

Enrichment of Multilingual Wikipedia Based on Quality Analysis

Włodzimierz Lewoniewski,
Poznań University of Economics and Business

Agenda

- Introduction
- Quality in Wikipedia
- Automatic Assessment of the Quality of Wikipedia Articles
- Quality Measures and Dimensions of Wikipedia Articles
- Building Quality Models for Automatic Quality Assessment
- Quality of Infoboxes
- Enrichment of Wikipedia
- Future Work



Introduction

- **Department of Information Systems**

(DIS) belongs to the Faculty of Informatics and Electronic Economy, which is acknowledged as outstanding by the Accreditation Committee by Polish Ministry of Science and Higher Education.



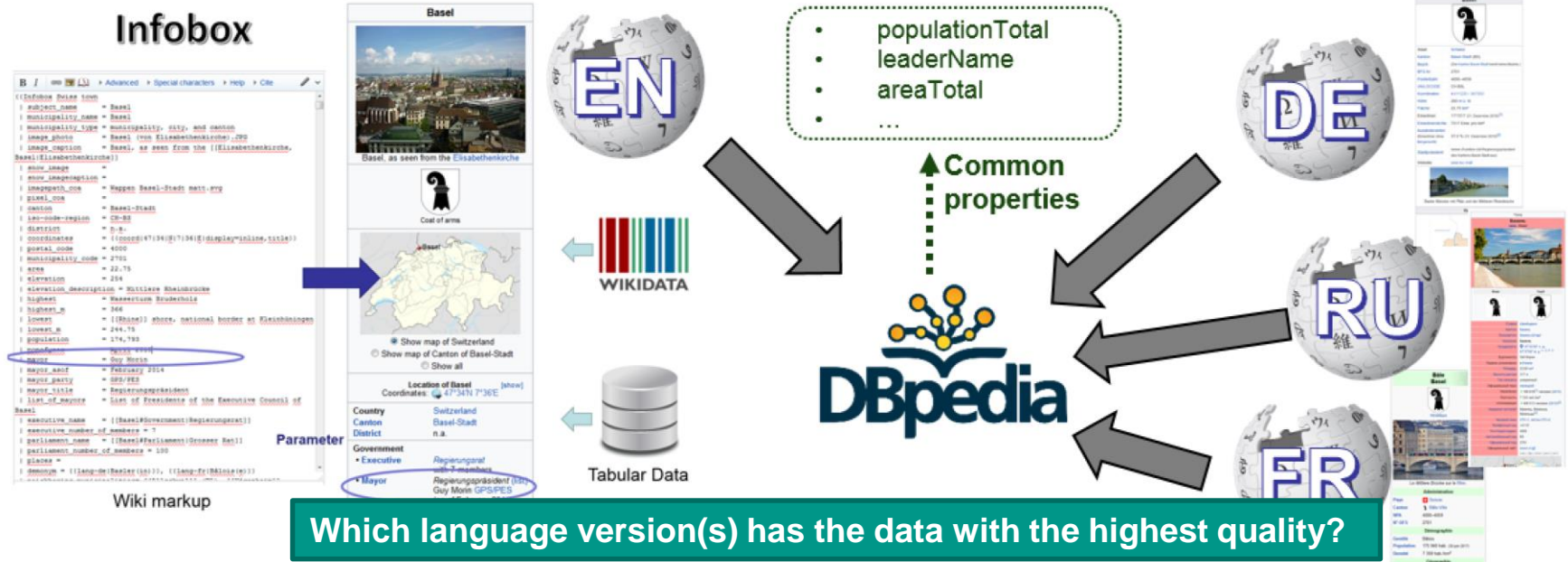
kie.ue.poznan.pl

- Head of the department: prof. **Witold Abramowicz**



POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Multilinguality of Wikipedia





Which language version(s) has the data with the highest quality?

Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2017). *Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. In Informatics (Vol. 4, No. 4, p. 43). Multidisciplinary Digital Publishing Institute.*

Quality of Articles

- Wikipedia articles can get quality grades from users.
- There are differences between grading schemes in language versions

Colors    are marked grades that have similar characteristics

| Grade / Language | BE 157,645 | DE 2,224,246 | EN 5,725,625 | FR 2,044,199 | PL 1,301,888 | RU 1,499,847 | UK 825,041 |
|---|---------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------------|
|  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 development | | | | | | X | X |
| Start | | | X | X | X | | |
| Stub | X | | X | X | X | X | X |
| Unassessed | 99,24% | 99,71% | 18,36% | 40,06% | 99,64% | 85,01% | 95,63% |

Automatic Assessment of the Quality of Wikipedia Articles

- It is possible to build models for quality assessment of Wikipedia articles based on different measures using data mining algorithms.
- There are different approaches, which use various measures and algorithms to assess quality of articles.



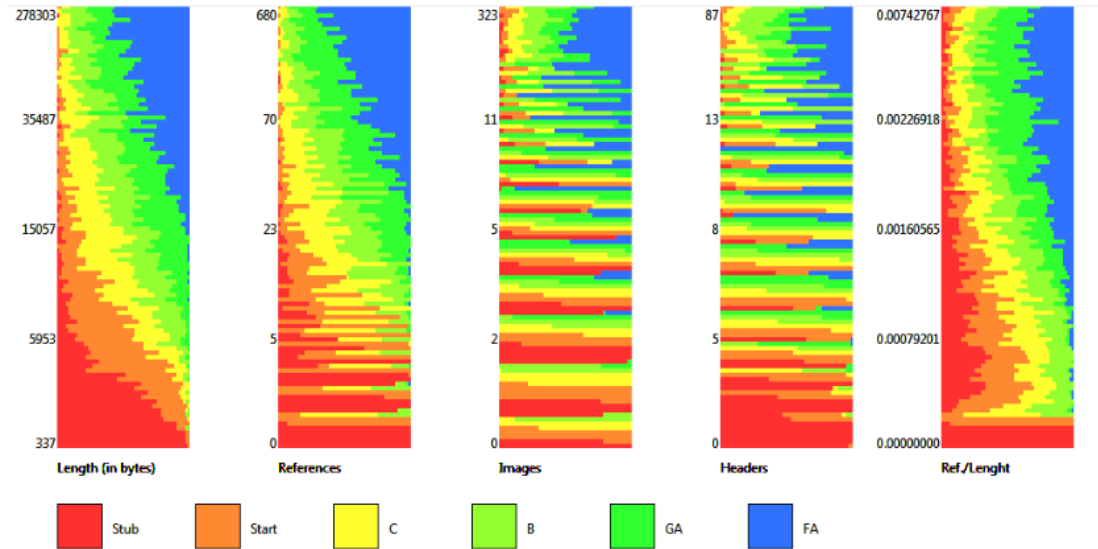
Related Work

- Most of the works focus on English Wikipedia
- One of the first studies showed that longer articles in Wikipedia often have higher quality grades (Blumenstock 2008).
- Often the best articles have more images, sections, use bigger number of references than articles with lower quality (Warncke-Wang et al., 2013; Węcel et al., 2015; Lewoniewski et al., 2016).
- Characteristics related to and edition history can also help to predict articles quality in Wikipedia (Dalip et al., 2014; Suzuki et al., 2016; Dang et al., 2016)



Measures Distribution

- Often we can observe a positive correlation between the article quality and the value of each measures.
- Figure show distribution of articles measures of each quality class in English Wikipedia.
- To build this chart we use randomly chosen 1000 articles from each quality class.

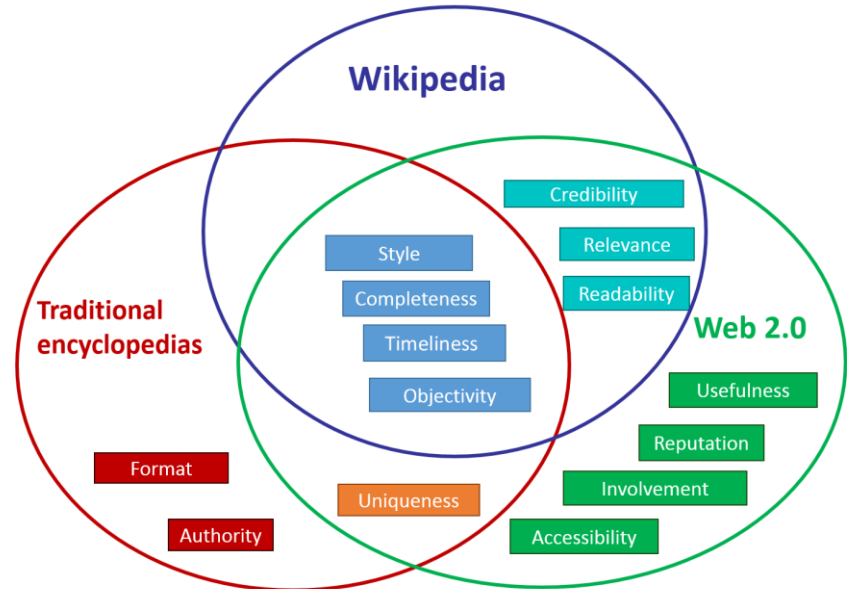


Source: Lewoniewski, W., Węcel, K. (2017). *Relative quality assessment of Wikipedia articles in different languages using synthetic measure*. In *International Conference on Business Information Systems* (pp. 282-292). Springer, Cham.

