Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

Article

Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles

Włodzimierz Lewoniewski[1,†,‡]*, Krzysztof Węcel[1,‡] and Witold Abramowicz[1,‡]

[1] Poznań University of Economics and Business;
{wlodzimierz.lewoniewski,krzysztof.wecel,witold.abramowicz}@ue.poznan.pl
* 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.

Abstract: Despite the fact that Wikipedia is often criticized for its poor quality, it continues to be one of the most popular knowledge base 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 showed that in language sensitive topics 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 manual assessing of the content quality. This makes automatic quality comparison of articles between various languages a challenging task. The paper presents results of relative quality and popularity assessment of over 28 million articles in 44 selected language versions. In addition, a comparative analysis of the quality and popularity of articles in some topics was conducted. The proposed method allows 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

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

Wikipedia is one the of best examples of success of such collaborative platforms. This encyclopedia became a popular source of information on different topics. Nowadays, it is the 5th most visited page in the world.[1] Wikipedia is free and anybody can make changes in the articles without peer reviewed procedure. The pages of this online knowledge base often appear among the first in search results in Google, Bing, Yandex, and other search engines. However, lack of professional editorial control and certain freedom in editing led to criticism of Wikipedia for the low quality.

There are about 300 language editions on Wikipedia with over 46 million articles.[2] The English edition is the largest and consist of over 5.4 million articles. Each language version of Wikipedia can have own community that defines 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" (FA) – they must be well-written with appropriate structure, comprehensive, well-researched with reliable sources, present views fairly and without bias.[3] Another award – "Good

1 https://www.alexa.com/siteinfo/wikipedia.org
2 https://meta.wikimedia.org/wiki/List_of_Wikipedias
3 https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Article" (GA) – can receive an article, which has not met the criteria for featured article 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 there are "Exzellente Artikel" and "Lesenswerte Artikel" respectively. However, share of such 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 grade". Polish Wikipedia have three additional grades: "Four", "Start", and "Stub". So, we can conclude that there are differences between Wikipedia languages in grading scheme and not all language versions have a developed system of quality grades for articles. For example, the second most largest version of Wikipedia is German, and it has only 2 highest grades – equivalent to FA and GA. Differences in quality grades do not allow to directly compare the quality of the articles between the various language versions. Additional challenge is a large number of articles without grades. For example, in Polish Wikipedia there are over 1.2 million unassessed articles (99% of all articles).

In this paper we present a method of quality assessment of Wikipedia articles as a synthetic measure, taking values between 0 and 100. This approach will be used to evaluate more than 28 million articles in 44 language version of Wikipedia. In addition, comparison of quality between the articles in different languages on selected topics will be made. Paper also presents result of relative popularity estimation of these articles.

2. Related work

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

First works concerning content-based methods concluded that longer articles in Wikipedia often had higher quality [2]. Other papers showed that high quality articles tend to have more images, sections, and references [3–5]. Some scientific works 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. 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 used in another study as the factors that can reflect articles quality – high-quality articles often used more nouns and verbs and less adjectives [9]. Finally, quality evaluation of Wikipedia articles can also base on special quality flaw templates [10].

Second group of studies – user-based – is related to editors' behavior. They try to analyze how the user skills, experience, and coordination of their activities affects to quality of Wikipedia articles. These methods use different metrics related to user reputation and changes that they 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 articles 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.

There are also a few works that try to combine metrics from articles content and edition history [16,17].

Although existing works propose various sets of metrics for assessing quality of Wikipedia articles, there is no universal feature set for this task [17]. Additional challenge is to consider different language versions, which can have different quality models [4,5]. Extraction rules of some metrics (eg. lexical) can also be language sensitive [6,7,9].

Concluding, by using different metrics and models it is possible to estimate the quality of an article. Majority of the approaches is focused only on the one (usually the largest – English) or several language versions. Additionally, these methods basically allow evaluation of articles and comparing 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 we decided to take into the account only important content-based metrics. Most of existing studies evaluates quality of Wikipedia articles as a binary classification problem, which is limited in comparing articles on similar quality. This study is continuation of works on building synthetic measure for quality assessment of Wikipedia articles in different language [18]. Preliminary results have shown the high efficiency of this method in assessing articles in language sensitive topics. Compared with our previous work [18], we decided to increase the number of considered language versions of Wikipedia (from 7 to 44), expand the rules for quality assessment and analyze popularity of the articles.

3. Quality measure

Many of existing studies solve the problem of automatic quality assessment of articles as classification task: articles can be marked as Complete or Incomplete [3–7,9]. This is a big limitation for comparing articles in different languages, as it is not possible to show in what degree one articles is better over 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, which earlier showed high prediction in quality assessment in English Wikipedia [3], as well as in other language editions of Wikipedia [4,5]:

- *len* - article length (in bytes)
- *ref* - number of references
- *img* - number of images
- *hdr* - number of 1st and 2nd level headers
- *ral* - the ratio of number of references and article length

In addition to the above metrics, which was used in our previous work [18], we also decided to take into account special quality flaw templates, which can indicate some problems as identified by Wikipedia editors in considered article. Such template in English Wikipedia can be divided into 12 types, e.g. verifiability, style of writing, structure, neutrality and others [10]. We conducted a preliminary analysis of the best articles for finding quality flaw templates. It turned out that articles with FA grade practically do not contain important quality flaw templates. Therefore, including this additional metric is important for decreasing quality score for articles with high values of content-based metrics and some quality problems at the same time.

3.1. Language versions

In our dataset we decided to include languages editions of Wikipedia with more than 100 000 articles and over 20 editing depth value, which show the depth of collaborativeness and how frequently articles are updated. These indicators correspond to 44 language versions. List of these languages with number of extracted articles and redirects is presented on table 1.

Table 1. Number of articles and redirects in considered language version of Wikipedia.

Code	Language	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 use own parser to extracts six considered metrics. This parser use some of the files from Wikipedia dumps[4]. Below is list of the files that were used by our parser for metrics extraction:

- `{lang}wiki-latest-pages-articles.xml.bz2` – recombine 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 determine articles name that redirects to other articles.
- `{lang}wiki-latest-langlinks.sql.gz` – wiki interlanguage link records. Used for determine name(s) of the article in in other language version(s).

In the above file names `{lang}` means language code[5] of Wikipedia edition. So, for each language version we must download and next process these five compressed files.

To get 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 links are mutual. Additional challenge is 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, based on the similarity analysis, reduced number of articles to 4.2 million interlanguage links sets.

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

In this paper we use Wikipedia dumps from September 2017.

3.3. Building quality measure

As described in [18] often we can observe a positive correlation between the article quality and the value of each of five considered quality metrics (*len*,*img*,*ref*,*hdr*,*ral*). Figure 1 shows how distribution of articles varies depending on metrics value considered by the example from the largest English Wikipedia that it is noticeable if we consider the same number of articles with different quality grades.

