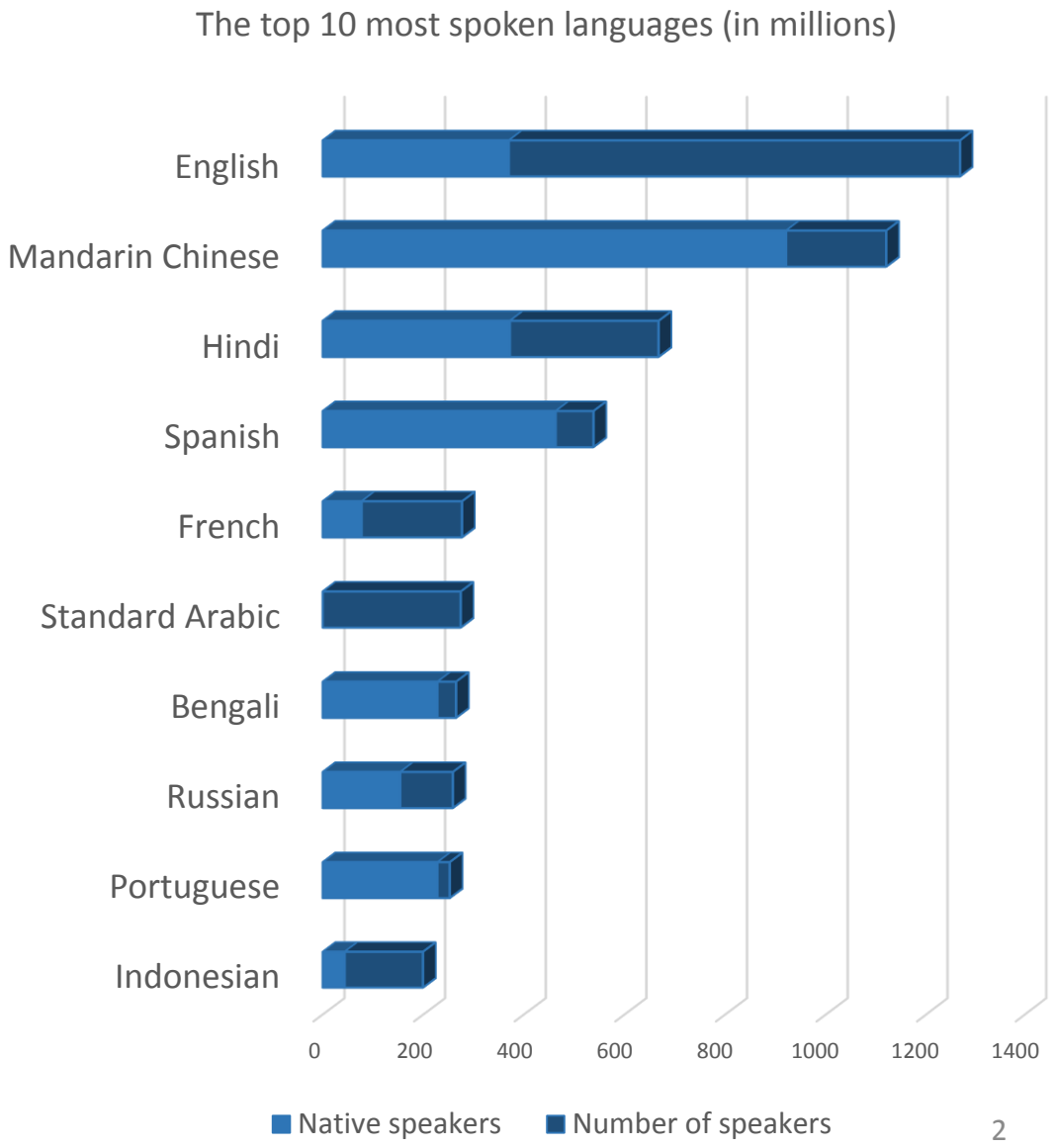
Quality assessment of Wikipedia and its sources

Dr. Włodzimierz Lewoniewski

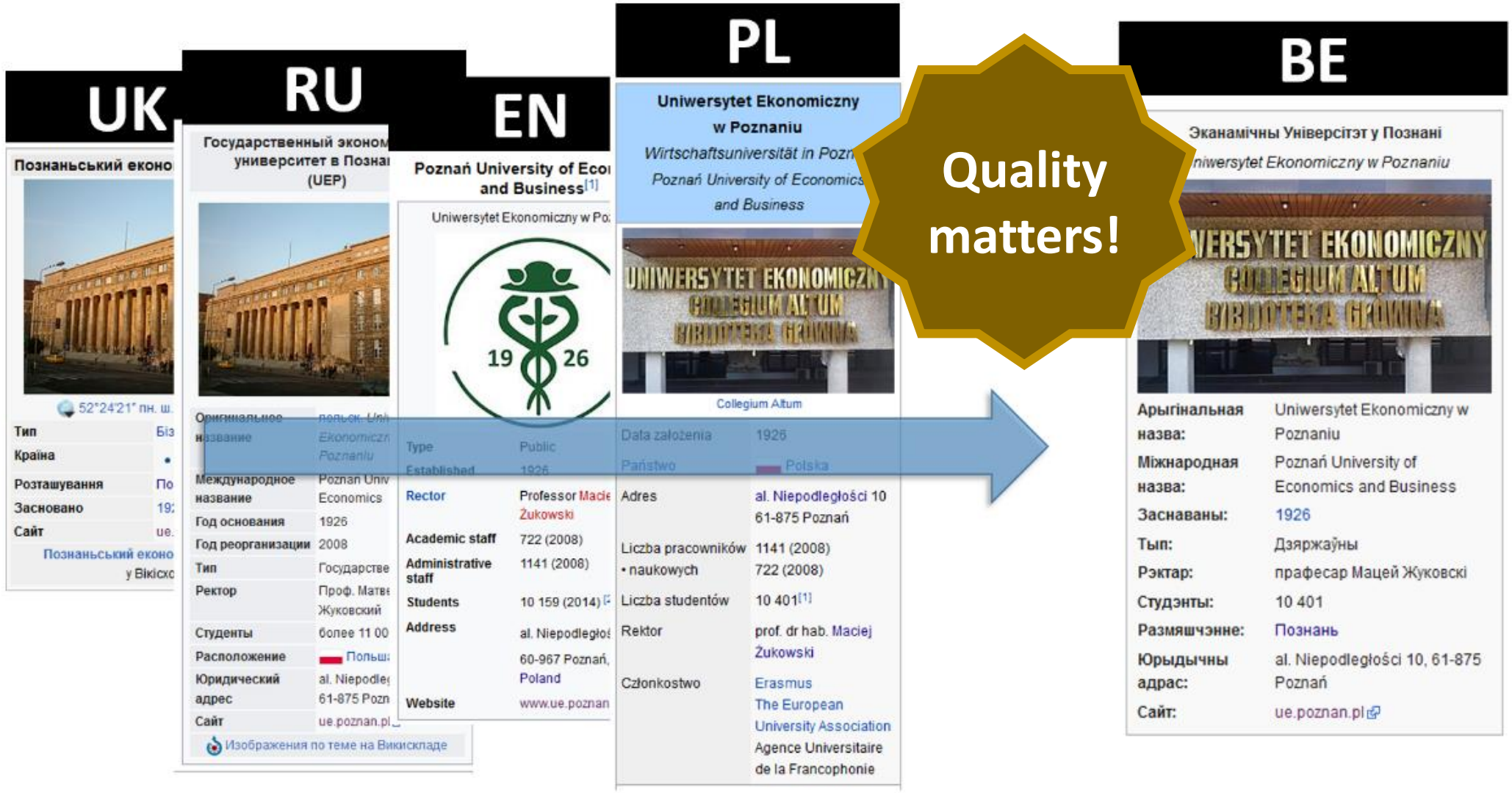
Languages of the world in 2020

- 7,117 languages are spoken.
- 2,926 languages are endangered.
- just 23 languages account for more than ½ the world's population
- Wikipedia articles have been created in 314 languages



