Modelling the quality of attributes in Wikipedia infoboxes

Krzysztof Węcel, Włodzimierz Lewoniewski

Poznań University of Economics

Extraction of attributes from infobox to DBpedia in one language

dbpedia-owl:areaCode	▪ +48 61
dbpedia-owl:areaTotal	▪ 261850000.000000 (xsd:double)
dbpedia-owl:country	▪ dbpedia:Poland
dbpedia-owl:isPartOf	▪ dbpedia:Greater_Poland_Voivodeship
dbpedia-owl:leaderName	▪ dbpedia:Ryszard_Grobelny
dbpedia-owl:leaderTitle	▪ Mayor
dbpedia-owl:maximumElevation	▪ 154.000000 (xsd:double)
dbpedia-owl:minimumElevation	▪ 60.000000 (xsd:double)
dbpedia-owl:populationMetro	▪ 1400000 (xsd:integer)
dbpedia-owl:populationTotal	▪ 551627 (xsd:integer)
dbpedia-owl:populationUrban	▪ 1100000 (xsd:integer)
dbpedia-owl:postalCode	▪ 60-001 to 61-890
dbpedia-owl:thumbnail	▪ http://commons.wikimedia.org/wiki/Sp
dbpedia-owl:timeZone	▪ dbpedia:Central_European_Time
dbpedia-owl:utcOffset	▪ +1 ▪ +2
dbpedia-owl:wikiPageExternalLink	▪ http://www.stat.gov.pl/cps/rde/xchg/ ▪ http://www.checkoutpoland.com/citie ▪ http://www.dariointernet.com/tr_pola ▪ http://www.ipoznan.pl/ ▪ http://www.mpk.poznan.pl/ ▪ http://www.poznan.pl/ ▪ http://www.poznan.pl/mim/public/inde ▪ http://www.youarein.poznan.pl/ ▪ http://www.poznan.pl
dbpedia-owl:wikiPageID	▪ 23723 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	▪ 606741220 (xsd:integer)
dbpprop:aprHighC	▪ 13.600000 (xsd:double)
dbpprop:aprHumidity	▪ 68 (xsd:integer)
dbpprop:aprLowC	▪ 6 (xsd:integer)
dbpprop:aprMeanC	▪ 9.800000 (xsd:double)
dbpprop:aprPrecipitationDays	▪ 9 (xsd:integer)
dbpprop:aprPrecipitationMm	▪ 38 (xsd:integer)
dbpprop:aprRecordHighC	▪ 30 (xsd:integer)
dbpprop:aprRecordLowC	▪ -6.800000 (xsd:double)
dbpprop:aprSun	▪ 179 (xsd:integer)
dbpprop:areaCode	▪ +48 61
dbpprop:areaTotalKm	▪ 261.850000 (xsd:double)
dbpprop:augHighC	▪ 24.500000 (xsd:double)
dbpprop:augHumidity	▪ 72 (xsd:integer)

http://dbpedia.org/resource/Poznań

English

Attributes in infobox

en.wikipedia.org/wiki/Poznań

Extraction of attributes from infoboxes to DBpedia in different languages

leaderName
populationTotal
establishedDate

Common properties

Properties of which language have the highest quality?

Properties

Properties

Polish

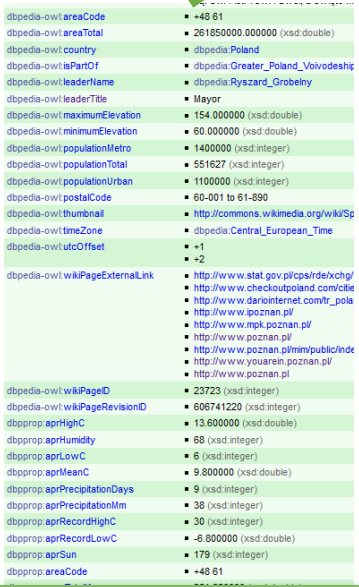
Attributes in infobox

English

Attributes in infobox

pl.wikipedia.org/wiki/Poznań

pl.dbpedia.org/resource/Poznań



dbpedia.org/resource/Poznań

en.wikipedia.org/wiki/Poznań

Quality of attributes vs. quality of the article

- Does the quality of the paper positively influences the quality of the attributes in the infobox?

Quality of the attribute

Quality of the article:

- Scope
- Format
- Uniqueness
- Accuracy
- Currency
- Accessibility

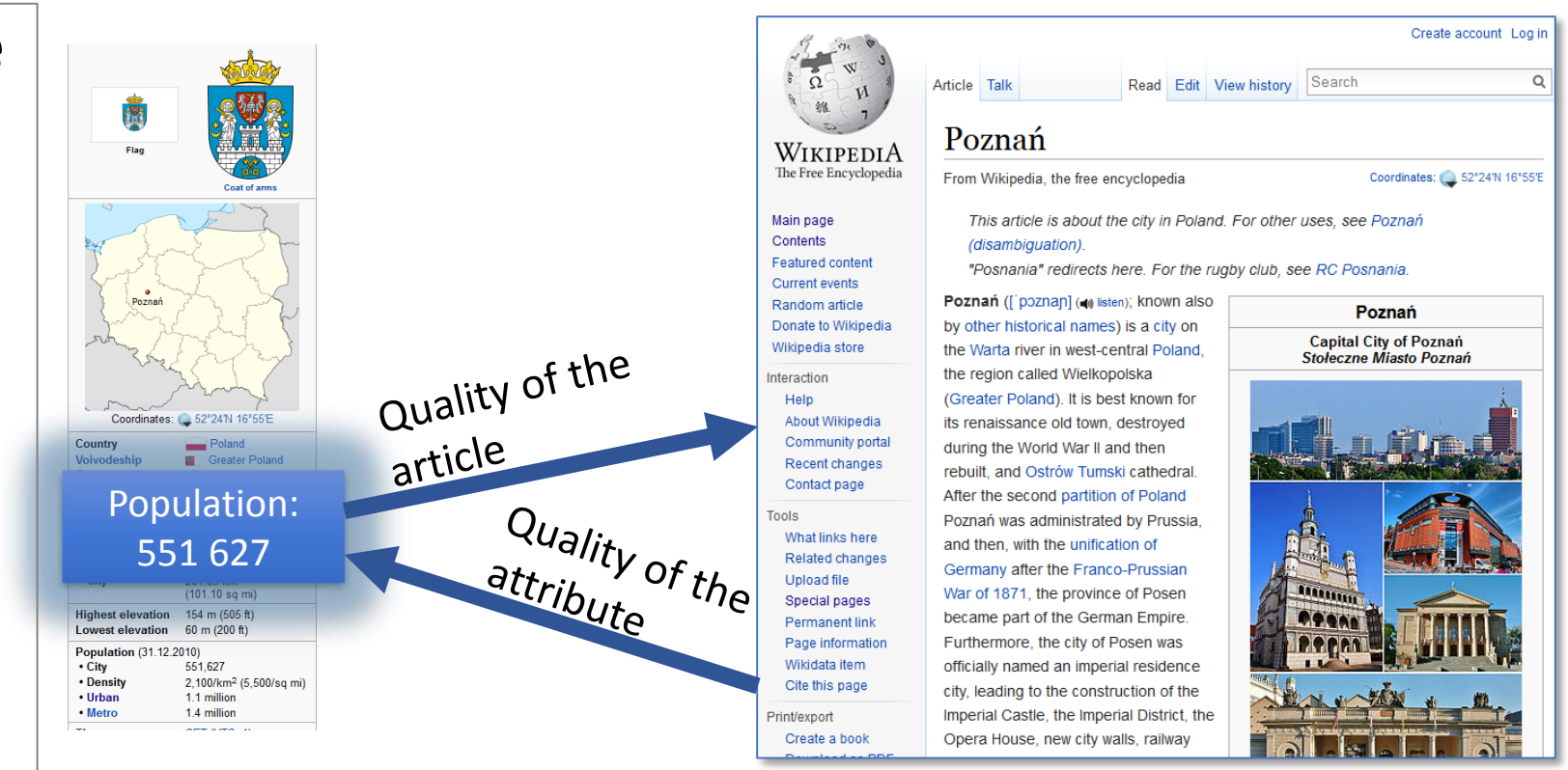
[Mark Anderka, 2013. Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia]