Quality Dimensions

Some of the measures related to dimensions:

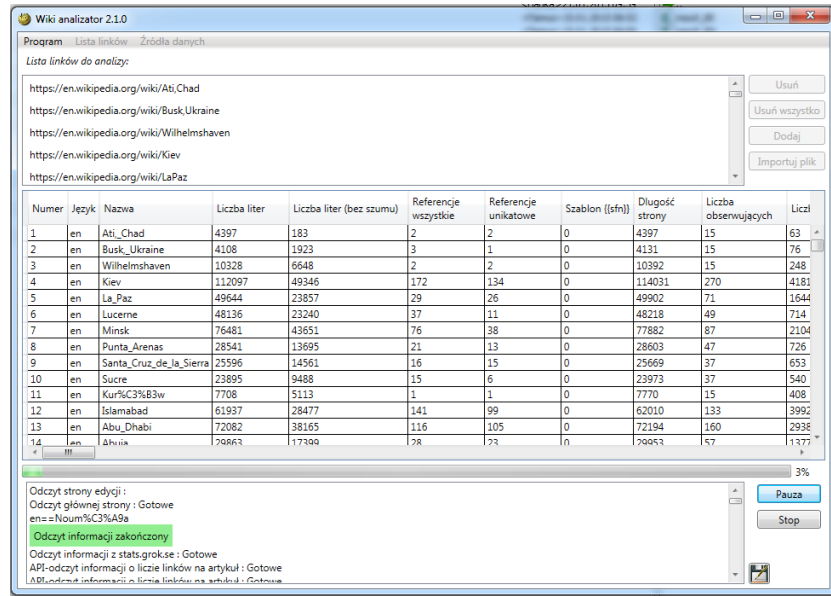
- Number of references (Credibility)
- Articles length (Completeness)
- Number of unique authors (Objectivity, Relevance)
- Automated Readability Index (Readability)
- Articles age (Timeliness, Relevance)
- Number of the sections (Style)
- Citation templates (Credibility, Completeness)
- Many more ...



Source: Lewoniewski, W. (2018). *Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia*. 21st International Conference on Business Information Systems, Berlin. (in press)

Measures Extraction

- We used different sources to extract measures for Wikipedia articles.
- Most of the measures are extracted from Wikimedia dump files
- We developed various applications to obtain measures (over 200) for all or selected articles in different language editions of Wikipedia



| Numer | Język | Nazwa | Liczba liter | Liczba liter (bez szumu) | Referencje wszystkie | Referencje unikatowe | Szablon {{sfn}} | Długość strony | Liczba obserwujących | Licz |
|-------|-------|-------------------------|--------------|--------------------------|----------------------|----------------------|-----------------|----------------|----------------------|------|
| 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 | Ahuac | 20663 | 17306 | 78 | 73 | 0 | 20663 | 47 | 1377 |

Database Dumps

- **enwiki-latest-pages-meta-current.xml.bz2**: recombine all pages (including articles), current versions only. This file is used for obtaining a majority of the articles measures.
- **enwiki-latest-pages-articles.xml.bz2**: consist articles, templates, media/file descriptions, and primary meta-pages. Can be used also for obtaining a majority of the articles measures (excluding statistics from discussion pages).
- **enwiki-latest-pagelinks.sql.gz**: wiki page-to-page link records. Used for network measures - for example incoming links from other articles.
- **enwiki-latest-categorylinks.sql.gz**: wiki category membership link records. Can be used for category count measure.
- **enwiki-latest-externallinks.sql.gz**: wiki external URL link records. can be used for external link count measure.
- **enwiki-latest-imagelinks.sql.gz**: wiki media/files usage records. Can be used to image count measure.
- **enwiki-latest-stub-meta-history.xml.gz**: contain only historical revision metadata. Can be used to extract number of the editors from different groups (bots, anonymous users, administartors etc.) and also number of the edits of various types (e.g. minor edits, edits comments).
- **enwiki-latest-iwlinks.sql.gz**: Interwiki link tracking records. Can be used to extract number of the unique internal links (links to other Wikipedia articles).
- **enwiki-latest-templatelinks.sql.gz**: Wiki template inclusion link records. Used for templates count measure, also it is possible to check if article has infobox
- **enwiki-latest-page.sql.gz**: base per-page data (id, title, old restrictions, etc). Can be used to extract last edit time, page length in bytes.
- Other....

Building the Models



- Quality of articles can be measured using features related to:
 - **Content:** text length, number of images, sections, references and others.
 - **Editors:** reputation, network of the users, comparison of edits and others.
- Quality can be measured as the probability of belonging to one of the specific classes (groups).



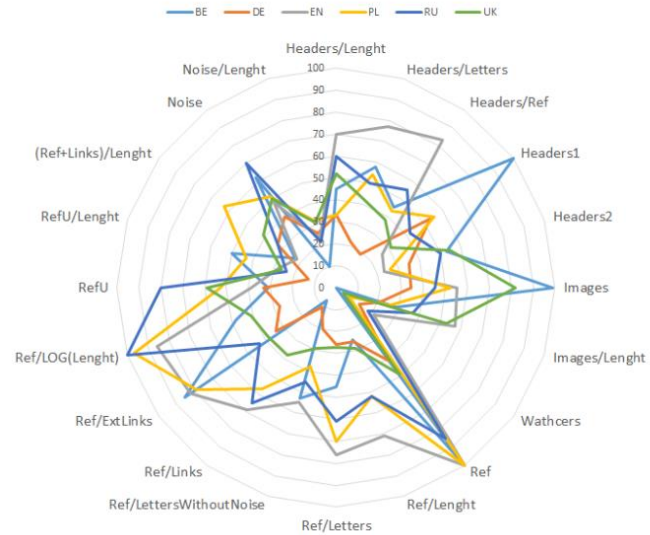
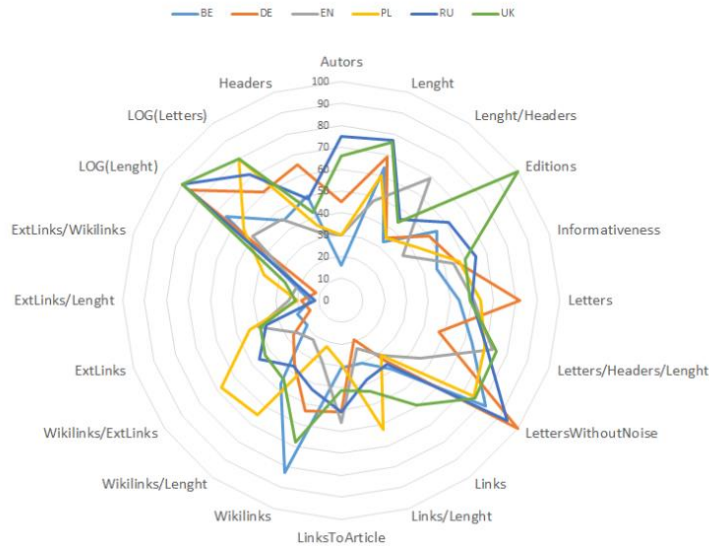
Binary Classification

Some of the approaches divide articles into two groups:

- **Complete:** articles with the highest quality grades (FA, GA)
- **Incomplete:** other articles, which have lower quality grades

| Grade / Language | BE 157,645 | DE 2,224,246 | EN 5,725,625 | FR 2,044,199 | PL 1,301,888 | RU 1,499,847 | UK 825,041 |
|---|---------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------------|
|  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 development | | | | | | X | X |
| Start | | | X | X | X | | |
| Stub | X | | X | X | X | X | X |
| Unassessed | 99,24% | 99,71% | 18,36% | 40,06% | 99,64% | 85,01% | 95,63% |

