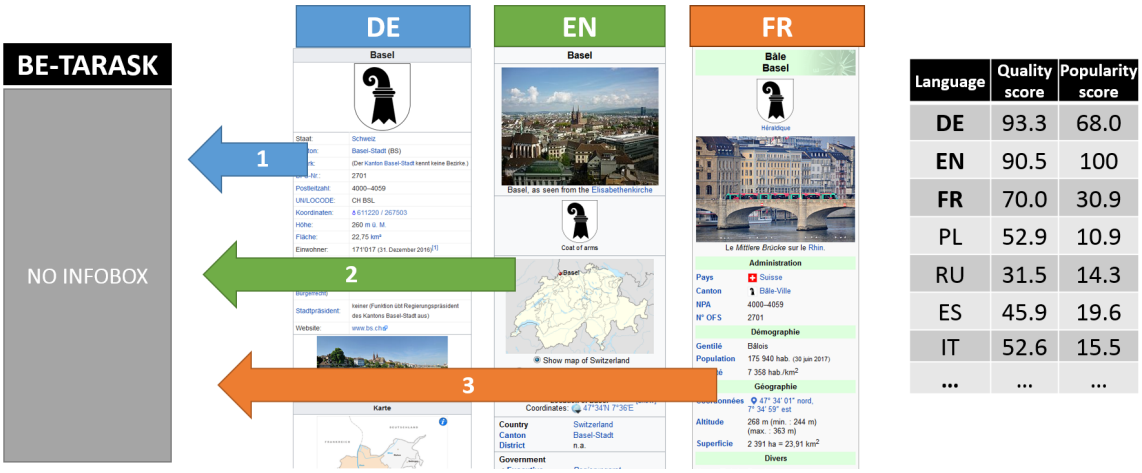

Figure 7. Scheme of information enrichment of Wikipedia infobox on the basis of quality and popularity assessment of other language versions by an example of Basel city. Source: own calculations from September 2017.

The methods proposed in the paper have practical implications. The synthetic quality measure is used in the WikiRank service (http://wikirank.net), which assesses and compare articles in the various language versions of Wikipedia. A quality and popularity assessment of an article can help to evaluate the quality of its important part—the infobox. Such evaluation is used in the Infoboxes service (http://infoboxes.net).

Some of the presented metrics can be expanded. For example, by analyzing the similarity of sources in Wikipedia articles across languages, we can also evaluate the quality of their content [25]. Furthermore, the references themselves can have their own quality metrics (e.g., impact factor), which can be used as an indirect indicator of the article's quality. For popularity measurementsm it can be useful to add some metrics related to link analysis in Wikipedia articles [26]. In the future, we plan to continue studies on new metrics and their extraction methods for improving the Wikipedia article quality assessment model.

Author Contributions: K.W. and W.L. conceived the research problem; W.L. conducted state of the art analysis; K.W. proposed research methodology and designed the experiments, starting from hypotheses to be verified statistically; W.L. collected data and performed the analysis; W.L. and K.W. interpreted the results; W.A. provided an overall guidance.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1

Table A1. Shares of Wikipedia articles with the highest quality score compared with other language versions (articles with at least three language versions were considered). Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
ar	0.0	0.002	0.001	0.008	0.001	0.003	0.007	0.001	0.004	0.001	0.007	0.004
az	0.0	0.0	0.001	0.001	0.001	0.004	0.001	0.006	0.001	0.0	0.004	0.0
be	0.0	0.0	0.0	0.003	0.0	0.001	0.0	0.0	0.0	0.0	0.0	0.0
bg	0.003	0.001	0.006	0.003	0.004	0.003	0.004	0.01	0.03	0.002	0.005	0.0
ca	0.001	0.003	0.03	0.012	0.004	0.004	0.04	0.005	0.033	0.004	0.004	0.008
cs	0.003	0.005	0.002	0.016	0.009	0.002	0.006	0.003	0.006	0.002	0.002	0.002
da	0.001	0.002	0.001	0.003	0.002	0.001	0.0	0.0	0.0	0.001	0.0	0.0
de	0.032	0.133	0.147	0.091	0.035	0.013	0.038	0.015	0.08	0.036	0.071	0.047
el	0.005	0.002	0.003	0.005	0.004	0.006	0.006	0.002	0.002	0.002	0.002	0.002
en	0.555	0.497	0.437	0.291	0.49	0.393	0.387	0.212	0.271	0.478	0.435	0.605