Quality of attributes vs. quality of the article

- Does the quality of the attribute in infobox influences the overall quality of the article?

Quality of the attribute (quality of data):

- Completeness
- Accuracy
- Format
- Timeliness



Quality of an article in Wikipedia

- article can be awarded a grade by users
- several grading systems exist:
  - community discussion
  - self-evaluation
- grading systems have specific rules and principles for evaluating articles
- rules can change in time

„Quality of information should be understood as a degree to which the user needs are satisfied”.
[W.Abramowicz, Filtrowanie informacji, 2008.]

Different languages – different quality classifications

Name/Language	BE	DE	EN	PL	RU	UK
Featured Article	X	X	X	X	X	X
Good Article	X	X	X	X	X	X
Solid article					X	
A-class			X			
Four				X		
Full article					X	X
B-class			X			
Developed					X	X
C-class			X			
In development					X	X
Start			X	X		
Stub	X		X	X	X	X

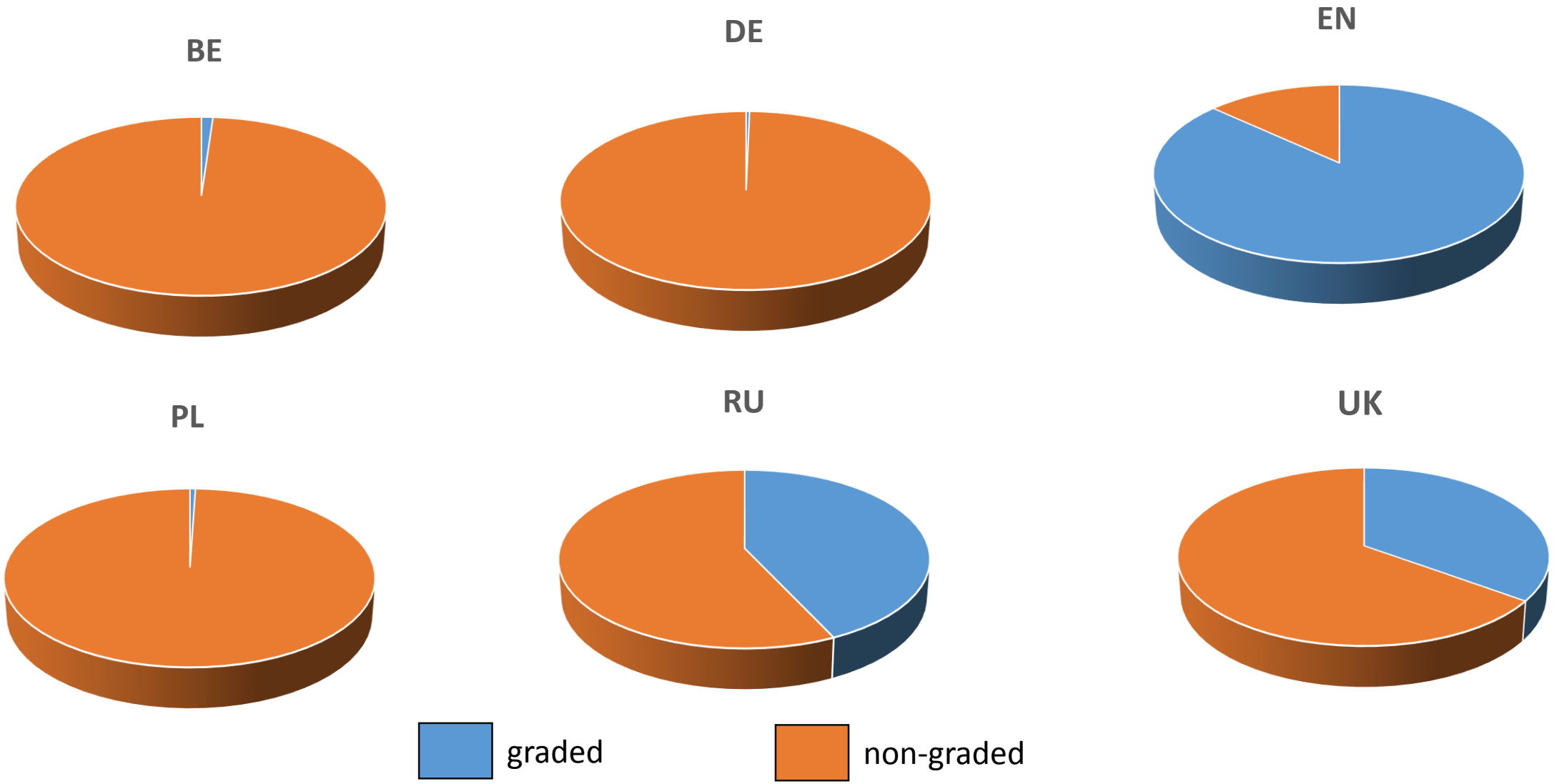
Colors mark similar classes

Number of articles in each class

Name/Language Number of articles	BE 89 923	DE 1 828 090	EN 4 741 168	PL 1 099 441	RU 1 198 199	UK 557 590
Featured Article	**54**	**2 383**	**4 481**	**648**	**768**	**204**
Good Article	**97**	**3 785**	**21 697**	**1 993**	**2 041**	**531**
Solid article					**1 661**	
A-class			**821**			
Four				**157**		
Full article					**3 978**	**195**
B-class			**80 919**			
Developed					**15 492**	**707**
C-class			**194 063**			
Developing					**60 801**	**3 490**
Start			**1 177 495**	**928**		
Stub	**860**		**2 634 800**	**2 071**	**429 926**	**189 742**