Measures Importance



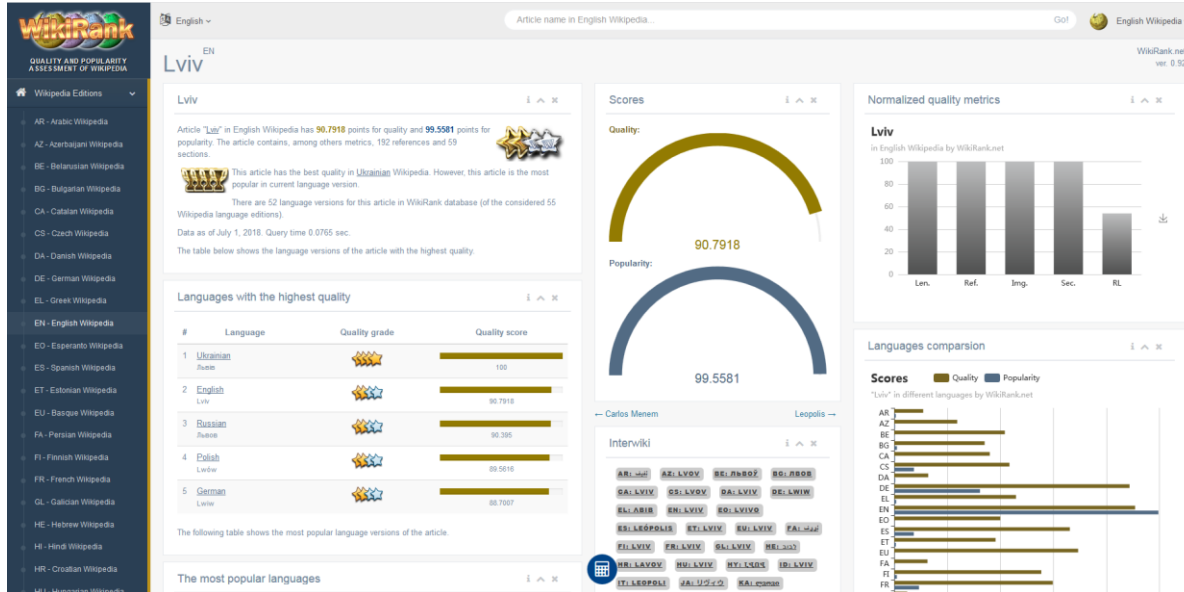
Source: Węcel, K., Lewoniewski, W. (2015). *Modelling the quality of attributes in Wikipedia infoboxes*. In *International Conference on Business Information Systems* (pp. 308-320). Springer, Cham.

Extended Assessment

- To build a model we can use more than two categories, depending on quality grades
 - Number of categories can be different in each language
- ORES score
 - Only for selected language versions
- Synthetic Measure
 - For all Wikipedia languages that have the highest grade (FA equivalent)



WikiRank



EN Lviv

Article "Lviv" in English Wikipedia has **90.7918** points for quality and **99.5581** points for popularity. The article contains, among others metrics, 192 references and 59 sections.

This article has the best quality in **Ukrainian** Wikipedia. However, this article is the most popular in current language version.

There are 52 language versions for this article in WikiRank database (of the considered 55 Wikipedia language editions).

Data as of July 1, 2018. Query time 0.0765 sec.

The table below shows the language versions of the article with the highest quality.

| # | Language | Quality grade | Quality score |
|---|------------------------------------|---------------|---------------|
| 1 | Ukrainian Львів | | 100 |
| 2 | English Lviv | | 90.7918 |
| 3 | Russian Львов | | 90.395 |
| 4 | Polish Lwów | | 89.5616 |
| 5 | German Lwiv | | 88.7507 |

The following table shows the most popular language versions of the article.

The most popular languages

| # | Language | Popularity grade | Popularity score |
|---|------------------------------------|------------------|------------------|
| 1 | English Lviv | | 99.5581 |
| 2 | Russian Львов | | 97.7383 |
| 3 | Polish Lwów | | 70.3037 |
| 4 | Ukrainian Львів | | 60.5037 |
| 5 | German Lwiv | | 32.3459 |

Scores

Quality: 90.7918
Popularity: 99.5581

Normalized quality metrics

Lviv in English Wikipedia by WikiRank.net

| Metric | Score |
|--------|-------|
| Len. | 100 |
| Ref. | 100 |
| Img. | 100 |
| Sec. | 100 |
| RL | 55 |

Languages comparison

Scores in different languages by WikiRank.net

| Language | Quality | Popularity |
|----------|---------|------------|
| AR | ~10 | ~10 |
| AZ | ~10 | ~10 |
| BE | ~10 | ~10 |
| BS | ~10 | ~10 |
| CA | ~10 | ~10 |
| CS | ~10 | ~10 |
| DA | ~10 | ~10 |
| DE | ~10 | ~10 |
| EL | ~10 | ~10 |
| EN | 90.7918 | 99.5581 |
| EO | ~10 | ~10 |
| ES | ~10 | ~10 |
| ET | ~10 | ~10 |
| EU | ~10 | ~10 |
| FA | ~10 | ~10 |
| FI | ~10 | ~10 |
| FR | ~10 | ~10 |

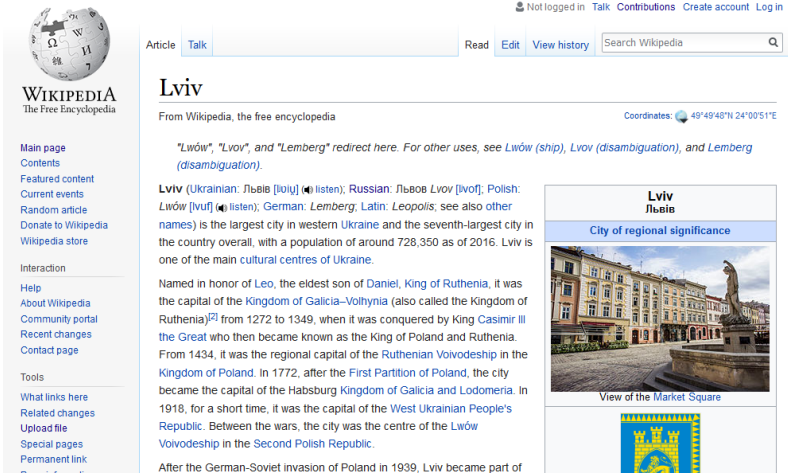
The most popular languages

| # | Language | Popularity grade | Popularity score |
|---|------------------------------------|------------------|------------------|
| 1 | English Lviv | | 99.5581 |
| 2 | Russian Львов | | 97.7383 |
| 3 | Polish Lwów | | 70.3037 |
| 4 | Ukrainian Львів | | 60.5037 |
| 5 | German Lwiv | | 32.3459 |

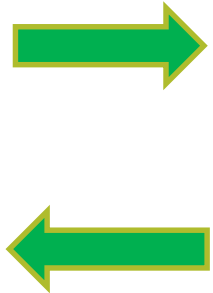
Source: <https://wikirank.net/en/Lviv>

Article and Infobox Quality

- Completeness
- Credibility
- Objectivity
- Readability
- Relevance
- Style
- Timeliness
- ...



The screenshot shows the Wikipedia article for Lviv. At the top, it says "Not logged in" with links for "Talk", "Contributions", "Create account", and "Log in". Below that is a search bar and navigation tabs for "Article" and "Talk". The title "Lviv" is prominently displayed, followed by a coordinate box. A red text block indicates that "Lviv", "Lvov", and "Lemberg" redirect here. The main text begins with "Lviv (Ukrainian: Львів [lɔjɪ] ⓘ listen); Russian: Львов Lvov [lʲɔvɔf]; Polish: Lwów [lɔvɔf] ⓘ listen); German: Lemberg; Latin: Leopoldis; see also other names) is the largest city in western Ukraine and the seventh-largest city in the country overall, with a population of around 728.350 as of 2016. Lviv is one of the main cultural centres of Ukraine." Below the text is a "View of the Market Square" image and a coat of arms. The infobox on the right contains a title, a description, a map, and various data fields.




The infobox for Lviv includes a title "Lviv", a description "City of regional significance", a map, and a table of data. The table contains fields for Country, Language, City status, Population, Area, and other relevant information.

- Completeness
- Credibility
- Relevance
- Timeliness
- ...