---

[4] A complete copy of all Wikimedia wikis, in the form of wikitext source, raw database tables in SQL and metadata embedded in XML - https://dumps.wikimedia.org/

[5] Language codes are described on table 1

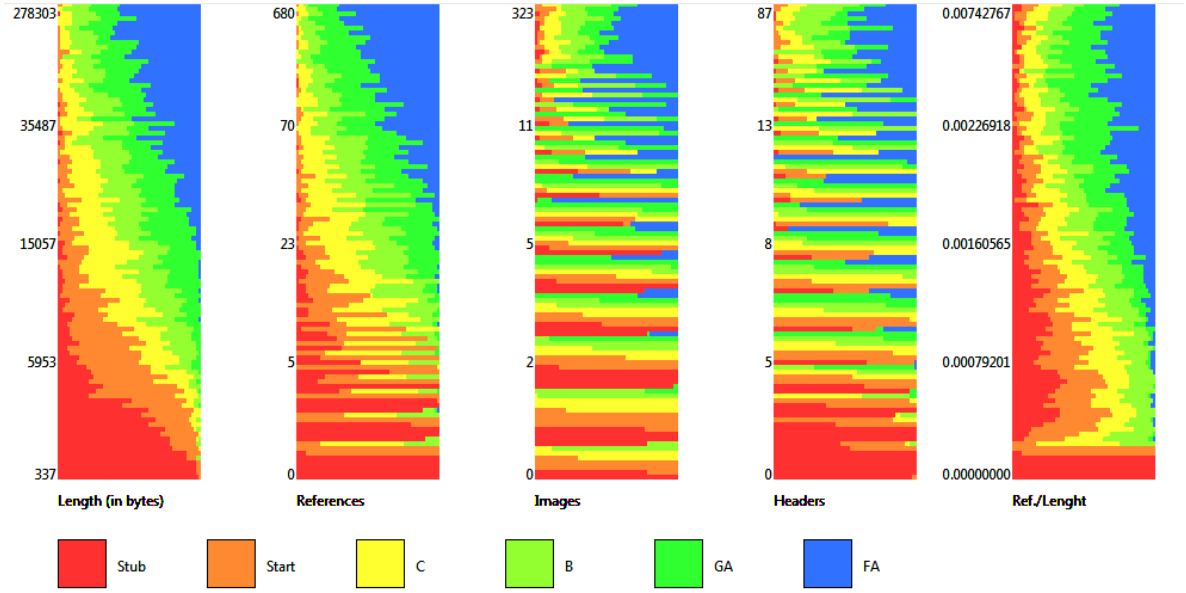

Figure 1. Distribution of metrics in articles of each quality class in English Wikipedia (FA - the highest grade, Stub - the lowest). Source: own calculation

Taking into account the presence of highest FA grade in all language versions of Wikipedia we calculate the median value of these best articles in each language. Medians for each considered metric and language versions are shown in Table 2.

Table 2. Median metrics values in the highest quality class in various Wikipedia languages. Source: own calculation.

Lang.	Length	References	Images	Headers	Ref./Len.
ar	120704.5	162.5	41.5	27.0	0.00133
az	76048.0	124.0	26.0	21.0	0.00162
be	170430.0	197.0	35.0	27.0	0.00113
bg	76416.0	60.0	22.0	21.0	0.00081
ca	47890.0	66.0	18.0	17.0	0.00144
cs	70012.0	123.0	18.0	21.0	0.00196
da	72937.5	125.0	22.0	29.5	0.00196
de	56438.0	55.0	17.0	21.0	0.00095
el	89168.0	83.5	13.0	18.0	0.00094
en	49316.0	113.0	13.0	14.0	0.00231
es	76565.5	99.0	19.0	21.0	0.00133
et	16834.0	27.0	10.0	12.5	0.00203
fa	102343.0	147.5	20.5	22.0	0.00141
fi	49264.0	113.0	15.0	20.0	0.00224
fr	90736.0	167.0	29.0	26.0	0.00186
gl	89990.0	157.0	21.0	22.0	0.00203
he	64263.0	38.0	17.0	19.0	0.0006
hi	74027.5	38.5	18.0	16.0	0.00057
hr	36925.0	25.0	14.0	17.0	0.00073
hu	59459.5	63.0	22.0	21.0	0.00114
hy	157587.0	169.0	38.0	33.0	0.00108
id	49018.0	92.0	14.0	16.0	0.00207
it	82750.0	141.0	29.0	23.0	0.00177
ja	97329.0	188.0	22.0	29.0	0.00198
ka	92822.0	46.0	21.0	20.0	0.00043
ko	72534.0	131.0	20.0	22.0	0.00186
lt	52274.0	44.0	27.0	22.0	0.00056
no	62999.0	77.0	20.0	23.0	0.00108
pl	59967.0	97.0	16.0	17.0	0.00168
pt	70432.5	146.0	23.0	17.0	0.00209
ro	83933.5	154.0	24.0	21.0	0.00197
ru	139812.0	164.0	24.0	22.0	0.00117
sh	55668.0	65.0	15.0	17.0	0.00116
simple	22231.0	51.0	8.0	9.0	0.00227
sl	40176.0	51.5	12.0	16.0	0.00135
sr	112775.0	109.0	29.0	24.0	0.00098
ta	96282.0	24.0	21.0	19.0	0.00017
th	122833.0	91.0	16.0	22.0	0.00088
tr	65254.0	98.0	18.0	17.0	0.00177
uk	84159.0	41.0	25.0	21.0	0.00051
ur	54045.5	31.5	17.5	21.0	0.00058
uz	55387.0	27.5	22.0	26.0	0.00081
vi	89724.0	138.0	21.0	20.0	0.00164
zh	43215.0	91.0	12.0	12.0	0.00219

The above values are then used as thresholds. As proposed in [18], based on the medians we normalize each metric in particular Wikipedia language version according to the following rule: if the value of the given metric in given language exceeds the threshold, it is set to 100 points, otherwise its value is 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 is 118, any article with a higher number of references will score 100 for this metric; article with 59 references will get proportionally 50 points after normalizing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2017 doi:10.20944/preprints201709.0130.v1

Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

Changing value of any metric in particular Wikipedia language version will have different effect on normalized value. For each language version of Wikipedia, each metric can play an important role in assessing the quality, therefore we first count normalized metrics average (NMA) by formula:

$$NMA = \frac{1}{c}\sum_{i=1}^{c} nm_i \tag{1}$$

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

Next we must take into account the number of quality flaw templates QFT in the considered article (if exist) and our final formula for quality measure will look like this:

$$QualityScore = NMA - NMA * 0.05 * QFT \tag{2}$$

In articles with high quality score value (for example 90 points) each quality flaw template will reduce Quality Score by 5% (for 1 such template in our example article will have 85.5 points). This way even if an article has the maximum values of particular metric, but at the same time has quality flaw template(s) - this will not allow to get maximum value of quality score (100).