Source: ethnologue.com, meta.wikimedia.org

Motivation – enrichment of multilingual information



Source: Lewoniewski, W. (2018). [The method of comparing and enriching information in multilingual wikis based on the analysis of their quality](). PhD thesis

Quality in Multilingual Wikipedia

- Wikipedia can be edited in each language independently
  - same subject can be described differently
  - user usually needs to understand those languages
- Information quality depends on language of Wikipedia
  - Each language defines own rules and standards
  - Standards may change over time
- Reliable sources are important
  - Assessment of the same source depends on language edition of Wikipedia
  - Reliablity of the same source may change over the time

Related works

- Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. *Information, 11(5), 263*
- Lewoniewski, W., Węcel, K., Abramowicz, W. (2019). Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. *Computers, 8(3), 60.*
- Lewoniewski, W. (2019). Measures for quality assessment of articles and infoboxes in multilingual Wikipedia. *In International Conference on Business Information Systems (pp. 619-633). Springer, Cham.*
- Lewoniewski, W. (2018). The method of comparing and enriching information in multilingual wikis based on the analysis of their quality. *PhD thesis*
- Lewoniewski, W., Węcel, K., Abramowicz, W. (2017). Relative quality and popularity evaluation of multilingual Wikipedia articles. *Informatics 2017, 4(4), 43.*
- Lewoniewski, W. (2017). Enrichment of information in multilingual Wikipedia based on quality analysis. *In International Conference on Business Information Systems (pp. 216-227). Springer, Cham.*
- Lewoniewski, W., Węcel, K., Abramowicz, W. (2017). Analysis of references across Wikipedia languages. *In International Conference on Information and Software Technologies (pp. 561-573). Springer, Cham.*

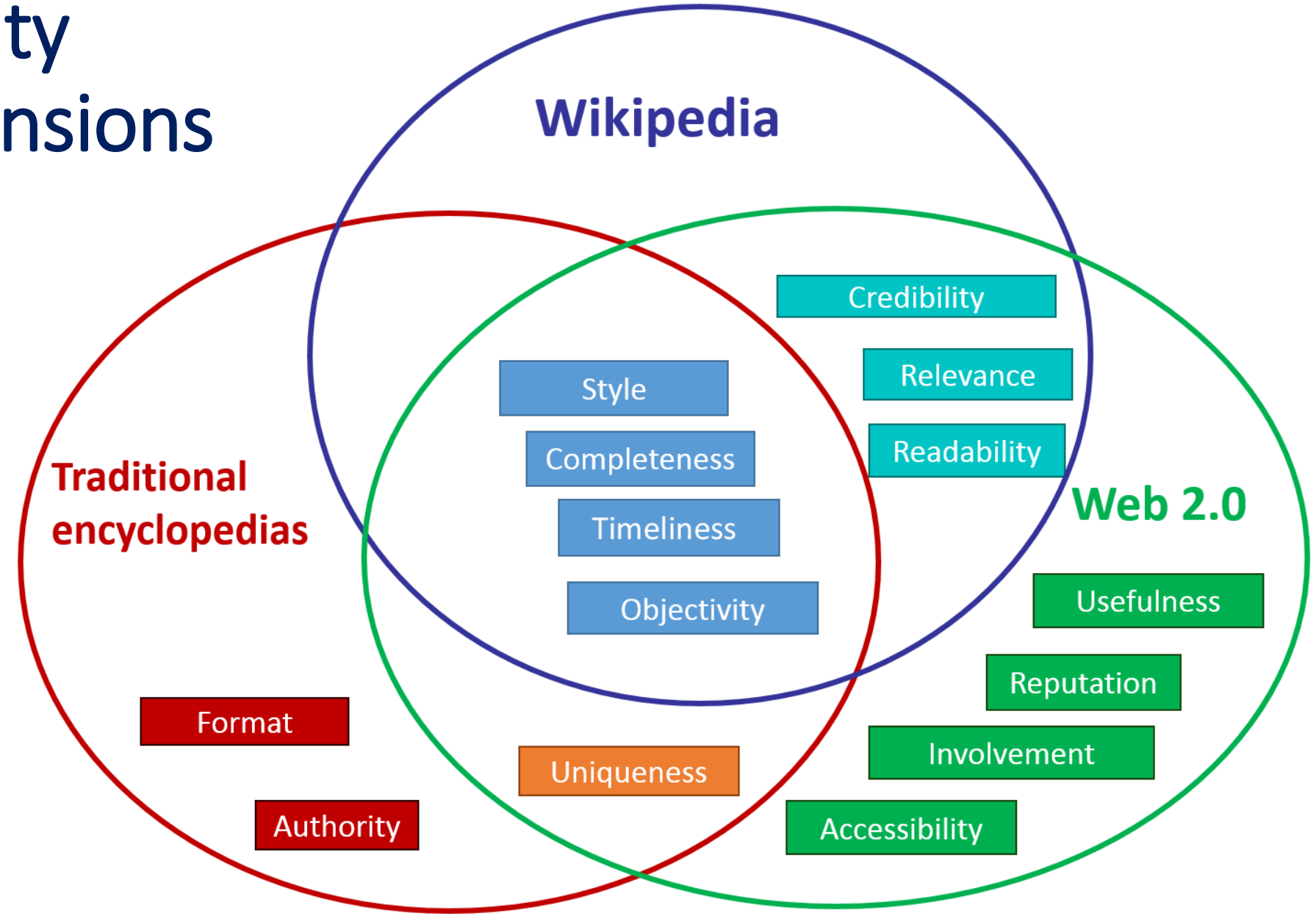
Quality classes in Wikipedia languages

Grade / Language	BE	DE	EN	FR	PL	RU	UK
Featured Article (FA)	X	X	X	X	X	X	X
Good Article (GA)	X	X	X	X	X	X	X
Solid Article						X	
A-class			X	X			
Four					X		
Full						X	X
B-class			X	X			
Developed						X	X
C-class			X				
In develpment						X	X
Start			X	X	X		
Stub	X		X	X	X	X	X
Unassessed	**99,34%**	**99,68%**	**10,16%**	**39,30%**	**99,50%**	**85,01%**	**97,04%**

Colors are marked grades that have similar characteristics

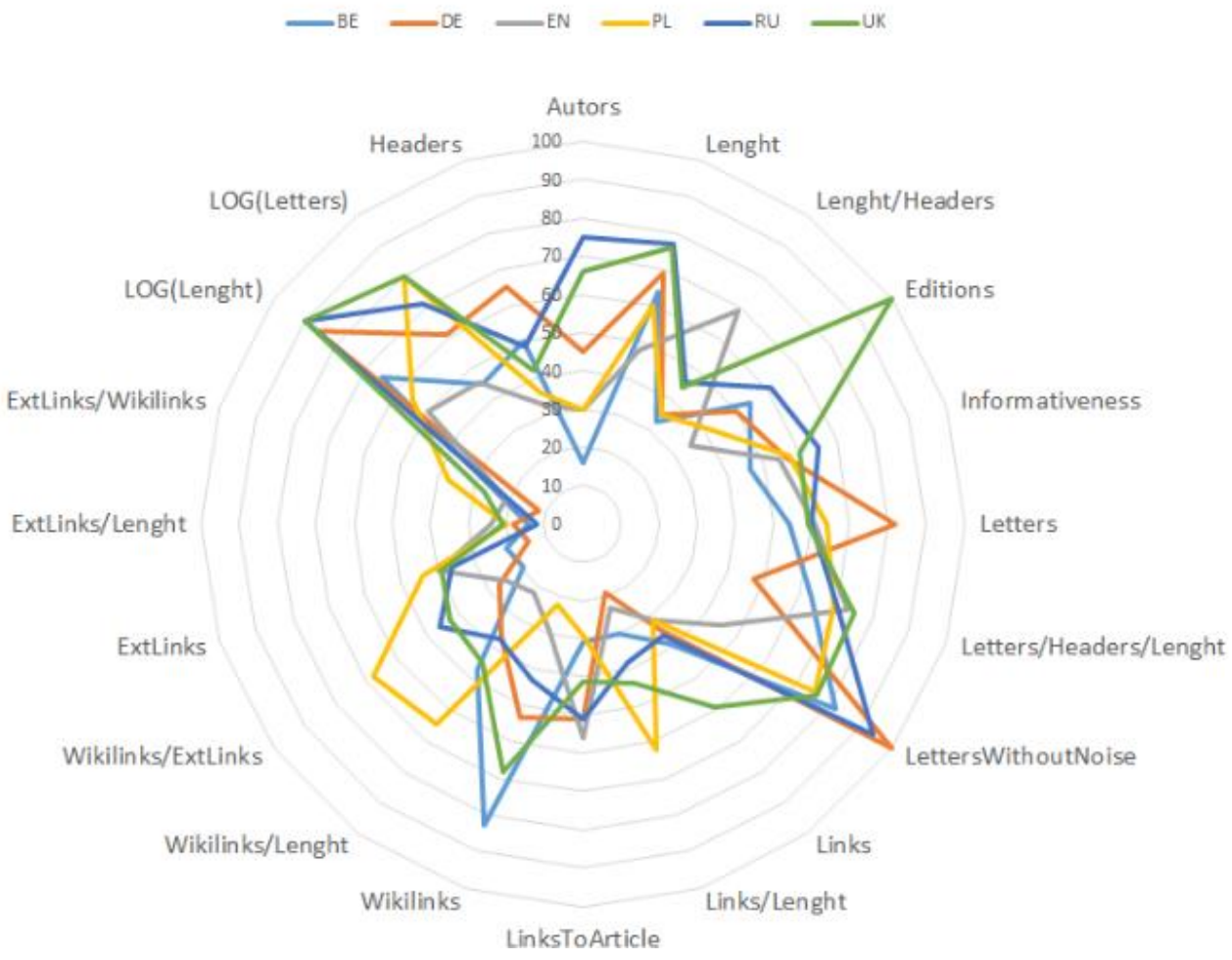
Source: Lewoniewski, W. (2017). Enrichment of information in multilingual Wikipedia based on quality analysis. In International Conference on Business Information Systems (pp. 216-227). Springer, Cham.