Colors mark similar classes

Number of graded articles



Motivation for automatic grading of articles

- majority of articles is not graded at all by users
- only small fraction is classified in the highest quality classes
- articles can be incorrectly classified in self-evaluation

Dataset for the experiment

- in English Wikipedia there are 7 quality classes
  - A-class is also often classified as GA, so the class was rejected
- 2000 articles have been chosen randomly for each class
- own application WikiAnalyzer have been used for data collection about the following measures:

- length (in bytes)
- number of letters
- num of letters without noise
- unique references
- all references

- observers
- num of editions
- num of authors
- displays in last 90 days
- incoming links
- num of figures

- internal links
- external links
- num of 1st level headers
- num of 2nd level headers

Dataset for the experiment

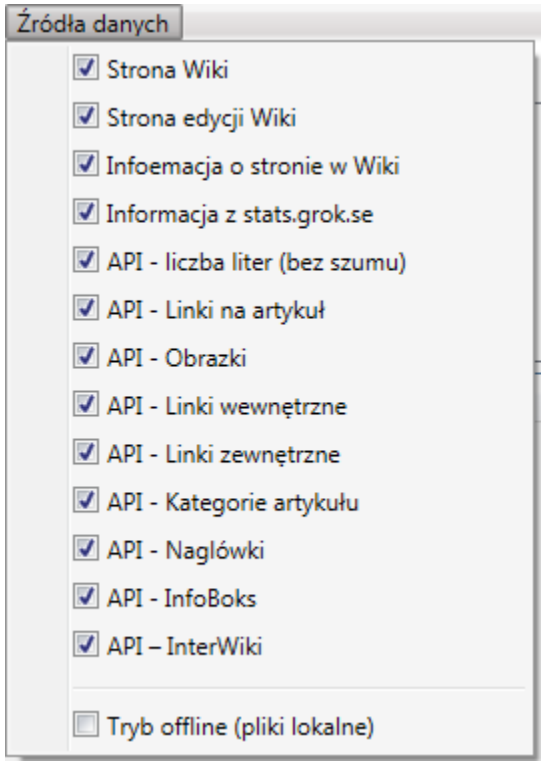
In some languages number of articles had to be lowered due to small number of graded articles

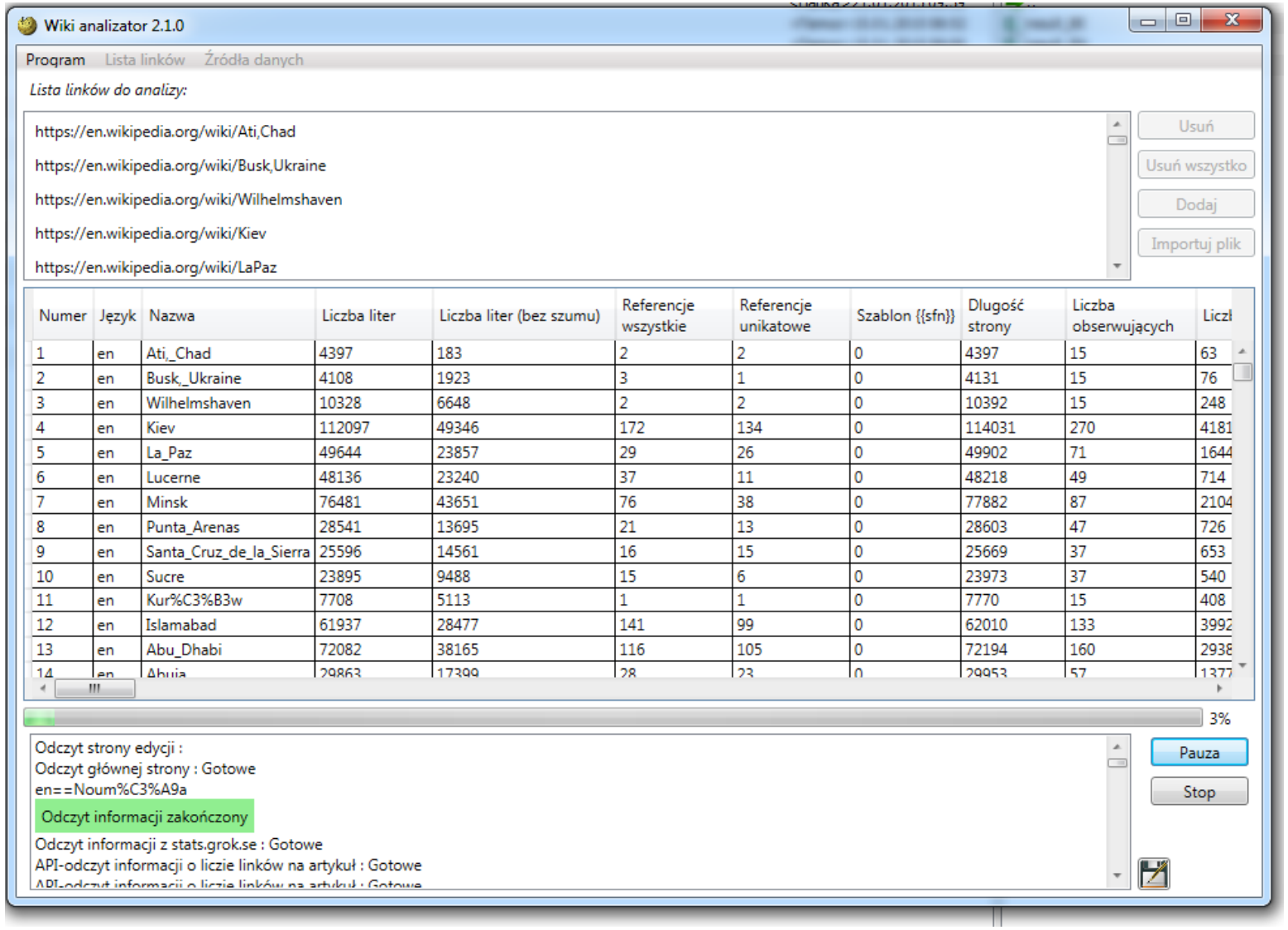
The following number of articles were chosen:

- German (DE) – 2000 articles in each class
- Russian (RU) – 744 articles in each class
- Ukrainian (UK) – 191 articles in each class
- Polish (PL) – 152 articles in each class
- Belarusian (BE) – 52 articles in each class

- *Experiment was conducted in Feb 2015*

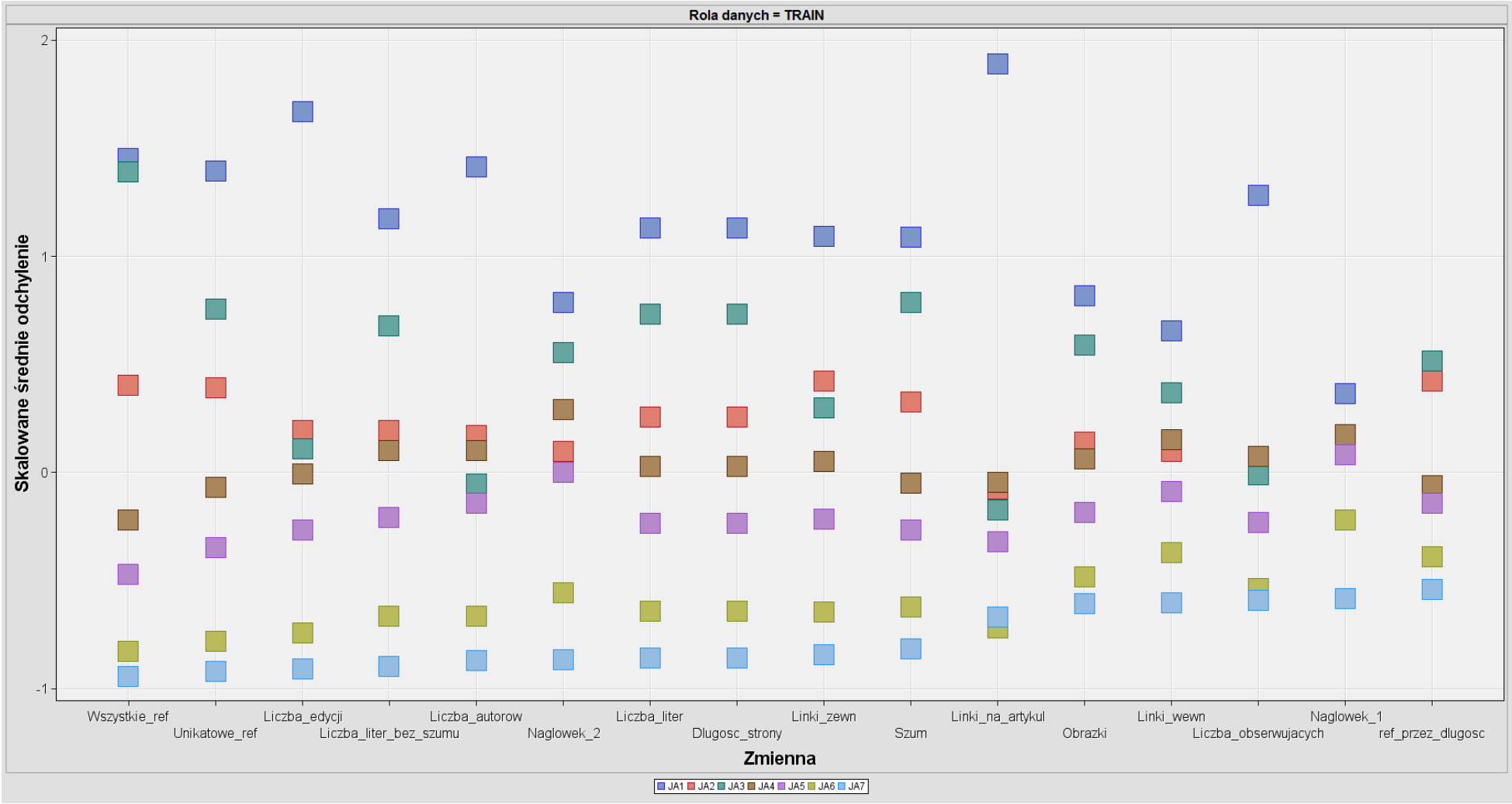
WikiAnalyzer 2





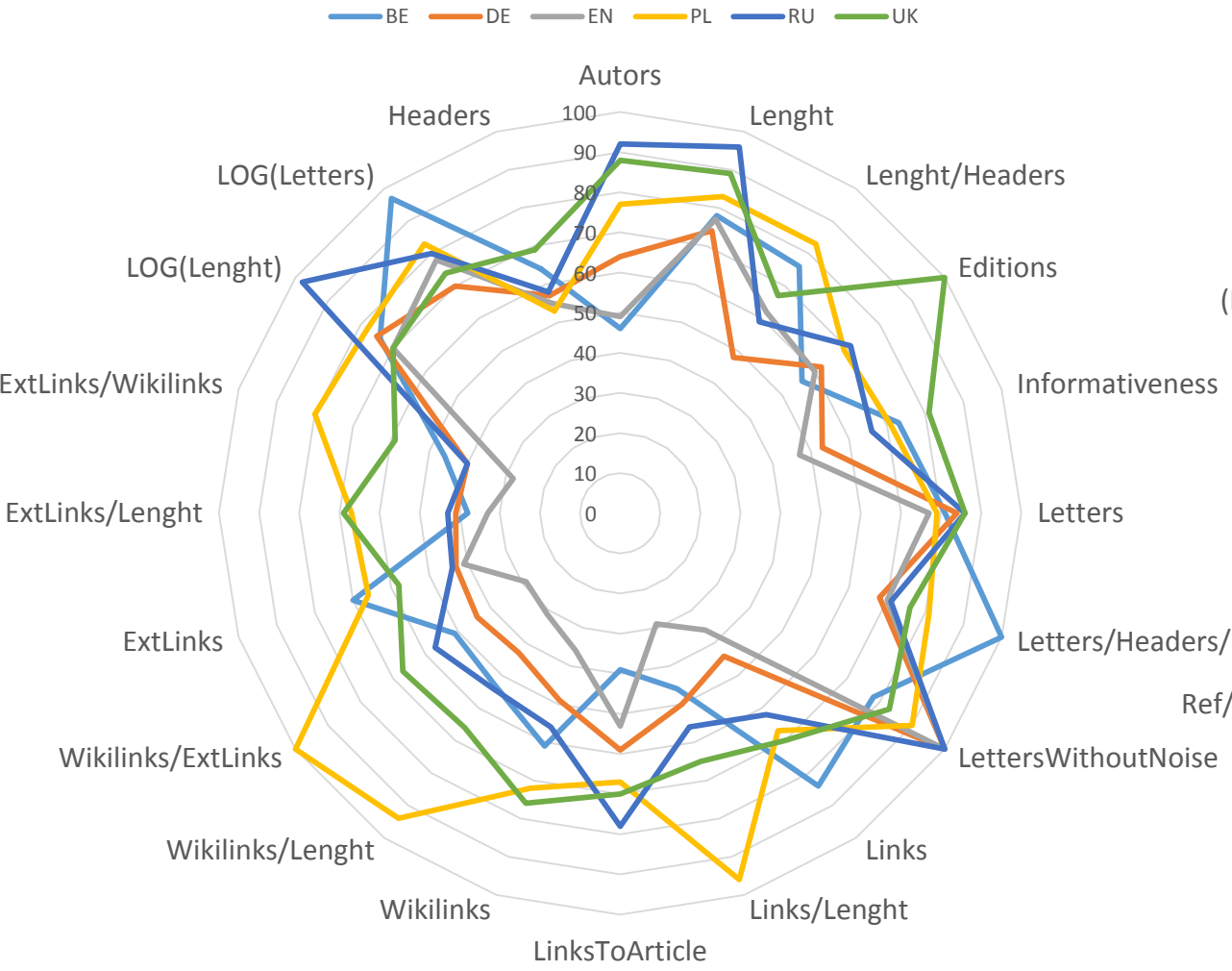
Numer	Język	Nazwa	Liczba liter	Liczba liter (bez szumu)	Referencje wszystkie	Referencje unikatowe	Szablon {{sfn}}	Dlugość strony	Liczba obserwujących	Liczł
1	en	Ati,_Chad	4397	183	2	2	0	4397	15	63
2	en	Busk,_Ukraine	4108	1923	3	1	0	4131	15	76
3	en	Wilhelmshaven	10328	6648	2	2	0	10392	15	248
4	en	Kiev	112097	49346	172	134	0	114031	270	4181