Table A1. *Cont.*

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
es	0.058	0.027	0.017	0.053	0.057	0.062	0.062	0.046	0.139	0.073	0.039	0.025
et	0.001	0.004	0.001	0.002	0.003	0.007	0.004	0.003	0.003	0.001	0.004	0.0
fa	0.001	0.002	0.002	0.003	0.002	0.008	0.003	0.012	0.0	0.002	0.005	0.001
fi	0.018	0.013	0.014	0.007	0.012	0.009	0.012	0.005	0.009	0.005	0.006	0.006
fr	0.032	0.05	0.06	0.052	0.067	0.065	0.077	0.115	0.026	0.038	0.06	0.051
gl	0.0	0.001	0.0	0.0	0.001	0.001	0.001	0.001	0.002	0.0	0.001	0.0
he	0.002	0.006	0.005	0.003	0.008	0.014	0.011	0.002	0.001	0.006	0.009	0.001
hi	0.001	0.002	0.006	0.0	0.001	0.002	0.001	0.002	0.0	0.002	0.002	0.0
hr	0.008	0.005	0.007	0.007	0.008	0.01	0.007	0.016	0.002	0.004	0.003	0.001
hu	0.011	0.006	0.008	0.015	0.01	0.013	0.01	0.015	0.009	0.011	0.003	0.005
hy	0.0	0.0	0.001	0.0	0.001	0.002	0.001	0.019	0.0	0.0	0.0	0.0
id	0.004	0.004	0.007	0.002	0.005	0.005	0.003	0.001	0.002	0.007	0.007	0.001
it	0.063	0.03	0.066	0.181	0.051	0.059	0.1	0.031	0.042	0.057	0.014	0.027
ja	0.011	0.046	0.017	0.036	0.035	0.019	0.018	0.004	0.006	0.043	0.047	0.053
ka	0.01	0.001	0.001	0.001	0.001	0.002	0.001	0.003	0.01	0.0	0.003	0.001
ko	0.003	0.007	0.001	0.003	0.005	0.005	0.003	0.001	0.002	0.011	0.013	0.004
lt	0.001	0.002	0.001	0.003	0.003	0.006	0.002	0.004	0.001	0.0	0.001	0.0
no	0.006	0.008	0.011	0.011	0.012	0.014	0.033	0.011	0.007	0.006	0.006	0.004
pl	0.038	0.016	0.018	0.027	0.038	0.06	0.04	0.088	0.034	0.024	0.025	0.015
pt	0.043	0.015	0.027	0.024	0.02	0.018	0.019	0.028	0.033	0.035	0.013	0.018
ro	0.003	0.003	0.002	0.004	0.003	0.004	0.003	0.017	0.029	0.002	0.003	0.001
ru	0.034	0.032	0.031	0.057	0.04	0.078	0.04	0.039	0.026	0.019	0.04	0.045
sh	0.0	0.001	0.005	0.0	0.001	0.005	0.002	0.036	0.0	0.001	0.002	0.0
simple	0.002	0.002	0.002	0.006	0.004	0.006	0.004	0.004	0.001	0.003	0.004	0.002
sl	0.003	0.001	0.001	0.0	0.003	0.004	0.003	0.017	0.0	0.0	0.0	0.001
sr	0.001	0.001	0.001	0.001	0.002	0.003	0.002	0.002	0.013	0.002	0.001	0.001
ta	0.0	0.002	0.001	0.0	0.001	0.005	0.002	0.002	0.001	0.001	0.003	0.0
th	0.001	0.001	0.001	0.001	0.002	0.003	0.001	0.001	0.002	0.002	0.008	0.002
tr	0.002	0.003	0.002	0.011	0.005	0.007	0.004	0.006	0.002	0.004	0.008	0.002
uk	0.019	0.019	0.021	0.041	0.017	0.037	0.02	0.151	0.011	0.007	0.023	0.01
ur	0.0	0.0	0.0	0.0	0.0	0.001	0.001	0.037	0.0	0.0	0.002	0.0
uz	0.0	0.0	0.0	0.0	0.0	0.001	0.0	0.001	0.0	0.0	0.0	0.0
vi	0.001	0.001	0.002	0.001	0.002	0.002	0.001	0.017	0.135	0.004	0.004	0.002
zh	0.022	0.044	0.032	0.012	0.029	0.032	0.019	0.009	0.025	0.105	0.109	0.052