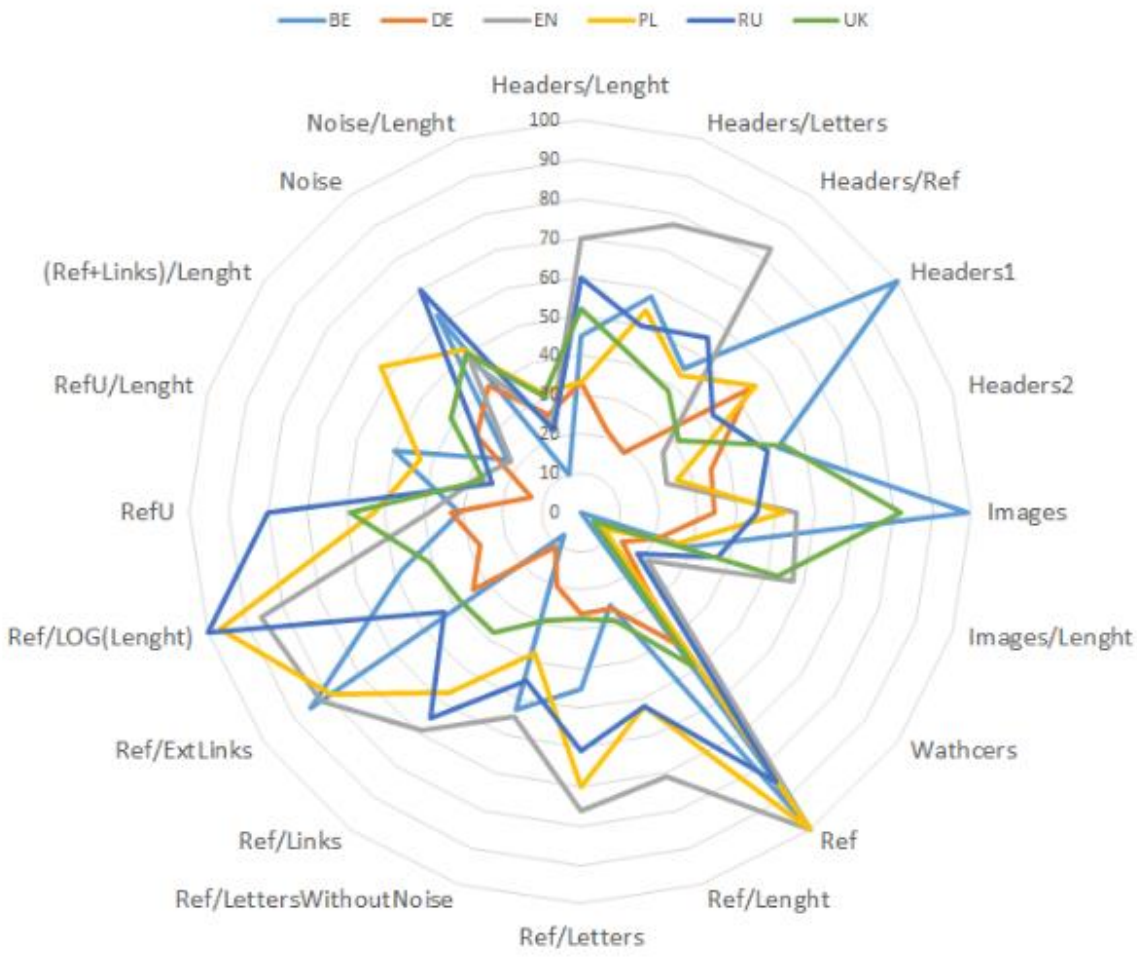
Quality dimensions



Source: *Lewoniewski, W. (2019).* [*Measures for quality assessment of articles and infoboxes in multilingual Wikipedia*](#).

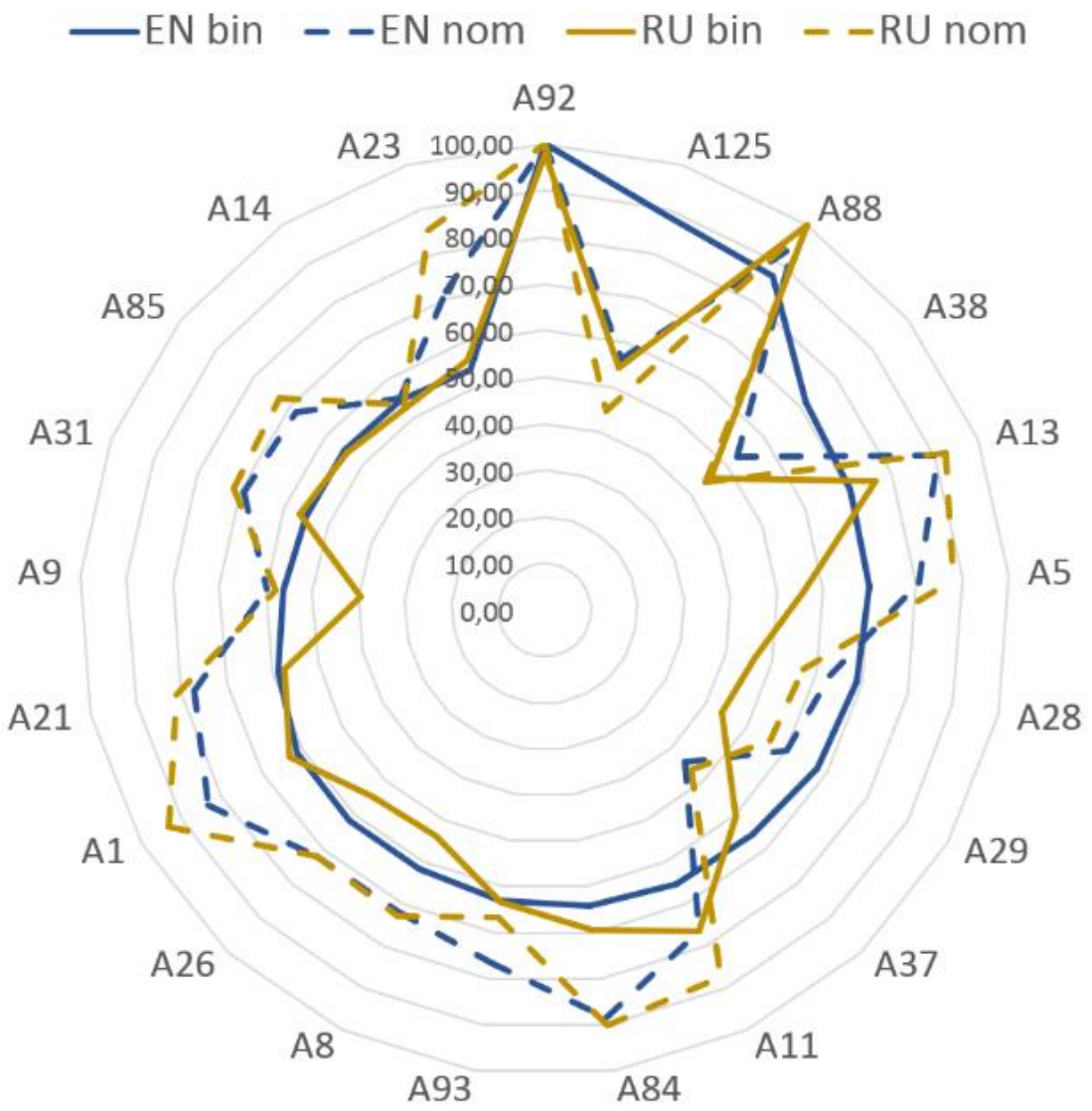
Significance of measures depending on language