5	en	La_Paz	49644	23857	29	26	0	49902	71	1644
6	en	Lucerne	48136	23240	37	11	0	48218	49	714
7	en	Minsk	76481	43651	76	38	0	77882	87	2104
8	en	Punta_Arenas	28541	13695	21	13	0	28603	47	726
9	en	Santa_Cruz_de_la_Sierra	25596	14561	16	15	0	25669	37	653
10	en	Sucre	23895	9488	15	6	0	23973	37	540
11	en	Kur%C3%B3w	7708	5113	1	1	0	7770	15	408
12	en	Islamabad	61937	28477	141	99	0	62010	133	3992
13	en	Abu_Dhabi	72082	38165	116	105	0	72194	160	2938
14	en	Abuja	29863	17399	28	23	0	29953	57	1377

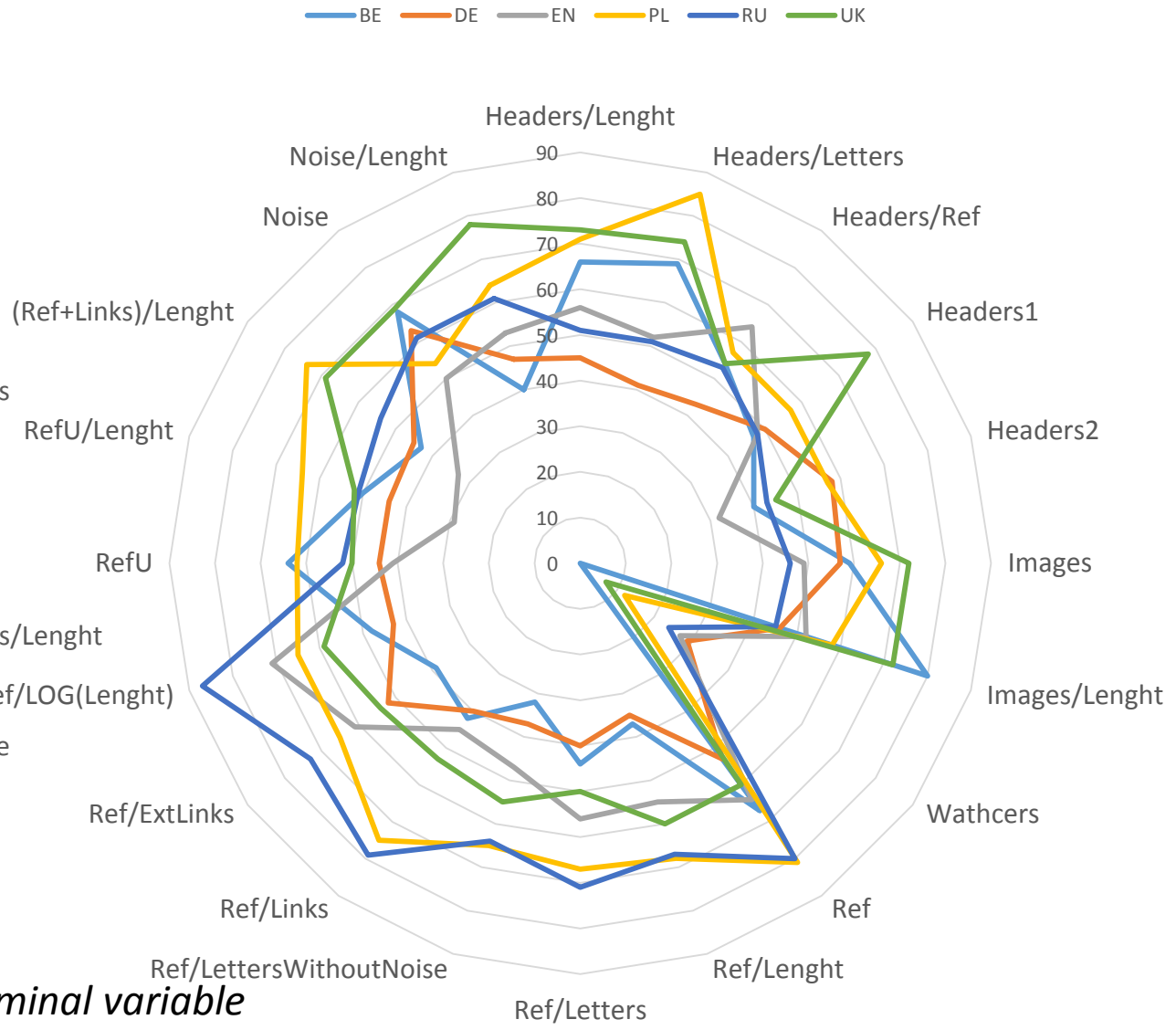
Layers of measures of quality



***Source**: SAS Enterprise Miner, boosting tree. Quality as nominal variable*

Significance of variables (nominal)