Appendix A.2

Table A2. Shares of Wikipedia articles with the highest popularity compared with other language versions (articles with at least three language versions were considered). Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
ar	0.0	0.001	0.002	0.004	0.003	0.003	0.006	0.0	0.001	0.001	0.01	0.0
az	0.0	0.0	0.0	0.0	0.001	0.002	0.0	0.0	0.0	0.0	0.004	0.0
be	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
bg	0.0	0.0	0.001	0.001	0.002	0.003	0.001	0.002	0.0	0.001	0.003	0.0
ca	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.0	0.0	0.0	0.0	0.0
cs	0.0	0.003	0.003	0.006	0.006	0.001	0.004	0.001	0.001	0.001	0.003	0.0
da	0.0	0.001	0.001	0.001	0.002	0.003	0.0	0.0	0.0	0.001	0.001	0.0
de	0.005	0.06	0.032	0.041	0.008	0.004	0.017	0.008	0.033	0.023	0.032	0.004
el	0.0	0.001	0.001	0.002	0.001	0.005	0.002	0.002	0.0	0.001	0.001	0.0
en	0.858	0.663	0.733	0.552	0.676	0.544	0.624	0.391	0.736	0.644	0.508	0.857
es	0.019	0.024	0.02	0.089	0.046	0.069	0.065	0.052	0.061	0.095	0.051	0.007
et	0.0	0.001	0.0	0.0	0.001	0.002	0.001	0.004	0.0	0.0	0.001	0.0
fa	0.0	0.001	0.002	0.004	0.002	0.006	0.004	0.005	0.0	0.001	0.008	0.0
fi	0.003	0.004	0.001	0.001	0.006	0.005	0.003	0.002	0.001	0.001	0.002	0.0
fr	0.011	0.032	0.043	0.035	0.029	0.037	0.049	0.113	0.033	0.022	0.051	0.005
gl	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
he	0.0	0.002	0.001	0.002	0.003	0.004	0.002	0.003	0.001	0.001	0.003	0.0
hi	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
hr	0.002	0.001	0.001	0.0	0.004	0.004	0.002	0.011	0.0	0.002	0.001	0.0
hu	0.0	0.001	0.001	0.005	0.003	0.006	0.004	0.019	0.001	0.001	0.003	0.0
hy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.0

Table A2. *Cont.*

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
id	0.0	0.002	0.0	0.0	0.003	0.004	0.001	0.002	0.008	0.0	0.005	0.0
it	0.011	0.018	0.037	0.035	0.019	0.027	0.038	0.024	0.006	0.012	0.012	0.004
ja	0.058	0.077	0.035	0.06	0.06	0.022	0.029	0.007	0.014	0.08	0.093	0.075
ka	0.0	0.0	0.0	0.0	0.0	0.002	0.0	0.0	0.0	0.0	0.0	0.0
ko	0.0	0.006	0.0	0.002	0.002	0.004	0.002	0.001	0.001	0.006	0.011	0.0
lt	0.0	0.001	0.0	0.0	0.001	0.002	0.001	0.001	0.0	0.0	0.002	0.0
no	0.0	0.002	0.001	0.003	0.003	0.003	0.004	0.003	0.001	0.001	0.001	0.0
pl	0.003	0.01	0.006	0.015	0.011	0.032	0.018	0.075	0.015	0.005	0.016	0.001
pt	0.007	0.012	0.004	0.03	0.014	0.012	0.013	0.025	0.008	0.019	0.017	0.002
ro	0.0	0.001	0.001	0.0	0.003	0.005	0.002	0.035	0.001	0.0	0.002	0.0
ru	0.013	0.05	0.051	0.096	0.059	0.142	0.084	0.169	0.052	0.029	0.076	0.036
sh	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.0	0.0	0.0	0.0
simple	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
sl	0.0	0.0	0.0	0.0	0.001	0.001	0.001	0.004	0.0	0.0	0.0	0.0
sr	0.001	0.0	0.004	0.0	0.003	0.005	0.003	0.014	0.0	0.002	0.005	0.0
ta	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
th	0.0	0.002	0.001	0.001	0.002	0.002	0.001	0.003	0.002	0.001	0.008	0.0
tr	0.001	0.004	0.003	0.01	0.006	0.011	0.005	0.012	0.0	0.006	0.012	0.0
uk	0.0	0.001	0.0	0.0	0.001	0.002	0.001	0.007	0.001	0.0	0.009	0.0
ur	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
uz	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
vi	0.0	0.001	0.0	0.0	0.002	0.002	0.001	0.002	0.007	0.001	0.006	0.0
zh	0.006	0.017	0.013	0.002	0.015	0.025	0.01	0.002	0.015	0.042	0.042	0.007