After assessing of more than 28 million articles in 44 considered language editions of Wikipedia we found that most of the articles get scores between 0 and 30 points. Figure 2 show distribution of articles in this scale[6].

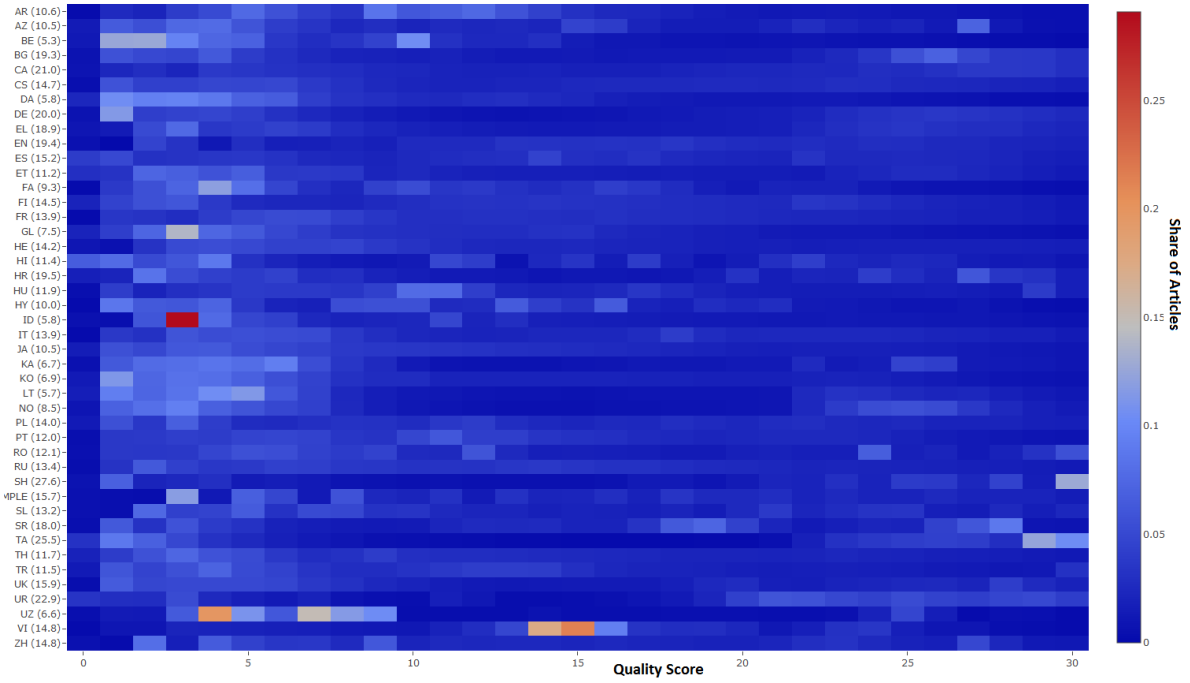

Figure 2. Distribution of articles depending on quality score in each language version of Wikipedia. The medians of the quality scores for each language edition in parentheses. Source: own calculation

4. Popularity measure

The quality of Wikipedia articles can change over time. An article that was once of a high quality can eventually become outdated. Based on the fact that users can edit articles, we can expect that more popular articles will be updated more often (if necessary). Therefore, it is necessary to include in the quality metrics also popularity dimension.

6 More detailed and interactive chart is placed on the webpage: http://data.lewoniewski.info/informatics2017/

Wikipedia records data on visiting their pages in all language versions every hour to special compressed files[7]. In order to measure popularity of articles we have downloaded these data files with statistics for the last year (from September 2016 till 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 visit during the considered period.

In order to calculate the relative popularity we normalize both metrics depending on maximum values of popularity metrics in corresponding articles in other languages. So, for popularity metric p of the particular article with v number of language version, the language v_p^* with the maximum value can be found by the formulas:

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

Now, to calculate the relative popularity RP on a scale from 0 to 100 of selected language version v of the article we count used average of normalized popularity metrics tp and sp:

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

5. Wikipedia articles assessment

In this section we present results of the quality 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 for each language, that allows grouping of articles. Thus, each language version of Wikipedia usually has own structure of categories and own practices concerning their assignment. For example, in some languages it is customary to tag article with more than 20 categories, in others the number can be limited to 2-5 categories. Quality of structure of categories also differ among languages. For example, in some language versions articles about people, events, transport and other topics can be assigned to just one category.

More reliable approach for classification is based on infobox system. 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 others popular knowledge bases (e.g. DBpedia[8]) to develop detailed ontology based on these Wikipedia templates [19]. Popular infoboxes usually have their own names in various languages. For purpose of our research we have chosen 12 different infobox types, based on popularity in English Wikipedia. Using interwiki links we extracted infobox names in other language versions. Table 3 shows that almost all languages of Wikipedia have equivalent of popular infoboxes in English version.

---

7 https://dumps.wikimedia.org/other/analytics/
8 https://dbpedia.org

Table 3. Number of considered language versions of Wikipedia with particular infobox. Source: own calculation.

Infobox name	Num. of lang.
album	41
company	41
film	43
football biography	38
musical artist	40
officeholder	35
person	41
settlement	42
taxobox	43
television	41
university	40
videogame	43

In order to define groups of the articles that describe the same topic we extract lists of article separately for each infobox in 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 infobox at all. However, it is often considered as an important element of article's quality. In such cases we can use interwiki links from identified articles in some languages to reach articles in other version. Results of the above procedure are presented in table 4, which presents number of articles on particular topic in analyzed Wikipedia languages.

Table 4. Number of articles on particular topic in various Wikipedia languages. Source: own calculations

