

Use of Linguistic Criteria for Estimating of Wikipedia Articles Quality

Anastasiia Kolesnik and Nina Khairova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

kolesniknastya20@gmail.com, nina_khajrova@yahoo.com

As far as the question of texts and articles quality is urgent today, in process of a research, the concept of quality for Wikipedia articles was analysed. There were marked out linguistic criteria of quality for technical documentation and scientific articles.

Nowadays everyone knows about such informational resource as Wikipedia. Since that day when Wikipedia was just an offshoot of Nupedia (project to produce a free encyclopedia), it has become the most well-known and popular internet encyclopedia with 282 active language editions such as German, French, Russian and Polish and of course the biggest one is English edition, that has more than 5 million articles. It is multilingual, web-based, free content encyclopedia project. It takes the 5th place according to the list of the most popular websites [6].

Wikipedia is written collaboratively by largely anonymous volunteers who write without pay. Anyone, with Internet access, can write and make changes to Wikipedia articles, except in limited cases where editing is restricted to prevent disruption or vandalism. Users can contribute anonymously, under a pseudonym, or, if they choose to, with their real identity. Some users visit Wikipedia to share their knowledge, others to get (acquire) [6].

Every day, hundreds of thousands of visitors from the various parts of the world collectively make tens of thousands of edits and create thousands of new articles to augment the knowledge held by the Wikipedia encyclopedia. All users, old or young, with different backgrounds and people of all cultures can make changes in articles or add their own one.

Wikipedia's greatest strengths, weaknesses, and differences all arise because it is open to anyone, it has a large contributor base, and its articles are written according to editorial guidelines and policies. According to the *Nature* (the first to use peer review that compares Wikipedia and Britannica's coverage of science), Wikipedia's strongest suit is the speed at which it can be updated, a factor not considered by *Nature's* reviewers. Of course it has large amount of uncovered flaws, different kinds of factual errors, omissions or misleading statements.

Quality issues, however, concern the creators of Wikipedia. That's why, in 2006 during the Opening plenary at Wikimania, Jimmy Wales suggested to concentrate on quality of the articles instead of their number [2]. They created assessment system *WP: ASSESS*. It uses a letter scheme which estimates how complete the article is, assigning to the definite article its grade. According to this system, Wikipedia has 9 grades: FA (Featured Article) [4], A, GA (Good Article), B, C, Start, Stub, FL (Featured List), List. Each of these grades has special criteria. Featured articles are

considered to be the best articles in Wikipedia [2]. This kind of article must be well-written, comprehensive, well-researched, neutral and stable.

The article with A grade is well-written, clear, appropriately structured, well referenced and it contains complete description of the topic. A good article (GA) [5] also must be well written, its spelling and grammar are correct, it complies with the manual of style guidelines and it mustn't contain copyright violations or plagiarism. The prose of the Start article is not fully un-encyclopedic but it should satisfy fundamental content policies.

All experts admit that there are some difficulties in determining the quality of the Wikipedia articles [3]. Such not easy task is connected with large number of articles (3.7 million articles). It is obvious that it is not easy task to search and evaluate all of them, especially when their amount keeps growing every day. Wikipedia isn't static resource. Anyone can make changes and it can well affect article quality.

The following linguistic resources, which are well-recommended at estimating of technical documentation quality, are proposed to be used for quality evaluation of Wikipedia articles [1]:

- Writing of digits from 1 through 9 in words.
- Use of numerals for 10 and greater.
- Use of numerals for all measurements, even if the number is less than 10.
- Use of one-word verbs instead of verb phrase
- Use of only international writing of terms.
- Use of only one gap after the punctuation mark.
- There is no coma in MMMM YYYY date format.
- Use punctuation mark without extra gap.
- Use of (*from i through*) instead of (*between i and*).
- Slash cannot be a substitute of "or".
- Use of MMMM DD, YYYY date format.
- No abbreviation of months (only full names).
- Use of italic formatting instead of upper-case.

Given linguistic criteria for estimating Wikipedia articles quality can be easily formalized, that will allow to raise efficiency of automatic estimating of articles quality.

References

1. Microsoft Manual of Style 4th edition / Published by Microsoft Press. – 2012. – 439 p.
2. Giles G. Internet encyclopedias go head to head. *Nature*, 438 (2005), 900-901. Wikipedia: Manual of Style: [Electronic source]. – Access mode: <http://en.wikipedia.org/wiki>
3. Wikipedia: WikiProject Articles for creation/Assessment: [Electronic source]. – Access mode: [http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Articles_for creation/ Assessment](http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Articles_for_creation/Assessment)
4. Wikipedia: featured articles: [Electronic source]. – Access mode: [http://en.wikipedia.org/wiki/Wikipedia:featu redarticles](http://en.wikipedia.org/wiki/Wikipedia:featu_redarticles).
5. Wikipedia: good_articles: [Electronic source]. – Access mode: [http://en.wikipedia.org/wiki/ Wikipedia:good_articles](http://en.wikipedia.org/wiki/Wikipedia:good_articles).
6. Stvilia B., Twidale M.B., Gasser L., Smith L.C. Information quality discussions in Wikipedia // In Proc. ICKM, 2005. –P. 101-113.