Appendix A.3

Table A3. Phi correlation coefficients of articles with the highest quality and popularity in selected Wikipedia languages. Source: own calculations.

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
ar	0.334	0.199	0.243	0.086	0.380	0.177	0.241	0.104	0.274	0.457	0.390	—
az	—	−0.006	0.463	0.101	0.350	0.263	0.226	0.063	0.158	−0.005	0.458	−0.010
be	0.304	0.263	—	—	0.269	0.099	0.161	0.028	0.349	—	—	—
bg	0.140	0.367	0.223	0.097	0.534	0.488	0.244	0.358	0.139	0.602	0.603	—
ca	0.180	0.051	0.041	0.186	0.073	0.033	0.170	0.167	0.053	0.288	0.181	0.088
cs	0.211	0.455	0.331	0.279	0.405	0.400	0.406	0.209	0.178	0.580	0.640	0.269
da	0.427	0.372	0.358	0.225	0.419	0.461	0.162	0.289	0.182	0.477	0.411	0.405
de	0.334	0.523	0.338	0.364	0.263	0.341	0.309	0.512	0.207	0.535	0.593	0.215
el	0.196	0.303	0.564	0.246	0.238	0.511	0.335	0.026	0.066	0.522	0.134	—
en	0.636	0.625	0.445	0.434	0.576	0.565	0.390	0.454	0.363	0.630	0.676	0.589
es	0.262	0.466	0.295	0.337	0.436	0.567	0.315	0.642	0.419	0.543	0.683	0.185
et	0.207	0.332	0.189	0.151	0.461	0.362	0.283	0.211	0.157	0.650	0.359	—
fa	0.298	0.402	0.386	0.220	0.595	0.455	0.458	0.161	0.141	0.655	0.601	0.181
fi	0.325	0.443	0.254	0.250	0.456	0.557	0.299	0.386	0.166	0.511	0.324	0.176
fr	0.238	0.430	0.237	0.263	0.343	0.362	0.330	0.370	0.204	0.457	0.636	0.205
gl	−0.001	0.103	0.065	−0.005	0.024	0.083	0.247	0.082	0.102	0.367	—	—
he	0.343	0.504	0.281	0.525	0.459	0.410	0.325	0.256	0.179	0.586	0.552	—
hi	—	—	0.074	—	−0.015	−0.006	0.074	0.015	0.200	−0.027	0.026	—
hr	0.262	0.421	0.257	0.402	0.593	0.400	0.352	0.409	0.341	0.443	0.456	—
hu	0.103	0.467	0.252	0.417	0.425	0.458	0.419	0.546	0.093	0.369	0.537	0.094
hy	0.163	−0.004	0.198	−0.002	0.220	0.245	0.124	−0.003	−0.007	0.168	0.341	—
id	0.249	0.362	0.190	0.212	0.435	0.582	0.322	0.656	0.033	0.466	0.705	—
it	0.248	0.438	0.137	0.344	0.337	0.440	0.313	0.466	0.283	0.254	0.536	0.159
ja	0.309	0.637	0.455	0.594	0.562	0.482	0.492	0.447	0.342	0.532	0.614	0.379
ka	0.052	—	0.268	0.020	0.176	0.201	0.195	0.501	−0.001	0.111	0.214	—
ko	0.213	0.454	0.190	0.319	0.369	0.567	0.417	0.552	0.176	0.426	0.511	0.150
lt	0.133	0.247	0.232	0.088	0.353	0.396	0.369	0.541	0.235	0.318	0.558	—
no	0.149	0.338	0.191	0.036	0.253	0.353	0.155	0.170	0.536	0.351	0.342	0.227
pl	0.319	0.607	0.256	0.292	0.414	0.424	0.339	0.524	0.134	0.304	0.717	0.160
pt	0.334	0.474	0.179	0.444	0.496	0.486	0.367	0.553	0.064	0.632	0.707	0.252
ro	0.129	0.445	0.441	0.101	0.498	0.370	0.398	0.113	0.435	0.259	0.473	—
ru	0.268	0.460	0.392	0.365	0.414	0.456	0.330	0.421	0.360	0.470	0.607	0.296