lang.	album	comp.	film	footb.	music.	offic.	person	settl.	taxobox	telev.	univ.	videog.
az	246	540	4692	1759	2042	3773	14818	9886	7755	204	956	218
be	168	589	251	2346	1157	5524	12944	10155	3870	60	436	111
bg	2644	1133	4919	4866	4395	3416	33340	27183	30660	1240	439	203
ca	1396	1483	8729	8082	3488	4130	128844	25239	27938	766	536	1214
cs	6919	3189	5471	10449	12007	3748	58212	18288	12716	1467	547	965
da	2859	2515	12849	6917	6745	3586	24469	6478	6077	1082	611	770
de	8699	23052	33079	37653	10977	6998	97836	33749	45935	6183	3643	3037
el	2020	781	2372	2578	2123	4064	28655	5641	2019	491	239	377
en	161207	67416	123962	149140	105658	142209	559453	513861	337211	43421	22934	22666
es	37487	9504	23071	28571	35908	36945	235382	168487	160470	12323	3927	6920
et	1437	1053	1625	2428	3016	5609	20821	17959	5803	397	507	178
fa	6680	5099	20037	16516	8979	12981	83842	150348	25017	3044	1462	1719
fi	22230	6432	10382	9950	15094	10948	79126	22885	19758	3700	950	3666
fr	42030	21845	51157	43026	39090	41593	278194	217022	111845	11041	5201	12364
gl	3883	1146	2458	2172	4080	4229	21557	11170	5087	611	222	498
he	4928	2552	4532	5310	6389	9351	41876	10703	7155	2421	603	654
hi	908	872	4307	62	626	5140	7829	7100	1333	563	623	69
hr	4875	1022	1991	3232	3593	4306	21691	25491	5127	531	200	345
hu	10453	2353	5980	16524	7723	10396	56162	101132	21410	2998	253	1076
hy	2874	855	3196	2473	3970	3286	22987	76528	3216	630	440	149
id	8567	4600	10519	13226	5360	12009	39419	93622	96843	4518	1561	673
it	71368	13114	60999	50138	31082	34395	331480	183633	37408	10590	1705	8790
ja	28375	31715	19029	16874	26501	16449	100936	43253	15758	4832	2917	8696
ka	4634	602	1690	1655	2111	5172	14554	30792	10582	384	248	204
ko	7234	7510	10446	11209	11703	9015	54498	24350	14142	7389	1721	2646
lt	2273	1387	2129	2644	3400	3974	14870	21297	9309	249	453	507
no	11565	5460	6822	10836	11341	14458	136405	36224	28405	1484	1742	1237
pl	30606	8185	19506	39589	20363	30018	172777	230483	42047	5972	2605	3372
pt	36065	9453	24044	26859	25360	18047	143961	153436	100580	10123	2232	5501
ro	4452	2593	4390	4743	5321	8085	36072	157473	32008	1417	484	763
ru	22059	12940	28386	32145	30248	48437	199057	244567	40238	5515	3229	6251
sh	1735	657	8898	2261	2445	6521	25881	119863	2578	1268	623	101
simple	3605	1488	2689	5713	5281	6061	28633	25203	4350	1258	836	968
sl	1536	965	730	2252	2194	3363	31796	27712	2441	217	274	140
sr	2571	1068	5621	2508	3105	5633	27707	102211	9605	1170	335	315
ta	1254	831	4960	166	652	2972	7537	7371	2501	239	724	34
th	2714	1439	2663	1739	3662	5126	13911	5535	5918	2687	687	796
tr	9641	3689	7294	17988	8510	10058	53068	57582	6127	2714	993	1318
uk	9880	6031	13967	15391	9170	18095	82276	176111	24649	1626	1817	1656
ur	143	467	322	69	384	1826	7717	64090	611	99	533	42
uz	90	184	132	177	567	1034	3427	71794	1024	43	177	15
vi	4231	1915	2706	2694	3297	6014	19693	201490	796749	2278	648	1038
zh	11059	11075	10129	11571	7663	19167	68975	148416	97553	11040	4669	4477

Table 5 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 5 (please observe the logarithmic scale on vertical axis).

Table 5. Number of articles which have a certain number of language versions in particular topics. Source: own calculations

langs	album	comp.	film	footb.	music.	offic.	person	settl.	taxobox	telev.	univ.	videog.
1	263745	129613	226052	187770	164297	198470	912559	842467	1139504	74482	31834	31437
2	90299	39045	92473	103279	74016	84295	527270	574990	425054	26976	12160	17700
3	54343	22640	53076	66452	47037	52638	368955	386807	180521	15538	6996	12385
4	38451	15753	36475	48100	34145	38797	278537	296388	100373	10656	4512	9059
5	28929	11951	27227	37441	26617	30934	219460	235161	70777	8009	3380	6962
6	22404	9498	21456	30388	21390	25496	176480	196752	53416	6199	2670	5514
7	17602	7800	17572	25042	17693	21374	143829	167865	42560	4930	2167	4496
8	14296	6482	14636	20365	14876	18262	118437	149570	34636	4021	1806	3687
9	11774	5504	12344	16987	12610	15663	98543	131890	28594	3324	1526	3063
10	9652	4696	10617	14315	10793	13625	82757	118162	23935	2808	1310	2536
11	7957	4017	9228	12065	9337	11884	70551	106233	20298	2379	1129	2094
12	6647	3481	8060	10178	8121	10491	60805	92093	17250	1999	994	1774
13	5579	3014	7056	8846	7116	9277	52603	76816	14855	1664	876	1513
14	4725	2635	6177	7746	6225	8303	46105	61660	12745	1428	779	1309
15	3973	2310	5411	6862	5525	7467	40671	48624	11085	1232	673	1112
16	3357	1994	4729	6154	4925	6725	36019	37011	9704	1051	599	935
17	2840	1753	4133	5507	4366	6103	32020	30276	8514	893	523	797
18	2398	1567	3642	4956	3861	5536	28548	25539	7543	763	470	669
19	2040	1393	3184	4486	3427	4996	25570	22281	6694	646	430	575
20	1734	1256	2787	4057	3027	4554	22873	19520	5951	543	381	492
21	1449	1130	2445	3656	2658	4181	20572	16850	5293	463	340	421
22	1250	995	2164	3276	2385	3841	18622	13717	4740	406	300	362
23	1049	879	1907	2971	2126	3503	16878	10561	4248	361	261	298
24	899	775	1662	2641	1895	3236	15245	8476	3801	309	240	249
25	744	690	1462	2371	1676	2955	13780	7031	3405	268	218	206
26	617	618	1282	2145	1500	2728	12553	6103	3030	237	191	171
27	504	559	1116	1942	1340	2498	11406	5407	2681	199	166	134
28	393	483	970	1745	1193	2289	10371	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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2017 doi:10.20944/preprints201709.0130.v1

Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

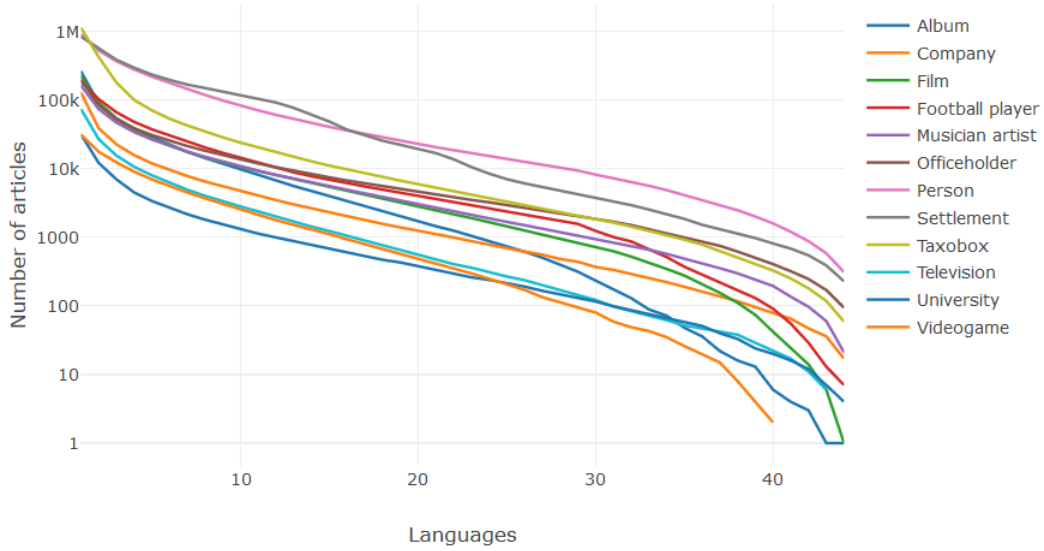

Figure 3. Number of articles which have a certain number of language versions in particular topics. Source: own calculations

Another possibility to analyze the data about language versions from table 4 is to show overlaps between a group of three languages using Venn diagrams. They show how many articles specific languages have in common (see figure 4).