Source: Lewoniewski, W. (2018). *Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia*. 21st International Conference on Business Information Systems, Berlin. (in press)

Simple Measures for Infobox

Number
of filled
parameters

8



| | |
|--------------|--|
| Directed by | Juliusz Machulski |
| Written by | Juliusz Machulski ^[1] |
| Starring | Jan Machulski Witold Pyrkosz Leonard Pietraszak Jacek Chmielnik Krzysztof Kiersznowski |
| Music by | Henryk Kuźniak |
| Release date | 1981 |
| Running time | 108 minutes ^[2] |
| Country | Poland |
| Language | Polish ^[2] |

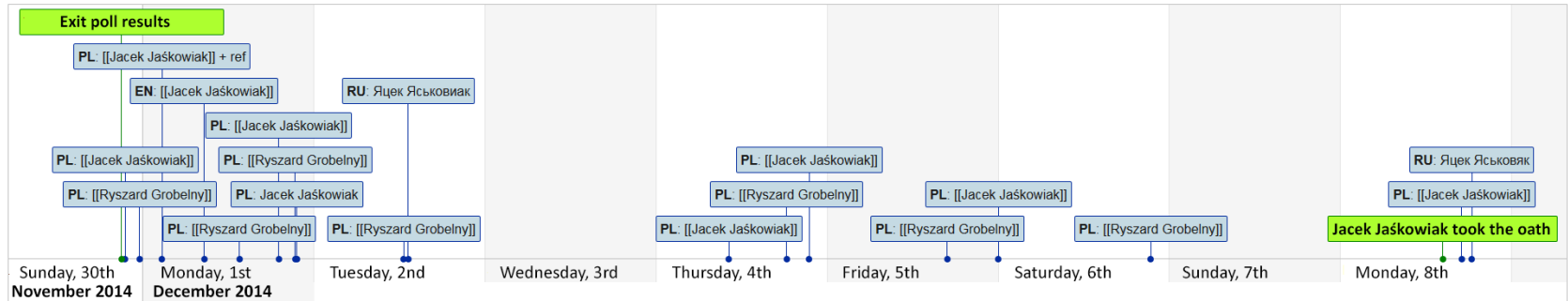
Number of
(unique)
references

3 (2)



Infobox Timeliness

- Timeliness measures can be related to currency and volatility of the infoboxes.
- Example: history of changes in the "leader name" parameter of the Poznań infobox



Source: Lewoniewski, W. (2018). *Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia*. 21st International Conference on Business Information Systems, Berlin. (in press)