***Source**: Statistica. random forest of 100 trees. Quality as nominal variable*

Error levels (quality – nominal variable)

- relatively high error rate in each language

	Misclassif.	Standard Error
BE:		
Train	0,150000	0,030178
Test	0,235294	0,051440
EN:		
Train	0,382743	0,008083
Test	0,436387	0,012508
RU:		
Train	0,382743	0,008083
Test	0,436387	0,012508

	Misclassif.	Standard Error
DE:		
Train	0,449976	0,005469
Test	0,465123	0,008382
PL:		
Train	0,272727	0,019587
Test	0,365145	0,031014
UK:		
Train	0,349068	0,016801
Test	0,502959	0,027196

***Source**: Statistic, random forest of 100 trees. Quality as nominal variable*

Quality as binary variable

All articles can be just split into two classes:

- complete (1, labeled GoodEnough): inclused FA and GA
- developing (0, NeedsWork): includes everything else

Motivation for the approach

- in all analysed language editions there are FA and GA classes
- FA and GA classes are the most reliable (voting)
- quality can be modelled as probability of belonging to one of the classes
- proven effectiveness of such split*

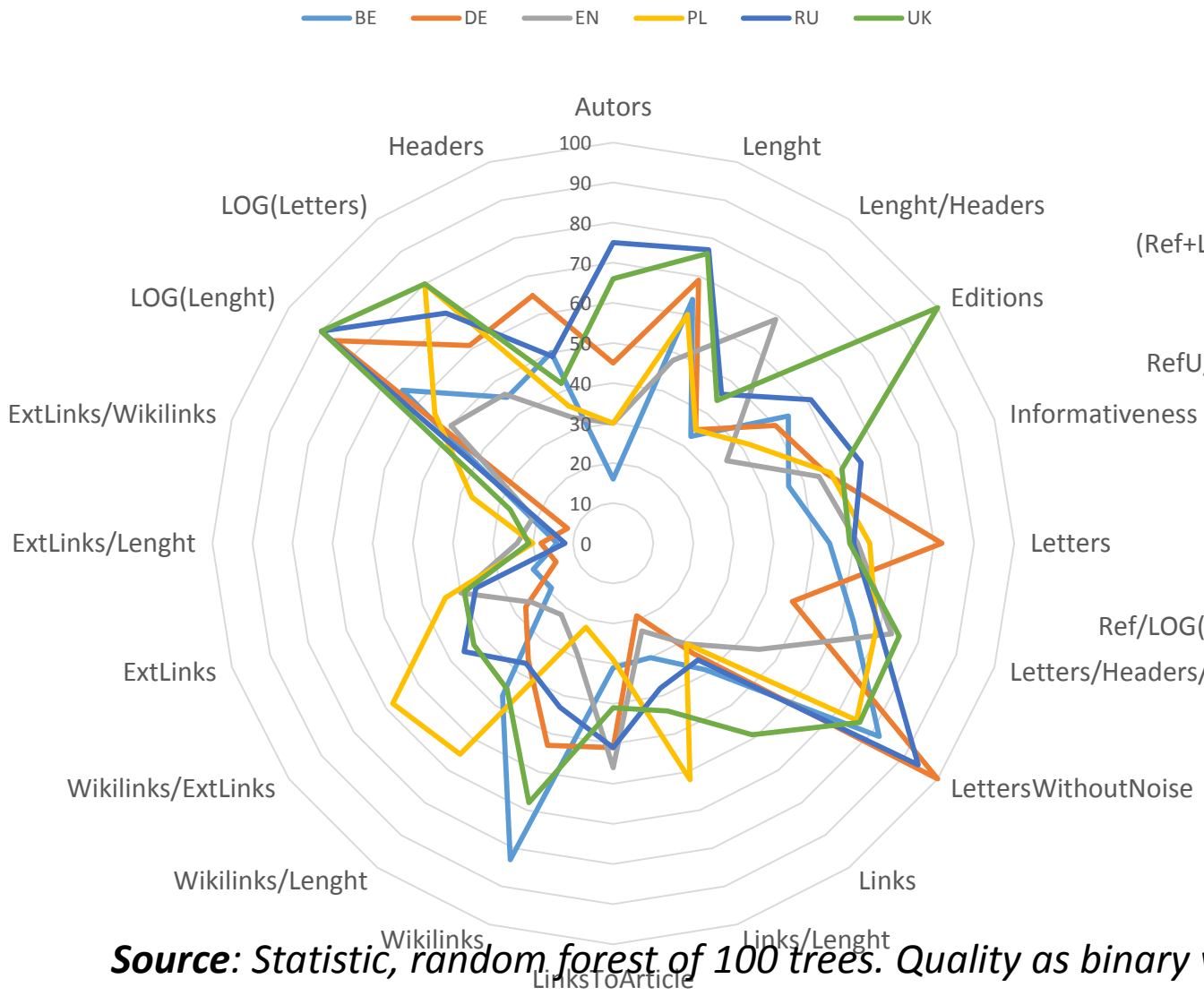
** Warncke-wang at all., Tell Me More: An Actionable Quality Model for Wikipedia. WikiSym 2013*