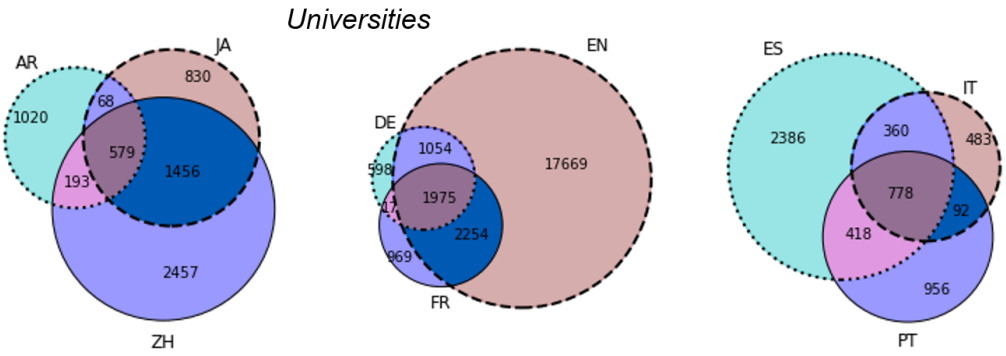

Figure 4. Coverage of articles that describe universities in different languages. Source: own calculation. Other interactive Venn diagrams for this paper with different topics and languages are available on the web page: http://data.lewoniewski.info/informatics2017/vn/

5.2. Quality assessment

For all articles from our dataset we have calculated synthetic measure of quality described in section 3.3. Table 6 presents the average quality scores of articles for each topic in 44 Wikipedia language editions.

Table 6. Average quality scores of articles for each topic in 44 Wikipedia language editions. Source: own calculations

langs	album	comp.	film	footb.	music.	offic.	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

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 distribution of quality scores for three Wikipedia language versions (EN, GE, FR) in 12 considered topics.[9]

[9] Charts for other languages are available from the webpage: http://data.lewoniewski.info/informatics2017/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2017 doi:10.20944/preprints201709.0130.v1

Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

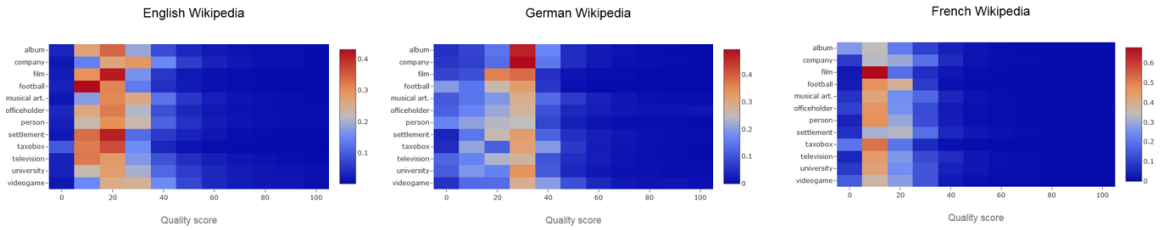

Figure 5. Distribution of quality scores for three Wikipedia language versions (EN, GE, FR) in 12 considered topics. Source: own calculation

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 7 we present the average popularity metric tp for articles in each topic in 44 Wikipedia language editions.

Table 7. Average popularity metric tp in articles for each topic in 44 Wikipedia language editions. Source: own calculations

langs	album	comp.	film	footb.	music.	offic.	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	11532.1	12524.1	6267.5	4579.8	2929.9	15321.5	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	11111.2	14451.0	16943.0	3250.7	18625.7	9016.0	14687.2	2491.9	2235.3	26019.4	7132.1	21296.7
es	3495.3	7508.5	7622.6	3110.3	7905.3	5143.7	4634.8	1369.0	1014.3	10001.4	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	10135.9	2112.0	11154.6	7079.9	7882.0	2509.4	6141.3	25687.4	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	12011.1	2904.5	8646.7	5561.7	6182.5	1464.4	3507.1	21073.4	4641.4	17428.5
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	13477.2	600.2	17115.2	4524.9	5499.5	361.8	495.2	17052.6	3535.2	8218.4

6. Association between quality and popularity

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

As there were additional requirements for relations between languages, we have conducted the analysis on the subset of Wikipedia articles. We selected only those articles in each topic that had at least three language versions (cf. table 5). We further analyze combinations of a language and a topic – a pair. Table 8 presents the top 25 pairs with share of articles, which have the highest quality in comparison to other language (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 are best described in English version.

Table 8. Top 25 pairs language—topic with share of articles, which have the highest quality in comparison to other language (articles with at least 3 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%

Analogous table was prepared for popularity. Table 9 presents the top 25 pairs language—topic with share of articles that attracts the highest popularity in comparison to other languages (full data is presented in table A2 in the appendix). Similarly to previous table, the first row of this table should be interpreted as follows: regarding the topic 'album' 85.8% of articles have English version as the most popular (attracts the biggest number of visits).

Table 9. Top 25 language versions and topics with share of articles, which have the highest popularity in comparison to other language. (articles with at least 3 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%

The goal of our research was 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 parametric test using Phi coefficient, calculated for each pair language—topic. It is a measure of association for two binary variables. Our variables are coded as follows: if an article about a specific topic in a given language is of the highest quality among all languages then it is assigned value 1 (high score), otherwise 0 (low score). Similarly for popularity: if an article about a specific topic in a given language is the most popular among all languages then it is assigned value 1 (high score), otherwise 0 (low score).

Then, Phi coefficient was calculated by formula:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}} \tag{5}$$

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

Depending on language and topic correlation may differ significantly. Table 10 shows top 25 pairs language—topic with the highest correlation coefficients (full data is presented in table A3 in the appendix).