Correlation of Measures related to Articles and Infoboxes

| | | Articles measures | | | | | | | | | | | | Infobox measures | | | | | | | | | | | | | | | | |
|-------------------|-----------------------|-----------------------|---------------|-------------|------------------|-------------|-----------|-----------|---------------------|------------|------------------|-------------|-----------|-----------------------|----------------|--------------------|------------------|-------------|-----------|----------------------------|---------------------------|------------|---------------------|----------------------------|--------------------|-------------------|--------------|-----------|-----------------------|-----------|
| | | Completeness | | | | Reliability | | | | Timeliness | | | | Completeness | | | | Reliability | | | | Timeliness | | | | | | | | |
| | | Means | Std.Dev. | Text length | Templa-tes count | Int. links | Images | Sections | Length without ref. | Ref. Count | Ref./text length | Ref. length | Last edit | Last edit (not a bot) | infobox length | Filled param-eters | Templa-tes count | Int. links | Images | Length of param-eter value | Length of value min. ref. | Ref. Count | Ref./infobox length | Ref./infobox param. length | Max. refs. on par. | Avg. ref. on par. | Ref. length. | Last edit | Last edit (not a bot) | |
| Articles measures | Completeness | Text length | 2.7702790E+01 | 30635 | 1.030030 | 0.015173 | 0.307812 | 0.757237 | 0.369460 | 0.966023 | 0.045134 | 0.045014 | 0.745113 | 0.177760 | 0.282328 | 0.370510 | 0.420869 | 0.301848 | 0.251058 | 0.405884 | 0.390345 | 0.230404 | 0.120254 | 0.203650 | 0.046214 | 0.136432 | 0.198950 | 0.021830 | 0.258521 | |
| | | Templa-tes count | 3.051441E+01 | 43 | 0.815173 | 1.030030 | 0.752804 | 0.758741 | 0.691985 | 0.817124 | 0.748092 | 0.128492 | 0.086557 | 0.133220 | 0.220753 | 0.307355 | 0.367134 | 0.075882 | 0.543281 | 0.218555 | 0.310052 | 0.250984 | 0.168952 | 0.284854 | -0.045740 | -0.013497 | 0.087732 | 0.191416 | -0.078858 | 0.434798 |
| | | Int. links | 1.820380E+02 | 218 | 0.837012 | 0.732694 | 0.800360 | 0.763353 | 0.801224 | 0.807630 | 0.748453 | 0.020814 | 0.030258 | 0.172824 | 0.263460 | 0.388752 | 0.415712 | 0.103705 | 0.287753 | 0.298205 | 0.436218 | 0.418227 | 0.249051 | 0.141014 | 0.200034 | 0.046700 | 0.154300 | 0.216187 | 0.026330 | 0.263338 |
| | | Images | 1.073564E+01 | 17 | 0.757237 | 0.758741 | 0.783833 | 1.000303 | 0.881830 | 0.783773 | 0.597040 | -0.029786 | 0.455009 | 0.097188 | 0.200369 | 0.320215 | 0.519531 | 0.068274 | 0.277802 | 0.246720 | 0.354340 | 0.342884 | 0.184580 | 0.285222 | -0.009585 | 0.035409 | 0.113038 | 0.203615 | -0.239528 | 0.291448 |
| | | Sections | 2.166529E+01 | 16 | 0.693480 | 0.694935 | 0.831224 | 0.691830 | 1.069360 | 0.673633 | 0.693462 | 0.026883 | 0.009253 | 0.197181 | 0.331168 | 0.346214 | 0.441264 | 0.199251 | 0.281127 | 0.310647 | 0.398412 | 0.338547 | 0.265472 | 0.115163 | 0.034115 | 0.094108 | 0.116486 | 0.178340 | 0.068222 | 0.603117 |
| | Reliability | Length without ref. | 2.870869E+04 | 29286 | 0.939025 | 0.817124 | 0.937838 | 0.783773 | 0.873872 | 1.000000 | 0.629585 | 0.032123 | 0.719556 | 0.179313 | 0.294469 | 0.377805 | 0.440539 | 0.028287 | 0.487440 | 0.280704 | 0.437486 | 0.391585 | 0.250479 | 0.134861 | -0.022989 | 0.047409 | 0.136385 | 0.220388 | 0.028451 | 0.800759 |
| | | Ref. Count | 2.461810E+01 | 38 | 0.840514 | 0.748092 | 0.748092 | 0.597340 | 0.664102 | 0.629495 | 1.000000 | 0.283030 | 0.782389 | 0.440355 | 0.260366 | 0.275307 | 0.361901 | 0.048405 | 0.274894 | 0.238836 | 0.294144 | 0.282167 | 0.198943 | 0.124376 | 0.015505 | 0.043677 | 0.118734 | 0.140071 | 0.034026 | 0.448454 |
| | | Ref./text length | 3.428219E-04 | 8 | 0.045814 | 0.126492 | -0.028914 | -0.029786 | -0.029963 | 0.032123 | 0.593000 | 1.001030 | 0.025922 | 0.015886 | -0.049212 | -0.212420 | 0.038752 | -0.003788 | 0.011594 | 0.011150 | -0.039592 | -0.027140 | 0.038761 | 0.265328 | 0.084913 | 0.026777 | 0.035136 | -0.013532 | 0.075382 | 0.048632 |
| | | Ref. length | 1.041610E+03 | 1841 | 0.745017 | 0.698957 | 0.698298 | 0.450368 | 0.569853 | 0.712683 | 0.038672 | 1.003030 | 0.088985 | 0.181455 | 0.292828 | 0.297289 | 0.078184 | 0.032426 | 0.182759 | 0.274735 | 0.268847 | 0.167315 | 0.167820 | 0.003887 | 0.015771 | 0.034445 | 0.108339 | 0.023803 | 0.338532 | |
| | | Last edit | 1.479503E+09 | 1072538 | 0.177736 | 0.133220 | 0.173204 | 0.397188 | 0.197151 | 0.179313 | 0.148856 | 0.015885 | 0.166995 | 0.003063 | 0.117803 | 0.114628 | 0.158722 | 0.001534 | 0.093563 | 0.069191 | 0.121776 | 0.114897 | 0.117440 | 0.168900 | 0.071296 | 0.015292 | 0.087150 | 0.068791 | 0.199550 | 0.213322 |
| Timeliness | Last edit (not a bot) | 1.479503E+09 | 4816890 | 0.262239 | 0.278783 | 0.263488 | 0.269388 | 0.331108 | 0.284489 | 0.208836 | 0.048581 | 0.181455 | 0.311983 | 0.303180 | 0.389405 | 0.167889 | 0.058486 | 0.127131 | 0.133811 | 0.118332 | 0.118830 | 0.043748 | 0.018477 | 0.028302 | 0.022253 | 0.037374 | 0.035115 | 0.034811 | 0.447236 | |
| | infobox length | 1.771798E+03 | 210 | 0.378515 | 0.337355 | 0.299752 | 0.328215 | 0.348214 | 0.377605 | 0.275387 | -0.012420 | 0.255895 | 0.186828 | 0.398955 | 0.280203 | 0.167889 | 0.009088 | 0.492448 | 0.133065 | 0.396059 | 0.368091 | 0.670400 | 0.434758 | 0.112600 | 0.030704 | 0.014500 | 0.027991 | 0.221835 | | |
| | Filled param-eters | 3.021216E+01 | 4 | 0.838989 | 0.365134 | 0.415112 | 0.318651 | 0.441264 | 0.443939 | 0.391893 | 0.038752 | 0.186732 | 0.182988 | 0.183781 | 1.003030 | 1.003030 | 0.183019 | 0.308485 | 0.218901 | 0.241634 | -0.018185 | 0.382374 | 0.173181 | 0.147696 | 0.026917 | 0.031188 | 0.018588 | 0.178541 | | |
| | Templa-tes count | 1.003249E+01 | 1 | 0.002849 | 0.076982 | 0.105795 | 0.068274 | 0.268856 | 0.002397 | 0.046450 | -0.003798 | 0.078144 | 0.081534 | 0.058346 | 0.145517 | 0.002006 | 0.068114 | 0.035777 | 0.482954 | 0.523240 | 0.424553 | 0.149105 | 0.004743 | 0.484003 | 0.264444 | -0.022179 | 0.076210 | | | |
| | Int. links | 2.781898E+03 | 2 | 0.818988 | 0.843691 | 0.818753 | 0.771882 | 0.287727 | 0.348547 | 0.214894 | 0.011594 | 0.030426 | 0.389353 | 0.727132 | 0.460448 | 0.183978 | 0.232816 | 1.003030 | 0.417087 | 0.488736 | 0.468602 | 0.252880 | 0.127887 | 0.104813 | 0.017894 | 0.158874 | 0.038478 | 0.178823 | | |
| Infobox measures | Completeness | Images | 2.204767E+00 | 1 | 0.278956 | 0.210955 | 0.266365 | 0.245763 | 0.318047 | 0.289704 | 0.001110 | 0.132229 | 0.069139 | 0.132911 | 0.102566 | 0.308495 | -0.091184 | 0.074087 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -0.000000 | -0.000000 | -0.000000 | -0.000000 | -0.000000 | 0.229973 |
| | | Length of param-eters | 4.388318E+02 | 208 | 0.405884 | 0.318982 | 0.438710 | 0.358438 | 0.388412 | 0.401488 | 0.284444 | -0.038942 | 0.274725 | 0.191710 | 0.115352 | 0.889129 | 0.108473 | 0.805777 | 0.488836 | 0.189221 | 1.000000 | 0.878810 | 0.718222 | 0.525784 | 0.172985 | 0.378885 | 0.086233 | 0.418837 | 0.025018 | 0.245883 |
| | | Length of par. value | 6.025279E+02 | 192 | 0.390545 | 0.289904 | 0.418227 | 0.343804 | 0.388547 | 0.391592 | 0.828167 | -0.007148 | 0.288047 | 0.114897 | 0.190000 | 0.668019 | 0.200501 | 0.482854 | 0.398905 | 0.171235 | 0.879516 | 0.000000 | 0.663165 | 0.472309 | 0.047269 | 0.028207 | 0.049360 | 0.028207 | 0.026230 | 0.263433 |
| | | Ref. Count | 1.288448E+03 | 1 | 0.218404 | 0.148856 | 0.244861 | 0.188488 | 0.289472 | 0.232439 | 0.181913 | 0.038181 | 0.166718 | 0.117440 | 0.043249 | 0.010403 | 0.548386 | 0.523548 | 0.282890 | 0.035796 | 0.710232 | 0.853788 | 1.000000 | 0.312884 | 0.128839 | 0.012283 | 0.414530 | 0.051448 | 0.158881 | 0.133388 |
| | | Ref./infobox length | 7.862633E-04 | 8 | 0.135324 | 0.054984 | 0.141214 | 0.095252 | 0.115163 | 0.134561 | 0.132426 | 0.005296 | 0.037830 | 0.060310 | 0.037487 | 0.454750 | -0.111610 | 0.424533 | 0.122897 | 0.033449 | 0.326754 | 0.472309 | 0.847056 | 0.000000 | 0.663889 | 0.140084 | 0.014924 | 0.018262 | 0.028630 | 0.058447 |
| | Reliability | Ref./infobox param | 7.862633E-03 | 8 | 0.028838 | -0.045740 | -0.038924 | -0.060386 | -0.054115 | -0.028969 | 0.015885 | 0.054913 | -0.003892 | 0.011286 | 0.016053 | 0.119800 | -0.062374 | 0.148105 | -0.104983 | -0.188884 | 0.157985 | 0.074928 | 0.378233 | 0.884484 | 0.863484 | 0.425881 | 0.728879 | 0.328189 | 0.110825 | -0.315216 |
| | | Max. refs. on par. | 9.945029E-01 | 8 | 0.046214 | -0.013497 | 0.047780 | 0.035499 | 0.304103 | 0.647400 | 0.041677 | 0.036777 | 0.015774 | 0.860329 | -0.327353 | 0.241387 | 0.113915 | -0.004743 | 0.071694 | 0.015060 | 0.312896 | 0.268207 | 0.414330 | 0.110384 | 0.423561 | 1.000000 | 0.491849 | 0.225465 | 0.148475 | 0.026330 |
| | | Avg. ref. on par. | 4.529307E-02 | 8 | 0.198432 | 0.087232 | 0.154388 | 0.119388 | 0.118488 | 0.118734 | 0.035398 | 0.163445 | 0.087150 | 0.081624 | 0.067384 | 0.067384 | 0.488408 | 0.390233 | 0.648688 | 0.385448 | 0.918588 | 0.365448 | 0.818588 | 0.729679 | 0.615989 | 1.000000 | 0.484785 | 0.088781 | 0.044883 | |
| | | Ref. length | 3.720335E+01 | 43 | 0.199953 | 0.191416 | 0.216187 | 0.200612 | 0.179343 | 0.202083 | 0.143071 | -0.013532 | 0.038330 | 0.069781 | 0.030115 | 0.143450 | 0.076617 | 0.265444 | 0.156674 | 0.001521 | 0.418037 | 0.222211 | 0.519861 | 0.451232 | 0.323109 | 0.040793 | 0.016080 | 0.028216 | 0.11344 | |
| | | Last edit | 1.478393E+09 | 11385306 | 0.028936 | -0.078888 | -0.068303 | -0.293816 | -0.088222 | 0.028481 | 0.034005 | 0.015885 | 0.028370 | 0.169899 | -0.034511 | 0.028791 | -0.101109 | -0.020178 | -0.008478 | -0.008436 | 0.026818 | 0.026290 | 0.071038 | 0.088658 | 0.110325 | 0.143445 | 0.089781 | 0.028214 | 0.036081 | 0.133551 |
| Timeliness | Last edit (not a bot) | 1.362110E+09 | 45661875 | 0.096531 | 0.424798 | 0.363038 | 0.391440 | 0.607317 | 0.603759 | 0.448434 | 0.094632 | 0.356672 | 0.210322 | 0.447236 | 0.221835 | 0.376541 | 0.076219 | 0.178063 | 0.323873 | 0.234383 | 0.226343 | 0.133966 | 0.805447 | 0.631216 | 0.026630 | 0.044563 | 0.11344 | 0.133551 | 0.000000 | |

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.