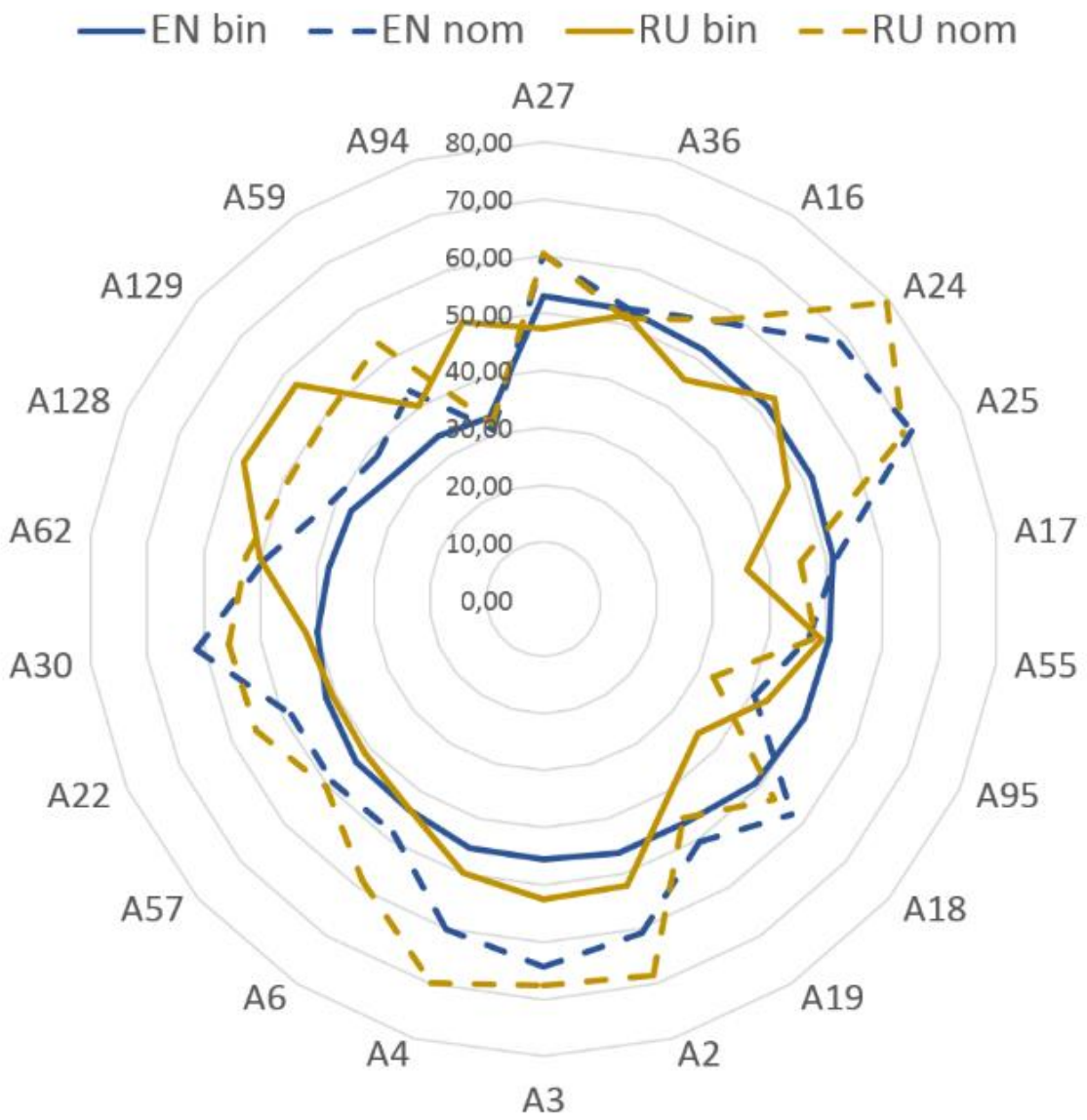




Source: Węcel, K., Lewoniewski, W. (2015). [Modelling the quality of attributes in Wikipedia infoboxes](#).

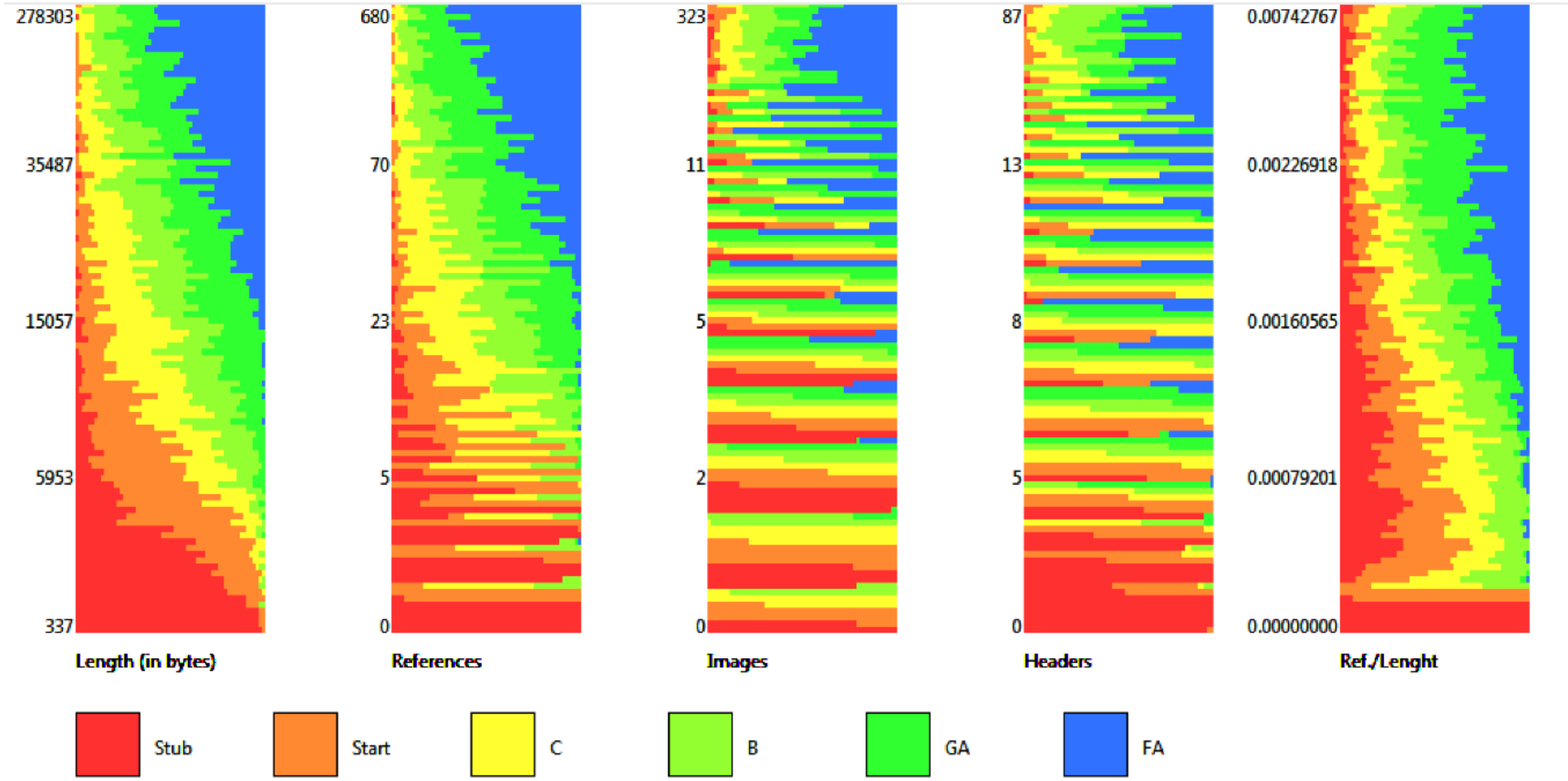
Significance of measures depending on language (2)