Table 10. Top 25 language versions and topics with the highest phi coefficients between articles with the highest quality and popularity. (articles with at least 3 language versions were considered). Source: own calculations

Lang - topic	Correlation coeff.
TH - university	.838
TH - officeholder	.762
VI - university	.719
PL - university	.717
PT - university	.707
ID - university	.705
TH - musical artist	.684
ES - university	.683
TR - university	.677
EN - university	.676
ID - settlement	.656
FA - television	.655
ET - television	.65
SL - film	.642
CS - university	.64
JA - company	.637
FR - university	.636
PT - television	.632
EN - television	.63
EN - company	.625
ZH - officeholder	.615
JA - university	.614
VI - musical artist	.607
BG - university	.603
BG - television	.602

The problem with Phi coefficient, which is a special case of Pearson correlation coefficient, is that results have high granularity and it cannot be easily generalized. Therefore we have also set up another experiment, in which we estimated the association between quality and popularity within a topic. For every topic we have prepared two lists of languages: one ordered by the share of articles that are of highest quality (see table A1) and the other ordered by the share of articles that are the most popular (see table A2). Those list are effectively ranks. We wanted to know if the order of languages is similar, which would support the hypothesis that quality and popularity are associated. For the purpose we used Spearman's rank correlation coefficient between shares of articles. Results are presented on table 11.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2017 doi:10.20944/preprints201709.0130.v1

Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

Table 11. Spearman's rank correlation coefficients for shares of the articles of the highest quality and the most popular ones, in various topics. Source: own calculation using [20].

Topic	Spearman's rank cor. coef.	2-sided p-value
album	0.7227	3.05e-08
company	0.8749	8.29e-15
film	0.6408	2.80e-06
football biography	0.7872	2.33e-10
musical artist	0.8453	5.27e-13
officeholder	0.7665	1.32e-09
person	0.8370	1.45e-12
settlement	0.6146	9.09e-06
taxobox	0.6997	1.26e-07
television	0.7950	1.15e-10
university	0.8362	1.60e-12
videogame	0.7436	7.35e-09

Spearman's rank correlation assesses the strength of a link between two sets of considered data, which in our case can reach 0.87 (for the topic 'company'). The results shows that depending on topic we can have different correlation between quality and popularity, but no less then 0.61 (for the topic 'settlement'). All associations are statistically significant (as shown by p-values). Overall, results of our calculations support the hypothesis that there is an association between articles of high quality with their popularity.

7. Conclusions and future work

Proposed methods can help in assessing quality and popularity of Wikipedia articles in different languages. Presented approach takes into account the specifics of the best articles of each language version of Wikipedia to calculate quality score. By considering the popularity measure we can improve the process of identification of language versions with the highest quality. The proposed quality metrics can be helpful particularly in automated knowledge extraction from Wikipedia articles. One of such solutions is DBpedia. The problem that is often encountered is a conflict resolution, which occurs when various language versions concerning the same subject have conflicting information [21]. 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.

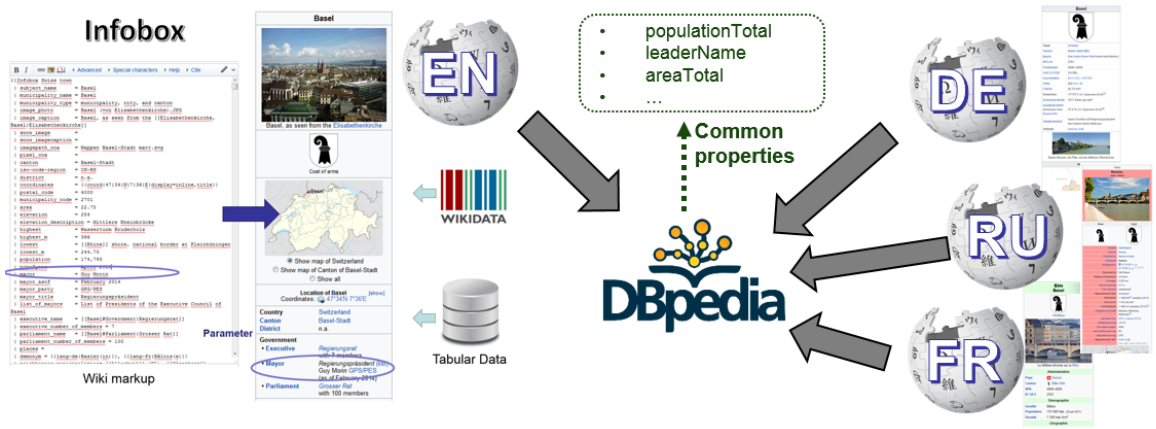

Figure 6. Infobox about Basel with its data sources and its extraction to DBpedia from different Wikipedia language versions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 September 2017 doi:10.20944/preprints201709.0130.v1

Peer-reviewed version available at Informatics 2017, 4, ; doi:10.3390/informatics4040043

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

Language	Quality score	Popularity score
DE	93.3	68.0
EN	90.5	100
FR	70.0	30.9
PL	52.9	10.9
RU	31.5	14.3
ES	45.9	19.6
IT	52.6	15.5
...	...	...

Figure 7. Scheme of information enrichment of Wikipedia infobox based on quality and popularity assessment of other language versions on an example of a Basel city. Source: own calculations in September, 2017.

The methods proposed in the paper are used in WikiRank.net service[10], which assesses and compare articles in the various language versions of Wikipedia.

In the future we plan to continue study new metrics and their extraction methods for improving the model. Some of the metrics can be expanded. For example, by analyzing similarity of sources in Wikipedia articles across languages we can also evaluate the quality of its content [22]. Furthermore, the references themselves can also have their own quality metrics (e.g. impact factor), which can be used as an indirect indicator of articles quality.

10 http://wikirank.net

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 3 language versions were considered). Source: own calculations