Infoboxes.net



English Wikipedia


According to the quality and popularity analysis*, the best infobox of the article Lviv is placed in English Wikipedia. This information can be used to improve quality of the relevant articles in less developed Wikipedia language editions and also enrich relevant resources in other popular open knowledge databases such as [DBpedia](#), [Wikidata](#), [YAGO](#) and others†.

Show top 10 langs


| English Wikipedia | | | Russian Wikipedia | | | Polish Wikipedia | | | Ukrainian Wikipedia | | | German Wikipedia | | | French Wikipedia | | | Italian Wikipedia | | |
|----------------------|--|---|-----------------------|--|--|----------------------|--|--|-----------------------|---|--|----------------------|--|--|----------------------|--|---|-------------------------|--|---|
| Article | Quality | Popularity | Article | Quality | Popularity | Article | Quality | Popularity | Article | Quality | Popularity | Article | Quality | Popularity | Article | Quality | Popularity | Article | Quality | Popularity |
| Lviv | 91 | 100 | Львов | 90 | 98 | Lwów | 90 | 70 | Львів | 100 | 61 | Lwiv | 89 | 32 | Lviv | 60 | 9 | Leopoli | 48 | 1 |

Lviv
Львів


City of regional significance



View of the [Market Square](#)



Flag







Coat of arms

Logo


Motto(s): [Сміється Львів](#)

Город
Львов



Льво́в

Lwów
Lwów



Panorama Lwowa w kierunku wschodnim z [wieży ratuszowej](#)

Hasło: **Flaga**

Państwo: 🇺🇦 **Ukraina**

Stożec: **brakowi**

Przez miasto: **1366**

Władca: **Andrii Sadowy**

Powierzchnia: **160,01 km²**



Miejscowość: **296 m n.p.m.**

Populacja (2016): **727 869** (11)

+ liczba ludności

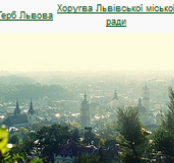
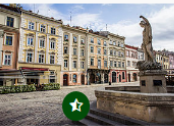
+ gęstość: **4295 os./km²**

Львів

Герб Львова



Хорова Львівської міської ради

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter [Lwiv \(Begriffsklärung\)](#) aufgeführt.

Lemberg ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter [Lemberg \(Begriffsklärung\)](#) aufgeführt.

Lwiv
Львів

Basisdaten

Oblast: **Oblast Lwiv**

Region: **Kreisfreie Stadt**

Höhe: **296 m**

Fläche: **171,01 km²**

Einwohner: **728.545** (1. März 2015)

4.260 Einwohner je km²



Bevölkerungsdichte: **km²**

Postleitzahlen: **79000–79490**

Vorwahl: **+380 322**

Geographische Lage: **49°51′N, 24°10′E**

Lviv
Львів

Państwo: 🇺🇦 **Ukraine**

Subdivision: **Oblast de Lviv**

Maire: **Andri Sadowy**

Code postal: **79000 — 79490**


Indicatif tél.: **+380 322**

Démographie



Population: **729 429 hab.** (2016)

Densité: **4 285 hab./km²**

Géographie



Leopoli
Leopoli

Localizzazione

Stato: 🇺🇦 **Ucraina**

Oblast: **Leopoli**

Distretto: **Non presente**


Amministrazione

Sindaco: **Andrii Sadowy**

Territorio

Coordinate: **49°51′N, 24°10′E** / **49.85°N, 24.01667°E**

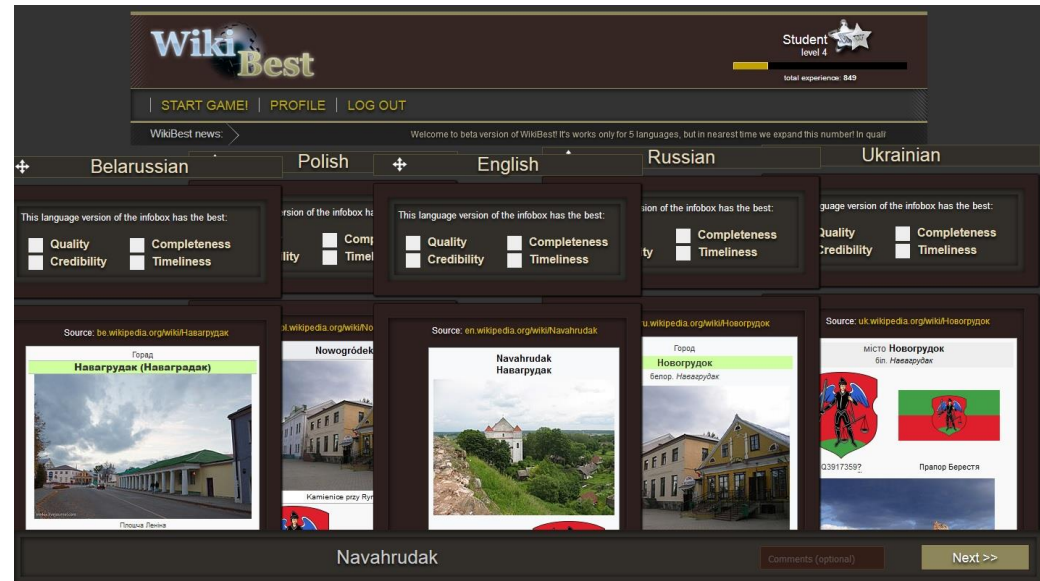
Coordinate:



WikiBest

Users can vote for the best infobox in four nominations:

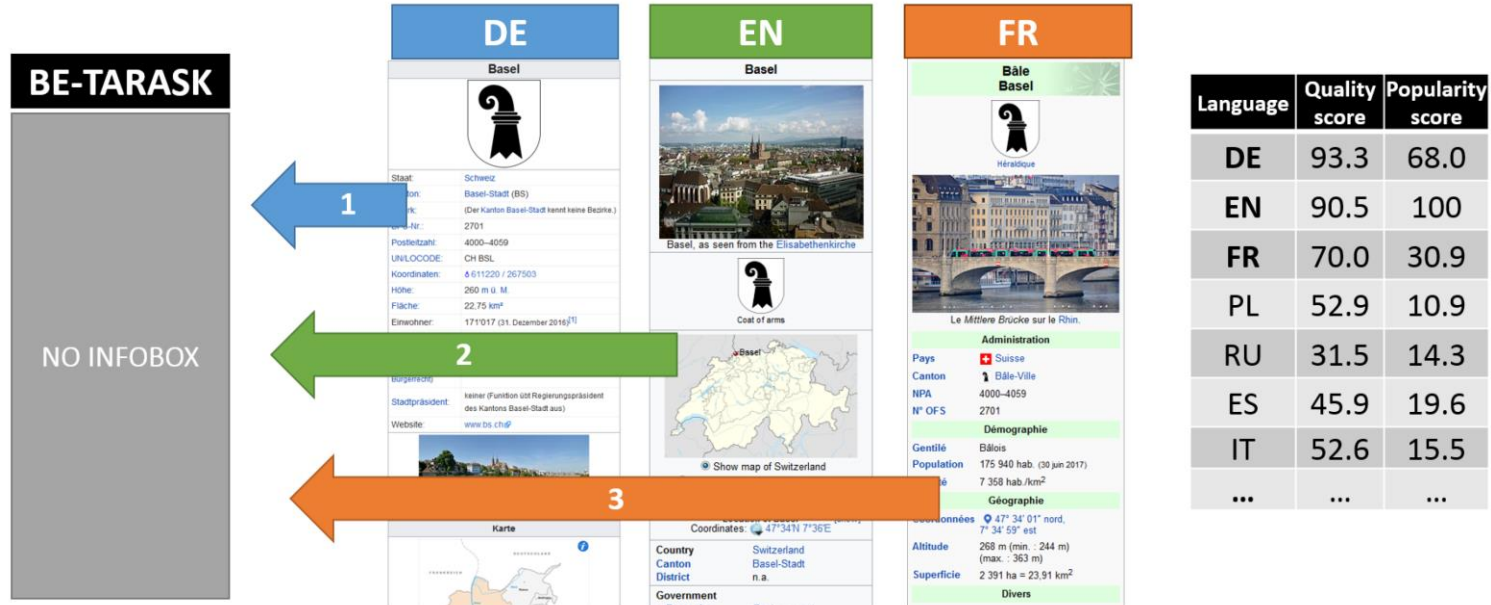
- the best quality
- the best completeness
- the best credibility
- the best timeliness



The screenshot displays the WikiBest voting interface. At the top, it shows the 'WikiBest' logo and a user profile for 'Student level 4' with a total experience of 849. Below this, there are navigation links for 'START GAME!', 'PROFILE', and 'LOG OUT'. The main content area is divided into four columns, each representing a different language: Belarusian, Polish, English, and Ukrainian. Each column contains a voting panel for the 'Navahrudak' infobox. The panels are identical in structure, with a header indicating the language version and four nomination buttons: Quality, Completeness, Credibility, and Timeliness. Below the voting panels, there are four image thumbnails for the 'Navahrudak' infobox, each corresponding to a language. The thumbnails show various views of the town of Navahrudak, including a street view, a building, a hillside, and a flag. The interface also includes a 'Next >>' button at the bottom right.