Source: Lewoniewski, W. (2018). [The method of comparing and enriching information in multilingual wikis based on the analysis of their quality](). PhD thesis

Distribution of measures in quality classes



Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2017). [Relative quality and popularity evaluation of multilingual Wikipedia articles](). Informatics 2017, 4(4), 43.

Article quality score - synthetic quality measure

Normalization of each measure m_i was conducted according to the following rule:

- if value of a given feature in a given language exceeded the threshold of the median value of the best articles in the same language version, it was set to 100 points;
- otherwise, its value was linearly scaled to reflect the relation of the value to the median value.

Normalized measures average (NMA):

$$NMA = \frac{1}{c}\sum_{i=1}^{c} \hat{m}_i$$

where $\hat{m}_i$ is a normalized measure m_i and c is the numer of measures.

Additionaly we need to take into account the numer of quality flaw templates (QFT) to measure the quality score:

$$QualityScore = NMA \cdot (1 - 5\% \cdot QFT)$$

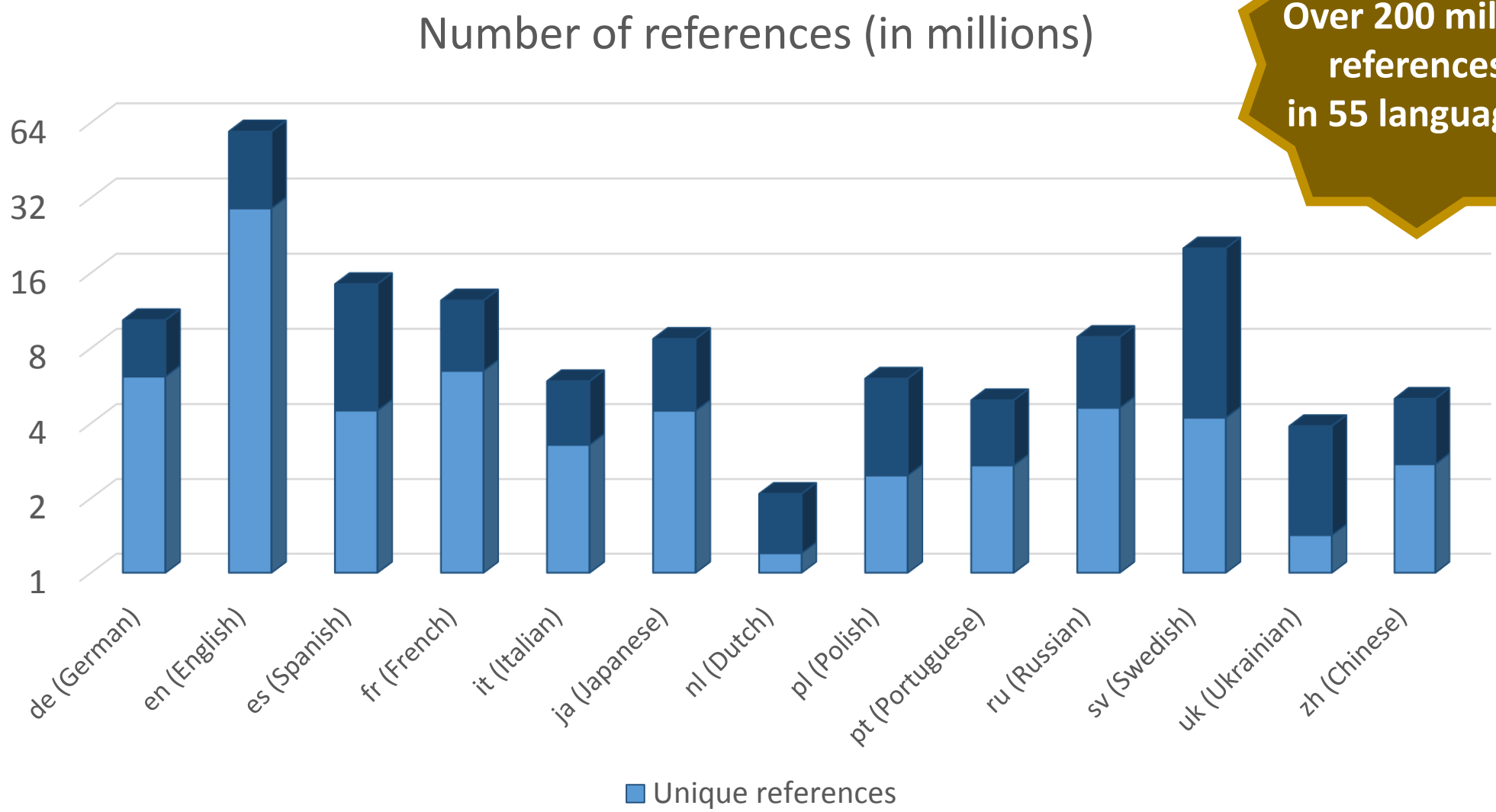
*For example, if the median for the number of references in Polish Wikipedia was **97**:*
- *any article with a larger number of references would score 100 for this feature;*
- *an article with **59** references would score proportionally 60.82 (**59/97**) points after normalizing.*

Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2019). Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. Computers, 8(3), 60.

Quality score – an example of implementation

Implementation of the quality score on WikiRank.net

Wikipedia references



Over 200 milion references in 55 languages

The calculation is based on Wikimedia dumps as of March 2020 using complex extraction of references. More languages:

Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

Wikipedia references – complex extraction

Epidemiology

Main articles: 2019–20 coronavirus pandemic by country and territory, cases/WHO situation reports, and deaths/WHO situation reports