langs	album	comp.	film	footb.	music.	offic.	person	settl.	taxobox	telev.	univ.	videog.
ar	.0	.002	.001	.008	.001	.003	.007	.001	.004	.001	.007	.004
az	.0	.0	.001	.001	.001	.004	.001	.006	.001	.0	.004	.0
be	.0	.0	.0	.003	.0	.001	.0	.0	.0	.0	.0	.0
bg	.003	.001	.006	.003	.004	.003	.004	.01	.03	.002	.005	.0
ca	.001	.003	.03	.012	.004	.004	.04	.005	.033	.004	.004	.008
cs	.003	.005	.002	.016	.009	.002	.006	.003	.006	.002	.002	.002
da	.001	.002	.001	.003	.002	.001	.0	.0	.0	.001	.0	.0
de	.032	.133	.147	.091	.035	.013	.038	.015	.08	.036	.071	.047
el	.005	.002	.003	.005	.004	.006	.006	.002	.002	.002	.002	.002
en	.555	.497	.437	.291	.49	.393	.387	.212	.271	.478	.435	.605
es	.058	.027	.017	.053	.057	.062	.062	.046	.139	.073	.039	.025
et	.001	.004	.001	.002	.003	.007	.004	.003	.003	.001	.004	.0
fa	.001	.002	.002	.003	.002	.008	.003	.012	.0	.002	.005	.001
fi	.018	.013	.014	.007	.012	.009	.012	.005	.009	.005	.006	.006
fr	.032	.05	.06	.052	.067	.065	.077	.115	.026	.038	.06	.051
gl	.0	.001	.0	.0	.001	.001	.001	.001	.002	.0	.001	.0
he	.002	.006	.005	.003	.008	.014	.011	.002	.001	.006	.009	.001
hi	.001	.002	.006	.0	.001	.002	.001	.002	.0	.002	.002	.0
hr	.008	.005	.007	.007	.008	.01	.007	.016	.002	.004	.003	.001
hu	.011	.006	.008	.015	.01	.013	.01	.015	.009	.011	.003	.005
hy	.0	.0	.001	.0	.001	.002	.001	.019	.0	.0	.0	.0
id	.004	.004	.007	.002	.005	.005	.003	.001	.002	.007	.007	.001
it	.063	.03	.066	.181	.051	.059	.1	.031	.042	.057	.014	.027
ja	.011	.046	.017	.036	.035	.019	.018	.004	.006	.043	.047	.053
ka	.01	.001	.001	.001	.001	.002	.001	.003	.01	.0	.003	.001
ko	.003	.007	.001	.003	.005	.005	.003	.001	.002	.011	.013	.004
lt	.001	.002	.001	.003	.003	.006	.002	.004	.001	.0	.001	.0
no	.006	.008	.011	.011	.012	.014	.033	.011	.007	.006	.006	.004
pl	.038	.016	.018	.027	.038	.06	.04	.088	.034	.024	.025	.015
pt	.043	.015	.027	.024	.02	.018	.019	.028	.033	.035	.013	.018
ro	.003	.003	.002	.004	.003	.004	.003	.017	.029	.002	.003	.001
ru	.034	.032	.031	.057	.04	.078	.04	.039	.026	.019	.04	.045
sh	.0	.001	.005	.0	.001	.005	.002	.036	.0	.001	.002	.0
simple	.002	.002	.002	.006	.004	.006	.004	.004	.001	.003	.004	.002
sl	.003	.001	.001	.0	.003	.004	.003	.017	.0	.0	.0	.001
sr	.001	.001	.001	.001	.002	.003	.002	.002	.013	.002	.001	.001
ta	.0	.002	.001	.0	.001	.005	.002	.002	.001	.001	.003	.0
th	.001	.001	.001	.001	.002	.003	.001	.001	.002	.002	.008	.002
tr	.002	.003	.002	.011	.005	.007	.004	.006	.002	.004	.008	.002
uk	.019	.019	.021	.041	.017	.037	.02	.151	.011	.007	.023	.01
ur	.0	.0	.0	.0	.0	.001	.001	.037	.0	.0	.002	.0
uz	.0	.0	.0	.0	.0	.001	.0	.001	.0	.0	.0	.0
vi	.001	.001	.002	.001	.002	.002	.001	.017	.135	.004	.004	.002
zh	.022	.044	.032	.012	.029	.032	.019	.009	.025	.105	.109	.052

Appendix A.2

Table A2. Shares of Wikipedia articles with the highest popularity compared with other language versions (articles with at least 3 language versions were considered). Source: own calculations

langs	album	comp.	film	footb.	music.	offic.	person	settl.	taxobox	telev.	univ.	videog.
ar	.0	.001	.002	.004	.003	.003	.006	.0	.001	.001	.01	.0
az	.0	.0	.0	.0	.001	.002	.0	.0	.0	.0	.004	.0
be	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
bg	.0	.0	.001	.001	.002	.003	.001	.002	.0	.001	.003	.0
ca	.0	.0	.0	.0	.0	.0	.001	.0	.0	.0	.0	.0
cs	.0	.003	.003	.006	.006	.001	.004	.001	.001	.001	.003	.0
da	.0	.001	.001	.001	.002	.003	.0	.0	.0	.001	.001	.0
de	.005	.06	.032	.041	.008	.004	.017	.008	.033	.023	.032	.004
el	.0	.001	.001	.002	.001	.005	.002	.002	.0	.001	.001	.0
en	.858	.663	.733	.552	.676	.544	.624	.391	.736	.644	.508	.857
es	.019	.024	.02	.089	.046	.069	.065	.052	.061	.095	.051	.007
et	.0	.001	.0	.0	.001	.002	.001	.004	.0	.0	.001	.0
fa	.0	.001	.002	.004	.002	.006	.004	.005	.0	.001	.008	.0
fi	.003	.004	.001	.001	.006	.005	.003	.002	.001	.001	.002	.0
fr	.011	.032	.043	.035	.029	.037	.049	.113	.033	.022	.051	.005
gl	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
he	.0	.002	.001	.002	.003	.004	.002	.003	.001	.001	.003	.0
hi	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
hr	.002	.001	.001	.0	.004	.004	.002	.011	.0	.002	.001	.0
hu	.0	.001	.001	.005	.003	.006	.004	.019	.001	.001	.003	.0
hy	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.001	.0
id	.0	.002	.0	.0	.003	.004	.001	.002	.008	.0	.005	.0
it	.011	.018	.037	.035	.019	.027	.038	.024	.006	.012	.012	.004
ja	.058	.077	.035	.06	.06	.022	.029	.007	.014	.08	.093	.075
ka	.0	.0	.0	.0	.0	.002	.0	.0	.0	.0	.0	.0
ko	.0	.006	.0	.002	.002	.004	.002	.001	.001	.006	.011	.0
lt	.0	.001	.0	.0	.001	.002	.001	.001	.0	.0	.002	.0
no	.0	.002	.001	.003	.003	.003	.004	.003	.001	.001	.001	.0
pl	.003	.01	.006	.015	.011	.032	.018	.075	.015	.005	.016	.001
pt	.007	.012	.004	.03	.014	.012	.013	.025	.008	.019	.017	.002
ro	.0	.001	.001	.0	.003	.005	.002	.035	.001	.0	.002	.0
ru	.013	.05	.051	.096	.059	.142	.084	.169	.052	.029	.076	.036
sh	.0	.0	.0	.0	.0	.0	.0	.001	.0	.0	.0	.0
simple	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
sl	.0	.0	.0	.0	.001	.001	.001	.004	.0	.0	.0	.0
sr	.001	.0	.004	.0	.003	.005	.003	.014	.0	.002	.005	.0
ta	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
th	.0	.002	.001	.001	.002	.002	.001	.003	.002	.001	.008	.0
tr	.001	.004	.003	.01	.006	.011	.005	.012	.0	.006	.012	.0
uk	.0	.001	.0	.0	.001	.002	.001	.007	.001	.0	.009	.0
ur	.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
vi	.0	.001	.0	.0	.002	.002	.001	.002	.007	.001	.006	.0
zh	.006	.017	.013	.002	.015	.025	.01	.002	.015	.042	.042	.007