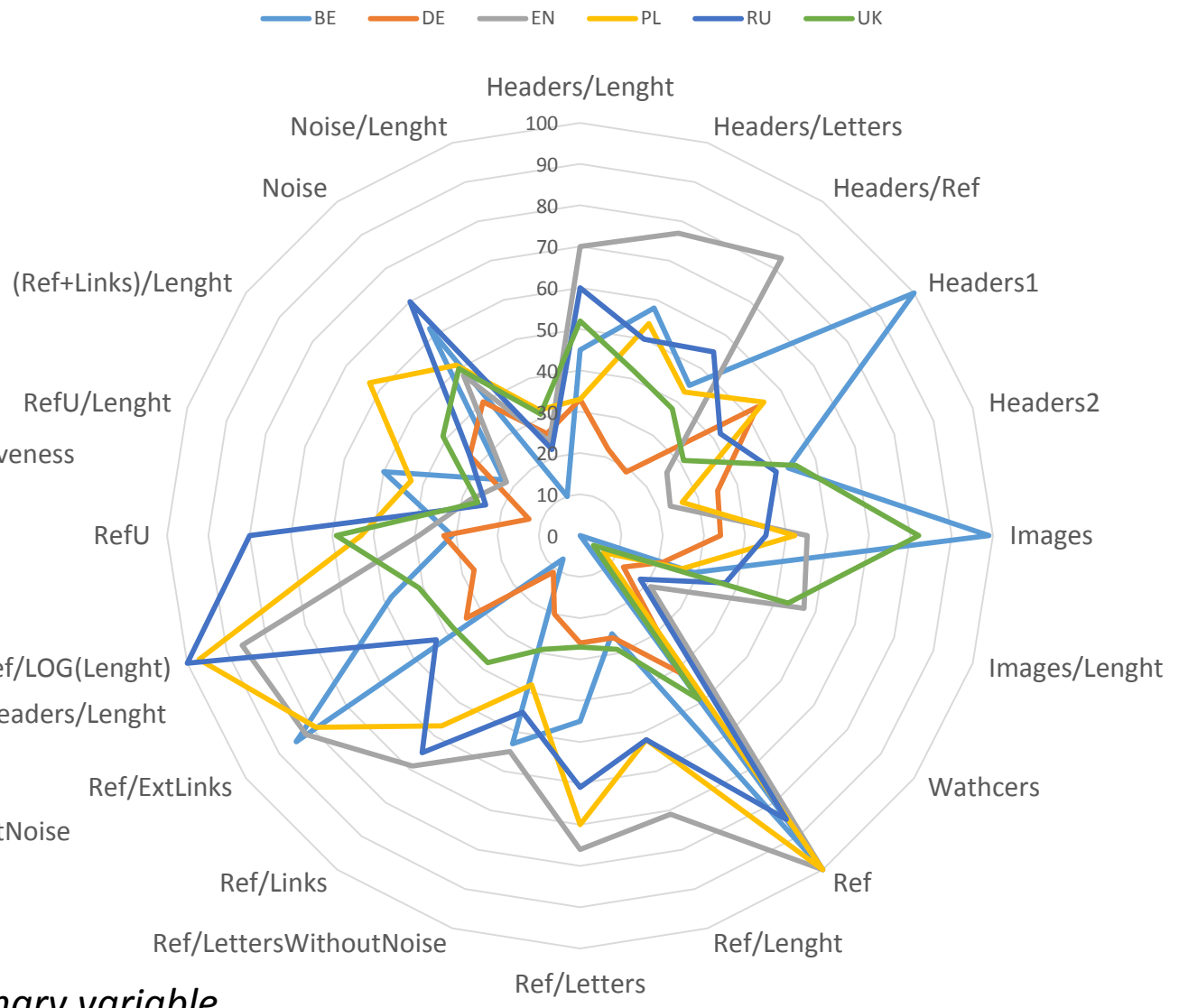
Quality as binary variable

Name/Language	BE	DE	EN	PL	RU	UK	
Featured Article	X	X	X	X	X	X	1
Good Article	X	X	X	X	X	X	
Solid article					X		
A-class			X				
Four				X			
Full article					X	X	
B-class			X				
Developed					X	X	0
C-class			X				
Developing					X	X	
Start			X	X			
Stub	X		X	X	X	X	

Colors mark similar classes

Significance of variables (binary)





***Source**: Statistic, random forest of 100 trees. Quality as binary variable*

Error levels (quality – binary variable)

- error much lower than in nominal data

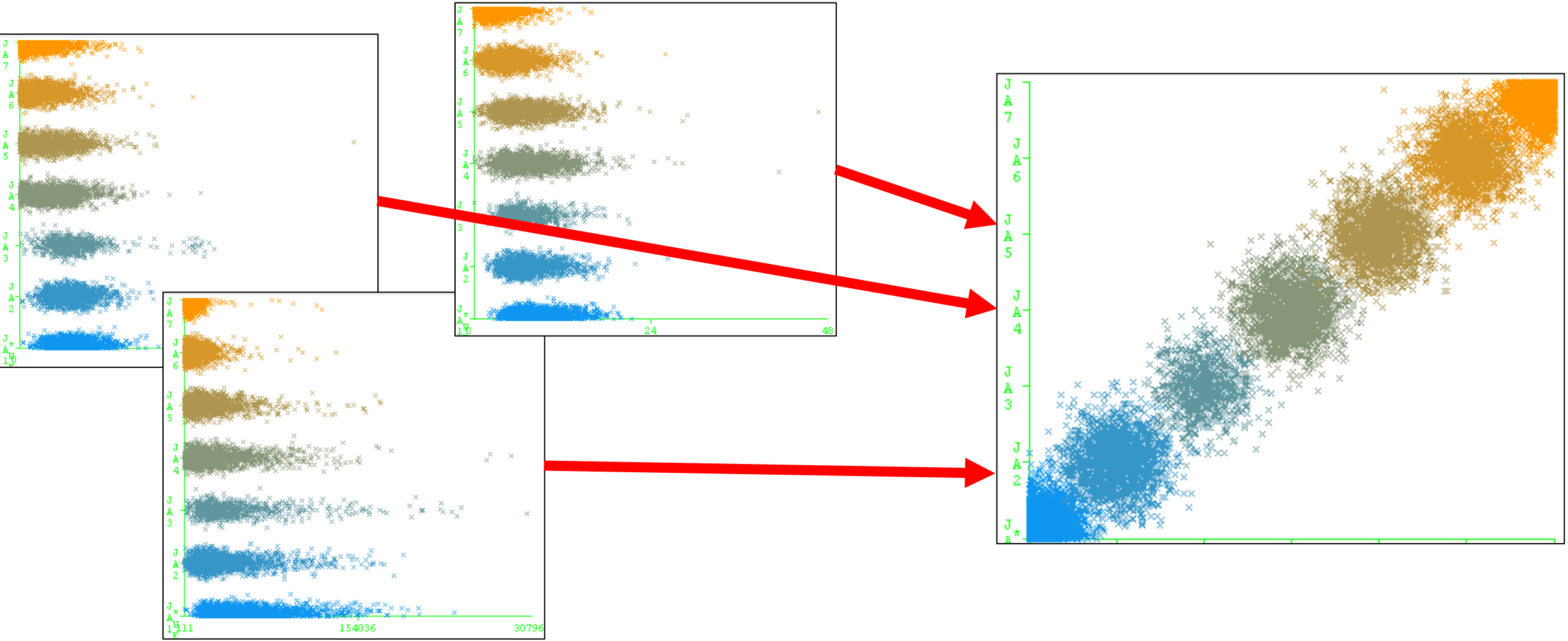
	Misclassif.	Standard Error
BE:		
Train	0,014286	0,010029
Test	0,000000	0,000000
EN:		
Train	0,115273	0,003510
Test	0,124823	0,005554
RU:		
Train	0,067478	0,004172
Test	0,080789	0,006873