Table A3. *Cont.*

Lang.	Album	Comp.	Film	Footb.	Music.	Offic.	Person	Settl.	Taxobox	Telev.	Univ.	Videog.
sh	—	−0.007	0.075	−0.001	0.067	0.027	0.175	0.011	−0.004	−0.005	0.495	—
simple	0.098	0.091	0.137	−0.007	0.049	—	0.071	0.077	0.074	—	—	—
sl	0.214	0.436	0.642	−0.005	0.329	0.230	0.329	0.113	0.140	−0.026	0.597	—
sr	0.122	0.189	0.305	0.045	0.330	0.251	0.301	0.113	0.081	0.398	0.556	—
ta	−0.012	−0.013	0.010	—	—	0.038	0.039	0.033	0.212	—	0.082	—
th	0.196	0.440	0.407	0.226	0.684	0.762	0.481	0.495	0.151	0.544	0.838	0.173
tr	0.451	0.367	0.453	0.366	0.507	0.397	0.405	0.507	0.290	0.560	0.677	0.163
uk	0.104	0.188	0.083	0.110	0.206	0.216	0.201	0.204	0.091	0.184	0.391	—
ur	—	—	—	—	—	0.104	0.044	0.013	—	—	—	—
uz	—	—	—	—	0.236	0.314	0.201	0.073	0.338	—	—	—
vi	0.121	0.302	0.261	0.420	0.607	0.481	0.369	0.142	0.372	0.209	0.719	0.096
zh	0.363	0.423	0.378	0.250	0.456	0.615	0.410	0.185	0.126	0.550	0.462	0.243

References

1. Staub, T.; Hodel, T. Wikipedia vs. Academia: An Investigation into the Role of the Internet in Education, with a Special Focus on Wikipedia. *Univ. J. Educ. Res.* **2016**, *4*, 349–354.
2. Blumenstock, J.E. Size matters: Word count as a measure of quality on wikipedia. In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 1095–1096.
3. Warncke-wang, M.; Cosley, D.; Riedl, J. Tell Me More: An Actionable Quality Model for Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration, Hong Kong, China, 5–7 August 2013; pp. 1–10.
4. Węcel, K.; Lewoniewski, W. Modelling the Quality of Attributes in Wikipedia Infoboxes. In *Business Information Systems Workshops*; Abramowicz, W., Ed.; Lecture Notes in Business Information Processing; Springer International Publishing: Cham, Switzerland, 2015; Volume 228, pp. 308–320.
5. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Quality and Importance of Wikipedia Articles in Different Languages. In *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings*; Springer International Publishing: Cham, Switzerland, 2016; pp. 613–624.
6. Lex, E.; Voelske, M.; Errecalde, M.; Ferretti, E.; Cagnina, L.; Horn, C.; Stein, B.; Granitzer, M. Measuring the quality of web content using factual information. In Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality—WebQuality'12, Lyon, France, 16–20 April 2012; p. 7.
7. Khairova, N.; Lewoniewski, W.; Węcel, K. Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In *Business Information Systems: 20th International Conference, BIS 2017, Poznan, Poland, June 28–30, 2017, Proceedings*; Abramowicz, W., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 28–40.
8. Lipka, N.; Stein, B. Identifying Featured Articles in Wikipedia: Writing Style Matters. In Proceedings of the 19th International Conference on World Wide Web (2010), Raleigh, NC, USA, 26–30 April 2010; pp. 1147–1148.
9. Xu, Y.; Luo, T. Measuring article quality in Wikipedia: Lexical clue model. In Proceedings of the 2011 3rd Symposium on Web Society (SWS), Port Elizabeth, South Africa, 26–28 October 2011; pp. 141–146.
10. Anderka, M. Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Ph.D. Thesis, Bauhaus-Universitaet, Weimar, Germany, 2013.
11. Wu, G.; Harrigan, M.; Cunningham, P. Characterizing Wikipedia Pages Using Edit Network Motif Profiles. In Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, Glasgow, UK, 24–28 October 2011; pp. 45–52.
12. Suzuki, Y.; Nakamura, S. Assessing the Quality of Wikipedia Editors Through Crowdsourcing. In Proceedings of the 25th International Conference Companion on World Wide Web, Montreal, QC, Canada, 11–15 April 2016; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2016; pp. 1001–1006.
13. Wilkinson, D.M.; Huberman, B.A. Cooperation and quality in wikipedia. In Proceedings of the 2007 International Symposium on Wikis WikiSym 07, Montreal, QC, Canada, 21–23 October 2007; pp. 157–164.
14. Ingawale, M.; Dutta, A.; Roy, R.; Seetharaman, P. Network analysis of user generated content quality in Wikipedia. *Online Inf. Rev.* **2013**, *37*, 602–619.