Appendix A.3

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

langs	album	comp.	film	footb.	music.	offic.	person	settl.	taxobox	telev.	univ.	videog.
ar	.334	.199	.243	.086	.380	.177	.241	.104	.274	.457	.390	-
az	-	-.006	.463	.101	.350	.263	.226	.063	.158	-.005	.458	-.010
be	.304	.263	-	-	.269	.099	.161	.028	.349	-	-	-
bg	.140	.367	.223	.097	.534	.488	.244	.358	.139	.602	.603	-
ca	.180	.051	.041	.186	.073	.033	.170	.167	.053	.288	.181	.088
cs	.211	.455	.331	.279	.405	.400	.406	.209	.178	.580	.640	.269
da	.427	.372	.358	.225	.419	.461	.162	.289	.182	.477	.411	.405
de	.334	.523	.338	.364	.263	.341	.309	.512	.207	.535	.593	.215
el	.196	.303	.564	.246	.238	.511	.335	.026	.066	.522	.134	-
en	.636	.625	.445	.434	.576	.565	.390	.454	.363	.630	.676	.589
es	.262	.466	.295	.337	.436	.567	.315	.642	.419	.543	.683	.185
et	.207	.332	.189	.151	.461	.362	.283	.211	.157	.650	.359	-
fa	.298	.402	.386	.220	.595	.455	.458	.161	.141	.655	.601	.181
fi	.325	.443	.254	.250	.456	.557	.299	.386	.166	.511	.324	.176
fr	.238	.430	.237	.263	.343	.362	.330	.370	.204	.457	.636	.205
gl	-.001	.103	.065	-.005	.024	.083	.247	.082	.102	.367	-	-
he	.343	.504	.281	.525	.459	.410	.325	.256	.179	.586	.552	-
hi	-	-	.074	-	-.015	-.006	.074	.015	.200	-.027	.026	-
hr	.262	.421	.257	.402	.593	.400	.352	.409	.341	.443	.456	-
hu	.103	.467	.252	.417	.425	.458	.419	.546	.093	.369	.537	.094
hy	.163	-.004	.198	-.002	.220	.245	.124	-.003	-.007	.168	.341	-
id	.249	.362	.190	.212	.435	.582	.322	.656	.033	.466	.705	-
it	.248	.438	.137	.344	.337	.440	.313	.466	.283	.254	.536	.159
ja	.309	.637	.455	.594	.562	.482	.492	.447	.342	.532	.614	.379
ka	.052	-	.268	.020	.176	.201	.195	.501	-.001	.111	.214	-
ko	.213	.454	.190	.319	.369	.567	.417	.552	.176	.426	.511	.150
lt	.133	.247	.232	.088	.353	.396	.369	.541	.235	.318	.558	-
no	.149	.338	.191	.036	.253	.353	.155	.170	.536	.351	.342	.227
pl	.319	.607	.256	.292	.414	.424	.339	.524	.134	.304	.717	.160
pt	.334	.474	.179	.444	.496	.486	.367	.553	.064	.632	.707	.252
ro	.129	.445	.441	.101	.498	.370	.398	.113	.435	.259	.473	-
ru	.268	.460	.392	.365	.414	.456	.330	.421	.360	.470	.607	.296
sh	-	-.007	.075	-.001	.067	.027	.175	.011	-.004	-.005	.495	-
simple	.098	.091	.137	-.007	.049	-	.071	.077	.074	-	-	-
sl	.214	.436	.642	-.005	.329	.230	.329	.113	.140	-.026	.597	-
sr	.122	.189	.305	.045	.330	.251	.301	.113	.081	.398	.556	-
ta	-.012	-.013	.010	-	-	.038	.039	.033	.212	-	.082	-
th	.196	.440	.407	.226	.684	.762	.481	.495	.151	.544	.838	.173
tr	.451	.367	.453	.366	.507	.397	.405	.507	.290	.560	.677	.163
uk	.104	.188	.083	.110	.206	.216	.201	.204	.091	.184	.391	-
ur	-	-	-	-	-	.104	.044	.013	-	-	-	-
uz	-	-	-	-	.236	.314	.201	.073	.338	-	-	-
vi	.121	.302	.261	.420	.607	.481	.369	.142	.372	.209	.719	.096
zh	.363	.423	.378	.250	.456	.615	.410	.185	.126	.550	.462	.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. *Universal Journal of Educational Research* **2016**, *4*, 349–354.
2. Blumenstock, J.E. Size matters: word count as a measure of quality on wikipedia. WWW, 2008, pp. 1095–1096.
3. Warncke-wang, M.; Cosley, D.; Riedl, J. Tell Me More : An Actionable Quality Model for Wikipedia. WikiSym 2013, 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.; Springer International Publishing, 2015; Vol. 228, *Lecture Notes in Business Information Processing*, 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, 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. *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12* **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, 2017; pp. 28–40.
8. Lipka, N.; Stein, B. Identifying Featured Articles in Wikipedia: Writing Style Matters. *Proceedings of the 19th International Conference on World Wide Web (2010)* **2010**, pp. 1147–1148.
9. Xu, Y.; Luo, T. Measuring article quality in Wikipedia: Lexical clue model. *IEEE Symposium on Web Society* **2011**, pp. 141–146.
10. Anderka, M. Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany, 2013.
11. Wu, G.; Harrigan, M.; Cunningham, P. Characterizing Wikipedia Pages Using Edit Network Motif Profiles. Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents; ACM: New York, NY, USA, 2011; SMUC '11, pp. 45–52.
12. Suzuki, Y.; Nakamura, S. Assessing the Quality of Wikipedia Editors Through Crowdsourcing. Proceedings of the 25th International Conference Companion on World Wide Web; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, Switzerland, 2016; WWW '16 Companion, pp. 1001–1006.
13. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international symposium on Wikis WikiSym 07* **2007**, pp. 157–164.
14. Ingawale, M.; Dutta, A.; Roy, R.; Seetharaman, P. Network analysis of user generated content quality in Wikipedia. *Online Information Review* **2013**, *37*, 602–619.
15. Halfaker, A.; Taraborelli, D. Artificial intelligence service "ORES" gives Wikipedians X-ray specs to see through bad edits **2015**.
16. Dalip, D.H.; Lima, H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Quality assessment of collaborative content with minimal information. IEEE/ACM Joint Conference on Digital Libraries, 2014, pp. 201–210.
17. Dang, Q.V.; Ignat, C.L. Quality assessment of Wikipedia articles without feature engineering. 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, pp. 27–30.
18. Lewoniewski, W.; Węcel, K. Relative quality assessment of Wikipedia articles in different languages using synthetic measure. 20th International Conference on Business Information Systems. (in press), 2017.
19. Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia-A crystallization point for the Web of Data. *Web Semantics: science, services and agents on the world wide web* **2009**, *7*, 154–165.
20. Wessa, R. Spearman Rank Correlation (v1.0.3) in Free Statistics Software (v1.2.1), Office for Research Development and Education. https://www.wessa.net/rwasp_spearman.wasp/. Accessed: 2017.
21. Bryl, V.; Bizer, C. Learning conflict resolution strategies for cross-language wikipedia data fusion. Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014, pp. 1129–1134.
22. 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, 2017; pp. 561–573.