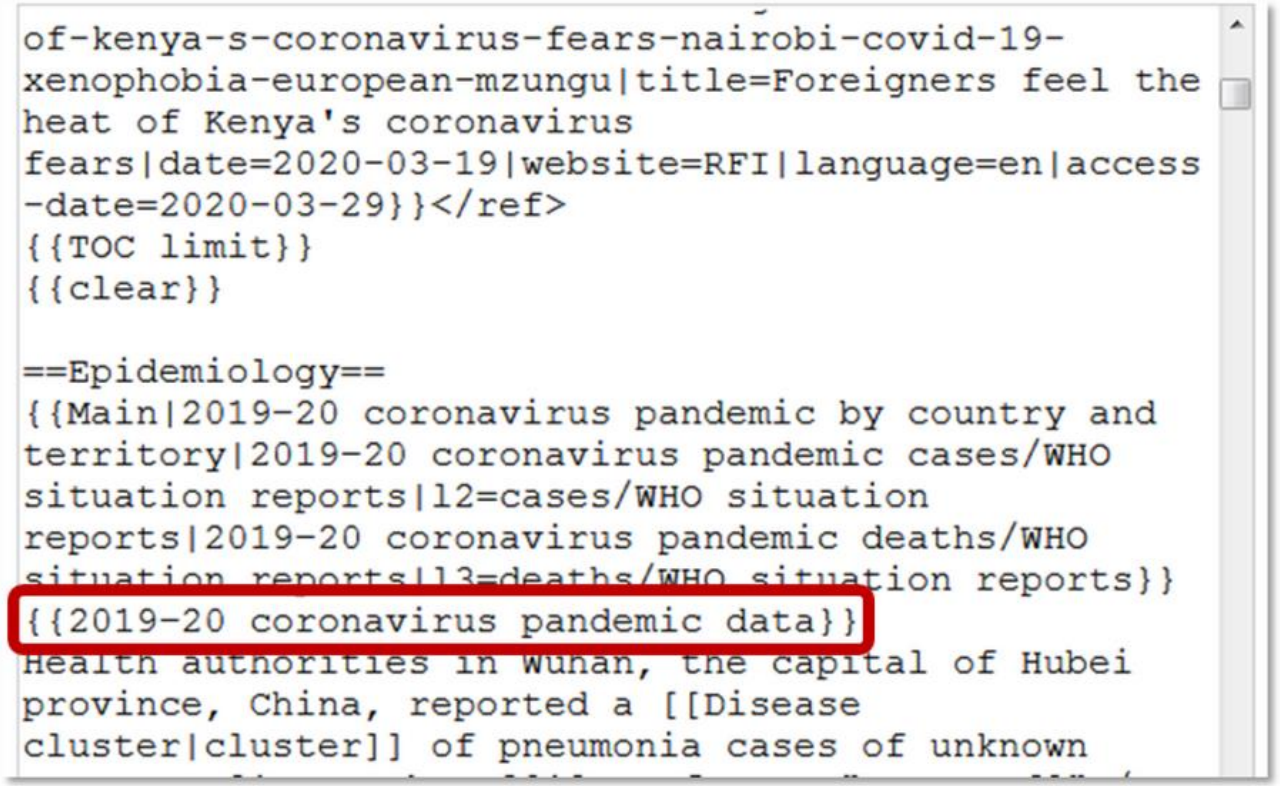
Health authorities in Wuhan, the capital of Hubei province, China, reported a cluster of pneumonia cases of unknown cause on 31 December 2019,[240] and an investigation was launched in early January 2020.[241] The cases mostly had links to the Huanan Seafood Wholesale Market and so the virus is thought to have a zoonotic origin.[242] The virus that caused the outbreak is known as SARS-CoV-2, a newly discovered virus closely related to bat coronaviruses,[243] pangolin coronaviruses,[244] and SARS-CoV.[245]

The earliest known person with symptoms was later discovered to have fallen ill on 1 December 2019, and they did not have visible connections with the later wet market cluster.[246][247] Of the early cluster of cases reported in December 2019, two-thirds were found to have a link with the market.[248][249][250] On 13 March 2020, an unverified report from the *South China Morning Post* suggested that a case traced back to 17 November 2019, in a 55-year-old from Hubei province, may have been the first.[251][252]

On 26 February 2020, the WHO reported that, as new cases reportedly declined in China but suddenly increased in Italy, Iran, and South Korea, the number of

2019–20 coronavirus pandemic by country and territory [show all]

Locations[b]	Cases[a]	Deaths[c]	Recov.[d]	Ref.
210	755,591	36,211	158,527	[41]
United States[e]	155,705	2,810	4,913	[44][45]
Italy[f]	101,739	11,591	14,620	[48][49]
Spain[g]	85,199	7,424	16,780	[51]
China[h]	81,470	3,304	75,770	[52]
Germany	64,527	568	6,522	[53]
Iran[i]	41,495	2,757	13,911	[54]
France[j]	40,174	2,606	7,202	[56][41]
United Kingdom[k]	22,141	1,408	135	[57]
Switzerland	15,923	354	2,105	[58]
Belgium	11,899	513	1,527	[59]
Netherlands[l]	11,750	864	–	[61]
Turkey	10,827	168	162	[62]
South Korea	9,661	158	5,228	[63]
Austria	9,541	108	636	[64][41]
Canada	7,274	70	956	[65]
Portugal	6,408	140	43	[66]
Norway[m]	4,445	32	–	[67]
Israel	4,347	15	132	[70]
Brazil	4,330	140	120	[44][41]
Australia[n]	4,247	18	226	[71]
Sweden[o]	4,106	161	18	[72][73]
Czech Republic	2,866	17	11	[74]
Denmark[p]	2,755	77	–	[77]
Malaysia	2,626	37	479	[78][79]
Ireland	2,615	46	5	[80]



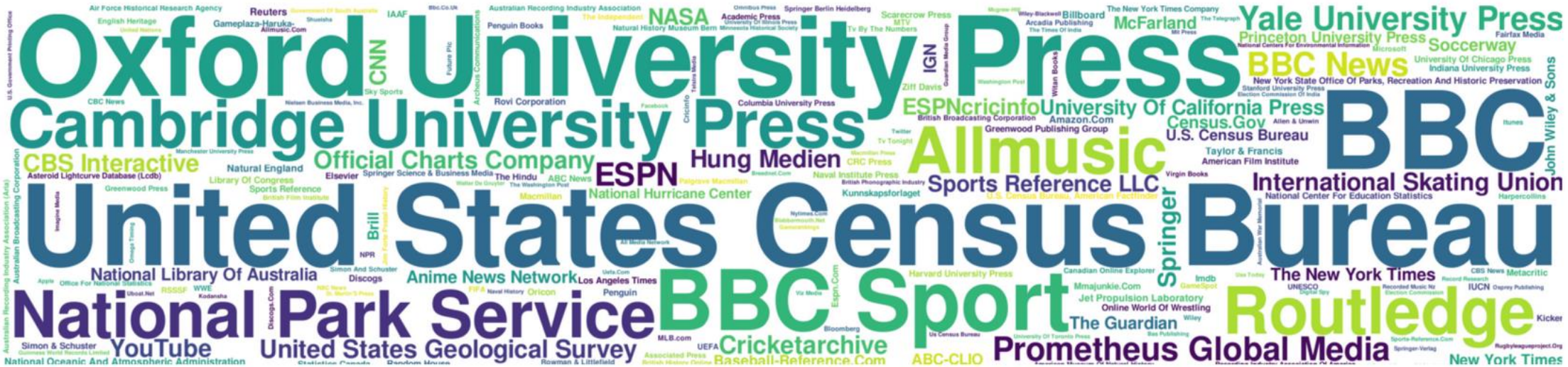
```
of-kenya-s-coronavirus-fears-nairobi-covid-19-
xenophobia-european-mzungu|title=Foreigners feel the
heat of Kenya's coronavirus
fears|date=2020-03-19|website=RFI|language=en|access
-date=2020-03-29}}</ref>
{{TOC limit}}
{{clear}}

==Epidemiology==
{{Main|2019-20 coronavirus pandemic by country and
territory|2019-20 coronavirus pandemic cases/WHO
situation reports|l2=cases/WHO situation
reports|2019-20 coronavirus pandemic deaths/WHO
situation reports|l3=deaths/WHO situation reports}}
{{2019-20 coronavirus pandemic data}}
Health authorities in Wuhan, the capital of Hubei
province, China, reported a [[Disease
cluster|cluster]] of pneumonia cases of unknown
```