Source: <https://wikibest.net>

Enrichment of Wikipedia



Potential for New Articles

- Despite the fact that English Wikipedia is the largest, it can be also enriched by other language versions.
- Table below presents potential number of articles in each language and each topic that can be created or enriched using infoboxes from other language version of Wikipedia

| Topic | BE | DE | EN | FR | PL | RU | UK |
|---------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Album | 170 793 | 157 925 | 22 538 | 130 712 | 143 118 | 151 548 | 162 086 |
| Companies | 83 829 | 58 169 | 22 783 | 65 154 | 77 409 | 72 932 | 79 470 |
| Films | 146 876 | 114 263 | 28 355 | 100 265 | 128 189 | 119 812 | 133 739 |
| Universities | 24 325 | 20 273 | 3 420 | 19 804 | 22 298 | 21 728 | 23 072 |
| Video games | 24 325 | 21 184 | 2 924 | 12 953 | 21 245 | 18 559 | 22 917 |

Source: Lewoniewski, W. (2017). *Completeness and Reliability of Wikipedia Infoboxes in Various Languages*. In International Conference on Business Information Systems (pp. 295-305). Springer, Cham.



Future work

- Expanding number of measures (including linguistic) for predicting quality of Wikipedia articles.
- Sentiment analysis of Wikipedia articles.
- Fact extraction from the content in different languages.
- Improving projects related to the quality of Wikipedia.
- Analysis of references measures on various granularity:
 - host level, domain, path and url.
- Detection of language sensitive topics in Wikipedia.



Thank You



POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Additional Information

Unification of Infobox Parameters

| DE | | EN | | PL | | RU | |
|--------------------|------|-----------|-------|--------------|------|-------------|------|
| Video Games | | | | | | | |
| Plattform | 2821 | platforms | 20345 | tytuł | 2926 | заголовок | 2774 |
| Genre | 2777 | genre | 20083 | data wydania | 2873 | разработчик | 2463 |
| Release | 2748 | developer | 20073 | platforma | 2868 | изображение | 2439 |
| Entwickler | 2730 | released | 19762 | producent | 2860 | жанр | 2252 |
| Spielmodi | 2615 | publisher | 19186 | gatunek | 2855 | издатель | 2179 |
| Titel | 2347 | modes | 18653 | tryby gry | 2776 | title | 2112 |
| Sprache | 2300 | title | 18178 | wydawca | 2749 | сайт | 2055 |
| Bedienung | 2269 | image | 17615 | www | 2166 | управление | 2049 |
| Medien | 2185 | caption | 9341 | nośniki | 1995 | developer | 2022 |
| Verleger | 1713 | composer | 6523 | kontrolery | 1635 | genre | 1985 |
| Info | 1335 | designer | 6205 | kategorie | | released | 1804 |
| USK | 1243 | series | 5635 | wiekowe | 1577 | publisher | 1774 |
| PEGI | 1213 | engine | 3412 | wymagania | 1240 | платформы | 1763 |
| Bild | 1123 | producer | 3085 | dystrybutor | 1157 | подпись | 1607 |
| Systemminima | 1044 | director | 2989 | seria gier | 1119 | серия | 1604 |
| | | ... | | ... | | ... | |

computingPlatform

developer

genre



releaseDate








publisher

foaf:name



SEO Measures

Mean of Visibility Index from different countries perspectives:

| | Country | Articles | | | |
|---|----------------|----------|------|------|------|
| | | PFA | UFA | PST | UST |
|  | France | .041 | .000 | .003 | .000 |
|  | Germany | .059 | .000 | .002 | .000 |
|  | Italy | .015 | .000 | .001 | .000 |
|  | Poland | .026 | .000 | .001 | .000 |
|  | Spain | .037 | .000 | .000 | .000 |
|  | United Kingdom | .255 | .000 | .020 | .000 |
|  | United States | .234 | .000 | .020 | .000 |

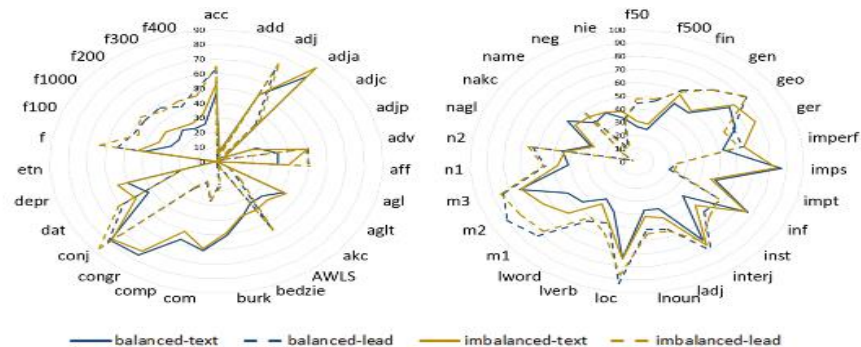
Mean of each social indicators:

| Indicator | Articles | | | |
|-----------|----------|-----|-----|-----|
| | PFA | UFA | PST | UST |
| FB | 2101.7 | 0.0 | 5.6 | 0.0 |
| FBI | 1138.0 | 0.0 | 2.4 | 0.0 |
| FBs | 517.9 | 0.0 | 1.9 | 0.0 |
| FBc | 391.8 | 0.0 | 1.2 | 0.0 |
| TW | 28.1 | 0.0 | 0.4 | 0.0 |
| LI | 16.9 | 0.0 | 0.2 | 0.0 |
| GP | 286.4 | 0.0 | 0.6 | 0.0 |
| PT | 174.5 | 0.0 | 0.0 | 0.0 |



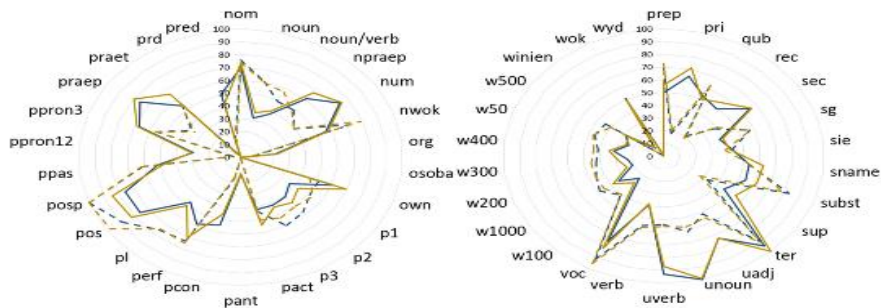
Linguistic Measures

- In Polish Wikipedia we extracted over 100 linguistic measures of articles
- Model shows over 93% classification precision