15. Halfaker, A.; Taraborelli, D. Artificial Intelligence Service "ORES" Gives Wikipedians X-Ray Specs to See Through Bad Edits. Available online: https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/ (accessed on 31 October 2017).
16. Dang, Q.V.; Ignat, C.L. Quality assessment of Wikipedia articles without feature engineering. In Proceedings of the 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), Newark, NJ, USA, 19–23 June 2016; pp. 27–30.
17. Dalip, D.H.; Lima, H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Quality assessment of collaborative content with minimal information. In Proceedings of the IEEE/ACM Joint Conference on Digital Libraries, London, UK, 8–12 September 2014; pp. 201–210.
18. Reinoso, A.J. Temporal and Behavioral Patterns in the Use of Wikipedia. Ph.D. Thesis, Universidad Rey Juan Carlos, Madrid, Spain, 2011. Available online: https://gsyc.es/~ajreinoso/thesis/ (accessed on 31 October 2017).
19. Lehmann, J.; Müller-Birn, C.; Laniado, D.; Lalmas, M.; Kaltenbrunner, A. Reader preferences and behavior on wikipedia. In Proceedings of the 25th ACM Conference on Hypertext and Social Media, Santiago, Chile, 1–4 September 2014; pp. 88–97.
20. Warncke-Wang, M.; Ranjan, V.; Terveen, L.G.; Hecht, B.J. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. In Proceedings of the ICWSM, Oxford, UK, 26–29 May 2015; pp. 493–502.
21. Lewoniewski, W.; Węcel, K. Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure. In *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers*; Abramowicz, W., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 282–292.
22. Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia-A crystallization point for the Web of Data. *Web Semant. Sci. Serv. Agents World Wide Web* **2009**, *7*, 154–165.
23. Wessa, R. Spearman Rank Correlation (v1.0.3) in Free Statistics Software (v1.2.1), Office for Research Development and Education. Available online: https://www.wessa.net/rwasp_spearman.wasp/ (accessed on 31 October 2017).
24. Bryl, V.; Bizer, C. Learning conflict resolution strategies for cross-language wikipedia data fusion. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 1129–1134.
25. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Analysis of References Across Wikipedia Languages. In *Information and Software Technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings*; Damaševičius, R.; Mikašytė, V., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 561–573.
26. Hanada, R.; Cristo, M.; Pimentel, M.D.G.C. How do metrics of link analysis correlate to quality, relevance and popularity in wikipedia? In Proceedings of the 19th Brazilian Symposium on Multimedia and the Web, Salvador, Brazil, 5–8 November 2013; pp. 105–112.