Wiki markup of the article

Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

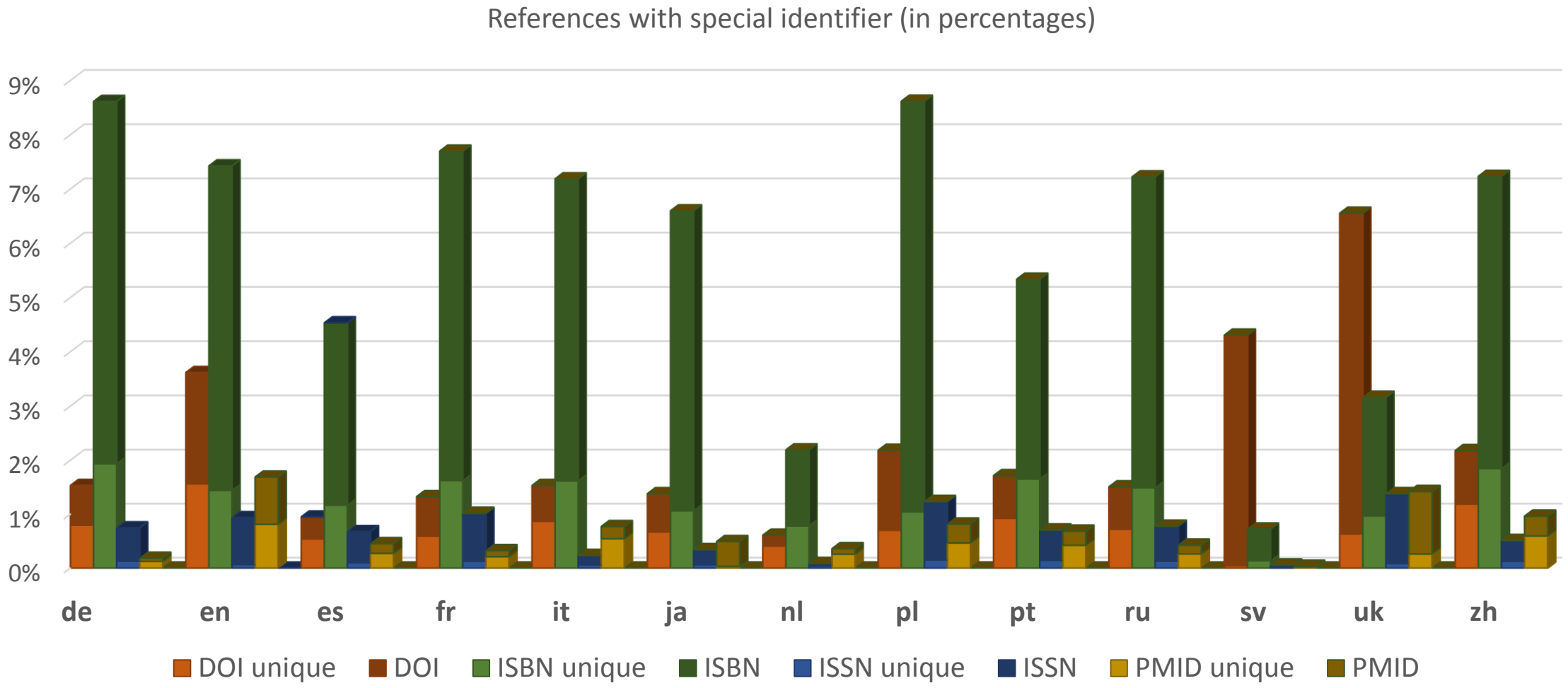
Templates in references on English Wikipedia

The most commonly used names in publisher parameter of citations templates:



Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

Wikipedia references with special identifiers



The calculation is based on Wikimedia dumps as of March 2020 using complex extraction of references. More languages and indentifiers:

Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

Popularity and reliability models

- **F** model—based on frequency (F) of source usage.
- **P** model—based on cumulative pageviews (P) of the article in which source appears.
- **PR** model—based on cumulative pageviews (P) of the article in which source appears divided by number of the references (R) in this article.
- **PL** model—based on cumulative pageviews (P) of the article in which source appears divided by article length (L).
- Models **Pm, PmR, PmL** are modified versions with daily pageviews median.
- Models **A, AR, AL** uses number of authors.

$$F(s) = \sum_{i=1}^{n} C_s(i)$$

$$P(s) = \sum_{i=1}^{n} C_s(i) \cdot V(i)$$

$$PR(s) = \sum_{i=1}^{n} \frac{V(i)}{C(i)} \cdot C_s(i)$$

$$PL(s) = \sum_{i=1}^{n} \frac{V(i)}{T(i)} \cdot C_s(i)$$

$$A(s) = \sum_{i=1}^{n} C_s(i) \cdot E(i)$$

$$AR(s) = \sum_{i=1}^{n} \frac{E(i)}{C(i)} \cdot C_s(i)$$

$$AL(s) = \sum_{i=1}^{n} \frac{E(i)}{T(i)} \cdot C_s(i)$$

Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

Publishers in references on English Wikipedia

Source	Position in the Ranking Depending on Model									
	F	P	PR	PL	Pm	PmR	PmL	A	AR	AL
AllMusic	8	28	8	8	26	8	9	14	6	7
BBC	3	4	3	3	5	4	3	2	3	3
BBC News	10	5	7	5	6	7	5	7	8	8
BBC Sport	4	11	15	12	16	17	13	5	7	5
Cambridge University Press	5	3	2	2	3	2	2	3	4	4
CBS Interactive	20	9	10	7	9	10	8	12	15	10
CNN	22	2	9	6	2	9	6	6	16	12
ESPN	13	8	17	14	8	19	16	10	13	14
IGN	32	37	29	24	34	29	23	22	17	15
National Park Service	7	94	38	48	89	47	58	60	12	11
Official Charts Company	16	30	24	18	31	26	18	18	21	18
Oxford University Press	2	1	1	1	1	1	1	1	2	1
Routledge	6	6	4	4	4	3	4	4	5	6
Springer	19	12	6	9	10	5	7	16	10	13
United States Census Bureau	1	27	5	11	24	6	12	9	1	2
University of California Press	24	16	19	19	12	18	17	15	19	22
Yale University Press	9	13	28	28	13	28	29	21	33	28

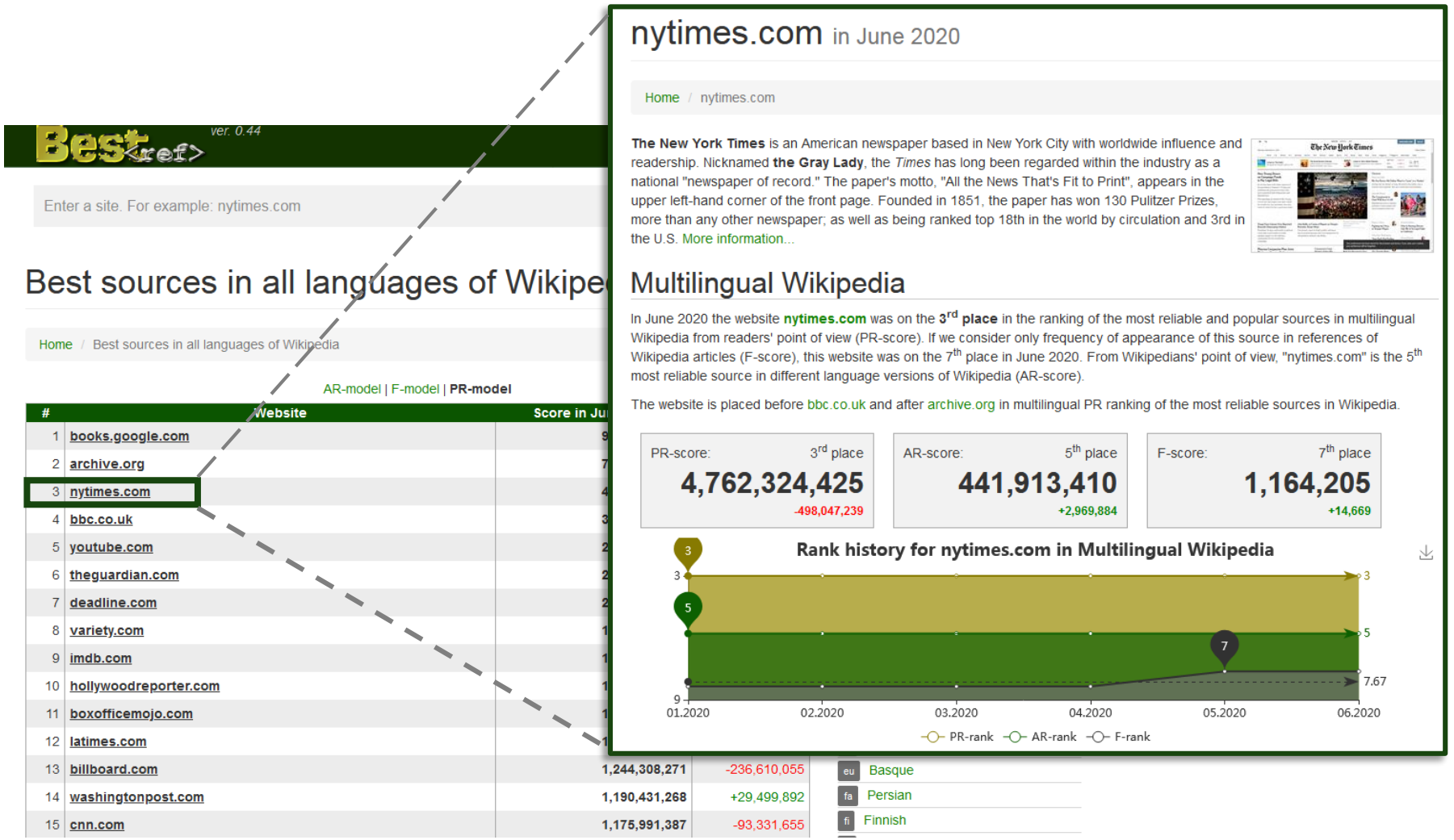
Position in rankings of publishers in English Wikipedia depending on popularity and reliability model in February 2020.