	Misclassif.	Standard Error
DE:		
Train	0,032972	0,002801
Test	0,036077	0,004427
PL:		
Train	0,087041	0,012398
Test	0,078838	0,017359
UK:		
Train	0,081988	0,009669
Test	0,091716	0,015699

***Source**: Statistic, random forest of 100 trees. Quality as binary variable*

Model estimation

- we need better discriminative power in order to more precisely distinguish various classes



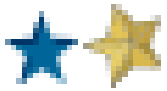
Article quality vs. timeliness of infobox attribute

517,447 English Wikipedia articles from class **PopulatedPlace**

→

10,000 articles in at least 6 languages: BE, DE, EN, PL, RU, UK

→



385 articles with only one featured article in one language

→

100 articles for manual verification of the timeliness and correctness of attribute **Population**

Precision 60%

Quality flaws

- additional hint about quality can be provided by additional templates for pointing quality flaws in articles

The ten quality flaws with the highest estimated flaw ratios in the English Wikipedia :

- Unreferenced
- Citation needed
- Orphan
- Refimprove
- Dead link
- BLP sources
- Multiple issues
- Empty section
- Notability
- No footnotes

[Maik Anderka, Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia, 2013]

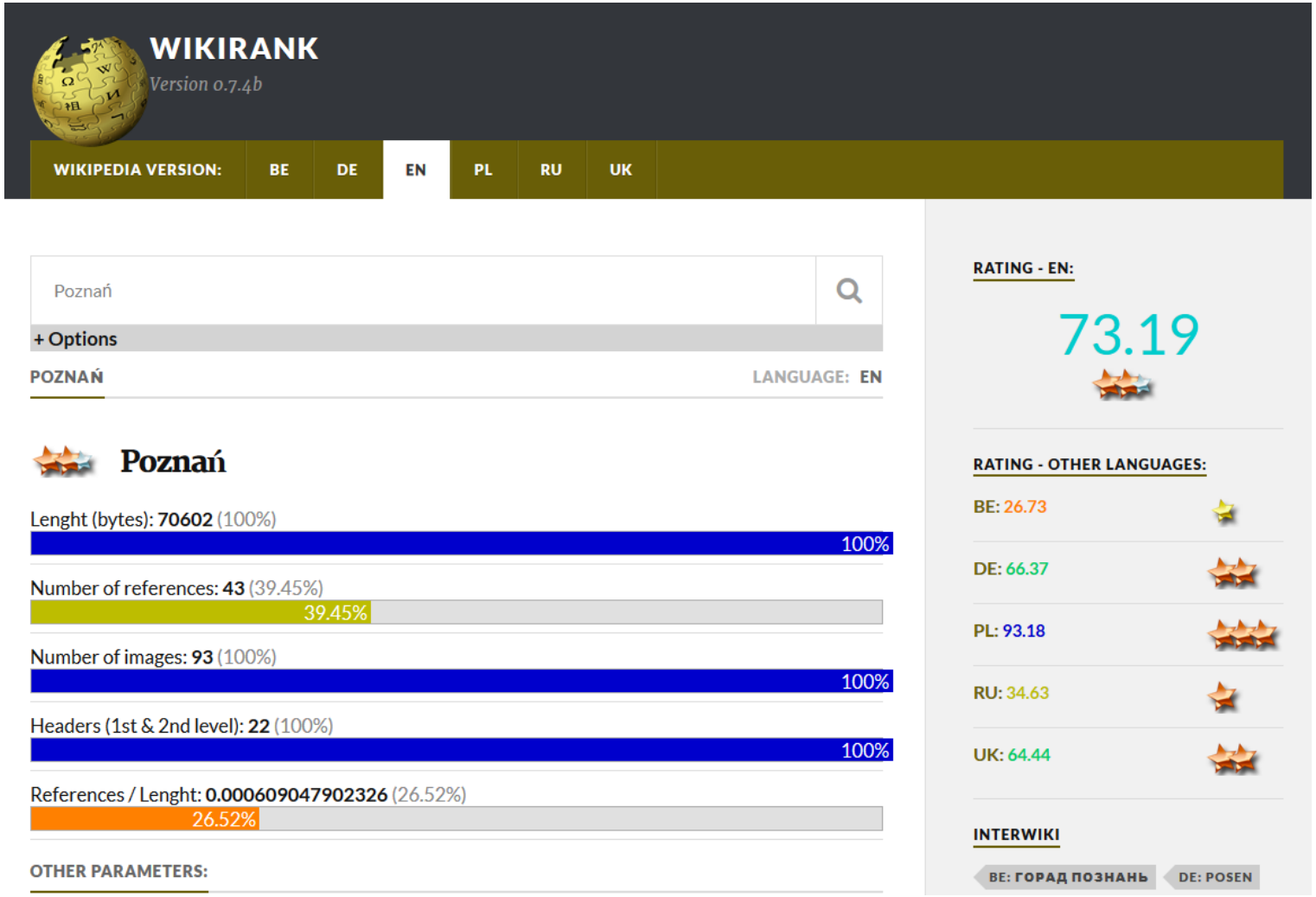
This article's *Criticism* or *Controversy* section **may compromise the article's neutral point of view of the subject**. Please integrate the section's contents into the article as a whole, or rewrite the material.

This article **may require cleanup to meet Wikipedia's quality standards**. No cleanup reason has been specified. Please help improve this article if you can; the talk page may contain suggestions.

This article has multiple issues. Please help **improve it** or discuss these issues on the **talk page**. [hide]

- The **neutrality of this article is disputed**. *(May 2009)*
- This article includes a list of references, related reading or external links, but **its sources remain unclear because it lacks inline citations**. *(Aug 2010)*
- This article **may contain improper references to self-published sources**. *(March 2011)*
- This biographical article **needs additional citations for verification**. *(March 2012)*

WikiRank



Screenshot of the WebSite http://wikirank.net

Workplan

- identification of the most significant technical characteristics of articles
- importance of the attribute
- timeliness of the attribute (correctness + promptness)
- authors’ reputation
- quality flaws
- modelling dependencies
- relative quality measure