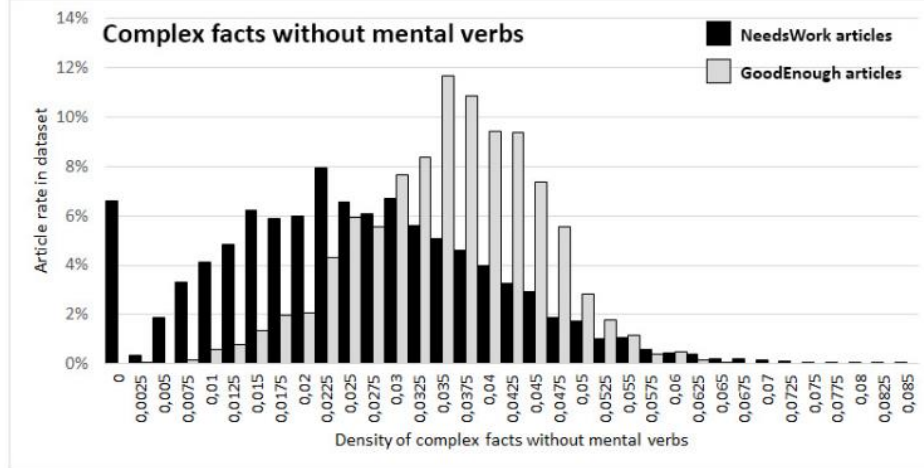
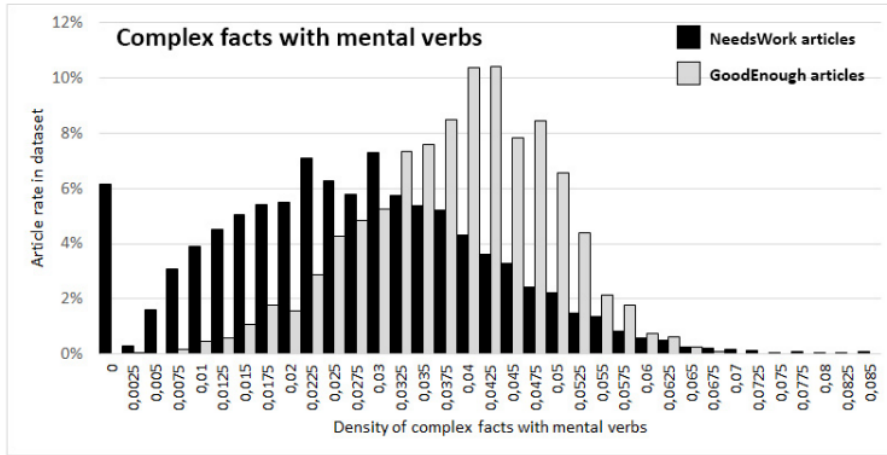
The most important features:

- impersonal verbs,
- third person words,
- unique nouns,
- unique verbs.

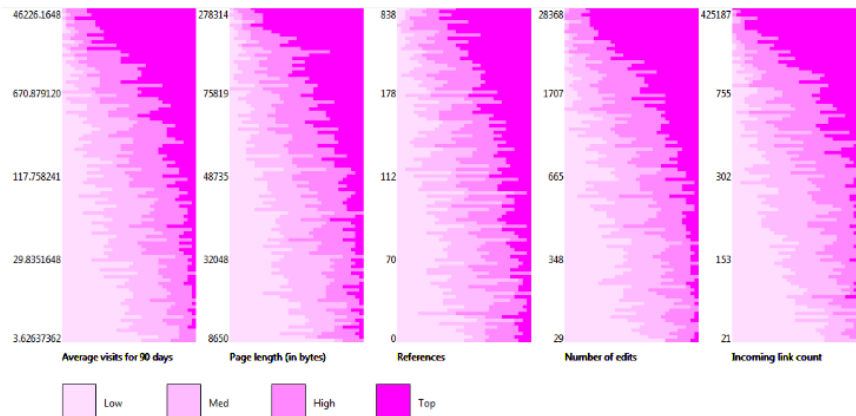
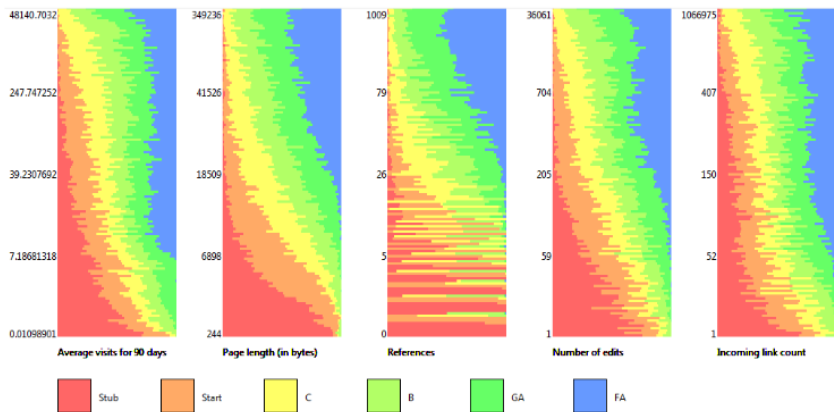


Fact Extraction

- Logical-linguistic model of fact extraction in Russian texts
- Density of simple and complex facts can determine the quality of Wikipedia articles



Quality and Importance Models



| Observed quality | Predicted quality | | | | | |
|------------------|-------------------|-------|-------|------|-------|-------|
| | ★ FA | ⊕ GA | B | C | Start | Stub |
| ★ FA | 2 859 | 277 | 52 | 11 | 1 | 0 |
| ⊕ GA | 575 | 2 302 | 207 | 92 | 24 | 0 |
| B | 111 | 417 | 1 261 | 853 | 454 | 104 |
| C | 35 | 262 | 856 | 1251 | 699 | 97 |
| Start | 8 | 81 | 246 | 609 | 1 734 | 522 |
| Stub | 1 | 12 | 37 | 97 | 563 | 2 490 |

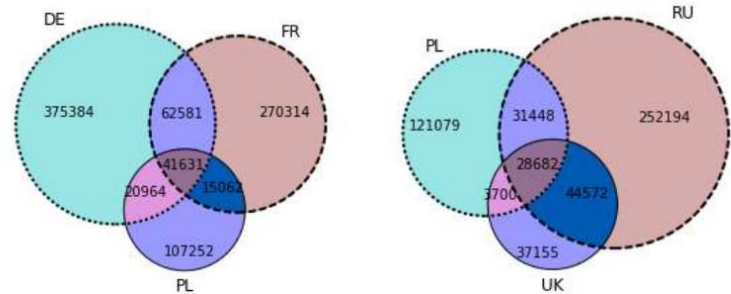
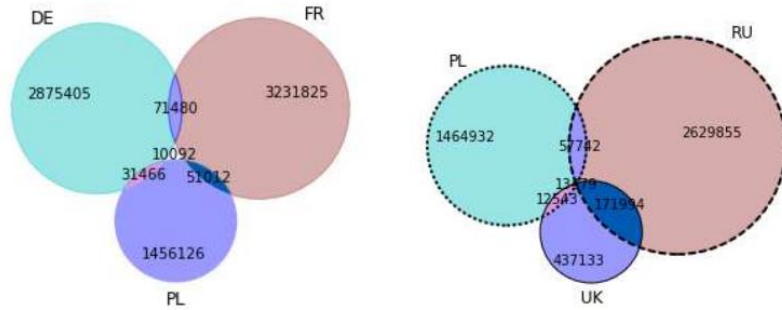
| Observed importance | Predicted importance | | | |
|---------------------|----------------------|-------|-------|------|
| | Top | High | Mid | Low |
| Top | 3 176 | 900 | 461 | 263 |
| High | 1 431 | 1 608 | 948 | 813 |
| Mid | 618 | 1 064 | 1 559 | 1559 |
| Low | 225 | 507 | 978 | 3090 |

Source: Lewoniewski, W., Węcel, K., Abramowicz, W. (2016). *Quality and importance of Wikipedia articles in different languages. In International Conference on Information and Software Technologies (pp. 613-624). Springer, Cham.*

Analysis of References in Wikipedia

Overlaps of unique references :

Overlaps of domains of references



| lang. | BE | DE | EN | FR | PL | RU | UK |
|-------|--------|-----------|------------|-----------|-----------|-----------|---------|
| BE | 82,295 | 3,522 | 19,116 | 6,127 | 5,043 | 47,931 | 13,100 |
| DE | - | 2,988,443 | 345,202 | 81,572 | 41,558 | 69,634 | 21,097 |
| EN | - | - | 18,470,130 | 584,037 | 244,120 | 635,546 | 160,408 |
| FR | - | - | - | 3,364,409 | 61,104 | 118,700 | 32,470 |
| PL | - | - | - | - | 1,548,696 | 71,221 | 26,022 |
| RU | - | - | - | - | - | 2,873,070 | 185,473 |
| UK | - | - | - | - | - | - | 635,149 |

| lang. | BE | DE | EN | FR | PL | RU | UK |
|-------|--------|---------|-----------|---------|---------|---------|---------|
| BE | 22,042 | 10,563 | 15,393 | 10,475 | 9,783 | 19,030 | 12,485 |
| DE | - | 500,560 | 219,536 | 104,212 | 62,595 | 90,361 | 41,407 |
| EN | - | - | 1,588,692 | 201,601 | 101,495 | 183,234 | 69,437 |
| FR | - | - | - | 389,588 | 56,693 | 86,071 | 39,426 |
| PL | - | - | - | - | 184,909 | 60,130 | 32,382 |
| RU | - | - | - | - | - | 356,896 | 73,254 |
| UK | - | - | - | - | - | - | 114,109 |