Source: own calculation based on Wikimedia dumps using complex extraction and using only values from publisher parameter of citation templates in references.

More publishers:

Lewoniewski, W., Węcel, K., Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. Information, 11(5), 263.

BestRef – analysis of sources on Wikipedia



Implementation of popularity and reliability models on BestRef.net

BestRef – analysis of sources on Wikipedia (2)

nytimes.com

Lang.	Models		
	F	PR	AR
all	7	3	5
ar	6	11	15
de	10	10	11
es	24	10	20
fr	14	18	18
it	12	11	13
ja	25	46	49
nl	23	17	23
pl	31	25	38
pt	10	11	22
ru	11	14	18
sv	46	23	32
uk	41	40	41
zh	13	27	32

spiegel.de

Lang.	Models		
	F	PR	AR
all	125	47	51
ar	270	423	415
de	2	1	2
es	413	457	584
fr	401	471	392
it	329	359	387
ja	692	992	899
nl	117	113	135
pl	534	392	421
pt	501	497	543
ru	355	333	324
sv	269	250	216
uk	342	501	340
zh	469	787	689

lemonde.fr

Lang.	Models		
	F	PR	AR
all	111	67	76
ar	343	556	418
de	341	393	355
es	280	398	506
fr	5	1	3
it	315	275	328
ja	701	1381	1244
nl	203	345	342
pl	730	689	675
pt	438	514	597
ru	622	753	875
sv	663	1023	630
uk	771	1522	1013
zh	526	1339	1128

elpais.com

Lang.	Models		
	F	PR	AR
all	100	56	62
ar	265	852	400
de	312	445	346
es	21	3	3
fr	109	248	199
it	248	281	256
ja	630	2096	1236
nl	351	620	435
pl	577	1018	785
pt	102	67	101
ru	635	832	741
sv	577	963	758
uk	720	1537	916
zh	371	1051	729

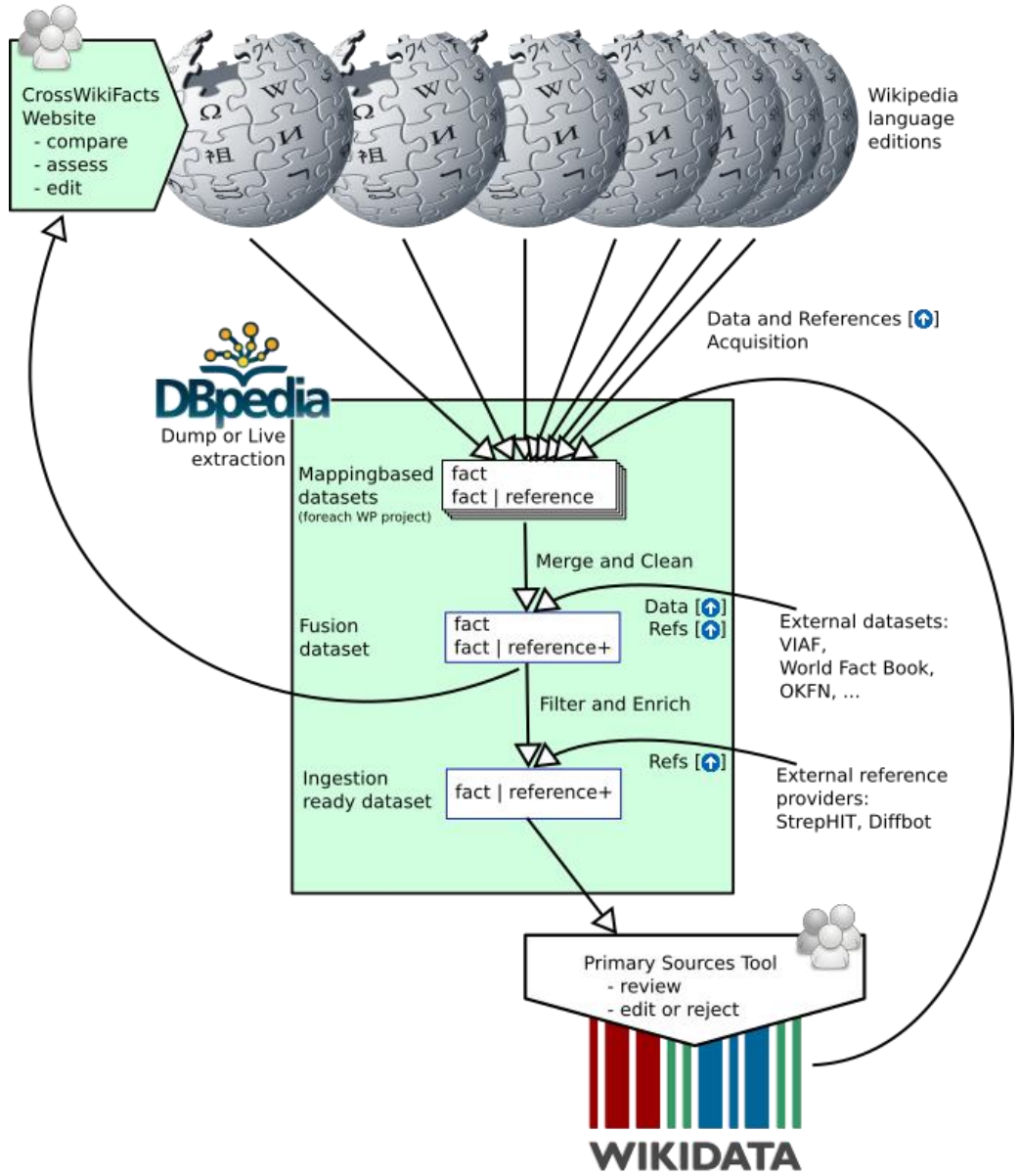
Source: BestRef.net

Browser extensions for quality assessment of Wikipedia

- Articles - WikiRank
  - Chrome: chrome.google.com/webstore/detail/wikirank/cnomlnphfhgijoghjcbdpmmhfgeooabd
  - Firefox: addons.mozilla.org/en-US/firefox/addon/wikirank
  - presentation: youtube.com/watch?v=jJdKw2gf1aA
- Infoboxes
  - Chrome: chrome.google.com/webstore/detail/infoboxes/njjjplipinhcglgjopmnnlphmlhdpkko
  - presentation: youtube.com/watch?v=HCfvx0wQ5oM
- Sources - BestRef
  - Chrome: chrome.google.com/webstore/detail/bestref/bnlfiilmigfboedocmjaejbklgbdmmio
  - presentation: youtube.com/watch?v=FXnfaAIaixc

Futher applications

- Quality models can help to enrich various language editions of Wikipedia and other knowledge bases with information of better quality.
- Some of the approaches are planned to be implemented on global.dbpedia.org



GlobalFactSync data flow.
Source: commons.wikimedia.org/wiki/File:GFS.png

Thank you

E-mail: wlodzimierz.lewoniewski@ue.poznan.pl

Web: kie.ue.poznan.pl