Is Wikipedia Too Difficult? Comparative Analysis of Readability of Wikipedia, Simple Wikipedia and Britannica

Adam Jatowt[1,2] and Katsumi Tanaka[1]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan

[2]Japan Science and Technology Agency
4-1-8 Honcho, Kawaguchi-shi, Saitama
332-0012 Tokyo, Japan

{adam, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Readability is one of key factors determining document quality and reader's satisfaction. In this paper we analyze readability of Wikipedia, which is a popular source of information for searchers about unknown topics. Although Wikipedia articles are frequently listed by search engines on top ranks, they are often too difficult for average readers searching information about difficult queries. We examine the average readability of content in Wikipedia and compare it to the one in Simple Wikipedia and Britannica. Next, we investigate readability of selected categories in Wikipedia. Apart from standard readability measures we use some new metrics based on words' popularity and their distributions across different document genres and topics.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

readability, web content analysis, web search

1. INTRODUCTION

With the increasing popularity of the Web people search for information which is not only related to their daily lives but also related to topics they are not familiar with. Upon encountering a new word (e.g., scientific term) users frequently turn to a search engine to learn about its meaning or simply to find more information. Often, the first page in search results is a Wikipedia article containing elaborate description of the term. Users then tend to refer to Wikipedia as a starting point in their learning process as it contains overview of most real world concepts and topics. Wikipedia articles are usually extensive and cover key aspects of described concepts and topics. They have been also reported to be of sufficient accuracy [8]. However, many Wikipedia articles such as the ones on scientific, technical or legal topics are too difficult for average users. Even though anyone can edit Wikipedia, the articles on difficult topics tend to be written by professionals, experts and hobbyists, thus generally by knowledgeable authors who possess sufficient knowledge. As the accuracy is the key objective, any simplifications, generalizations or intuitive explanations may never be provided following the requirement of correct and accurate content, or they may be simply removed by other knowledgeable editors who attempt to correct any inaccurate or ambiguous content. In result, the final product after many revisions often becomes quite unreadable and inaccessible to average readers who lack knowledge, expertise or cognitive capabilities.

Thus users viewing Wikipedia may have trouble understanding the content, especially the one on difficult topics (e.g., mathematics, physics, engineering, economics, and medicine). This is not surprising considering the average education level and literacy in society. According to the National Adult Literacy Survey approximately 21% of the adult population in USA have low literacy skills, defined as reading at the 6 grade level or below, while another 27% have limited literacy ability, defined as lacking general reading and numeric proficiency to function adequately in society [9]. In addition, for many users of Wikipedia English may not be the first language.

To shed light on the scope and character of this problem we conduct comparative analysis of the readability of Wikipedia. As comparison sources we use Simple Wikipedia[1] and Britannica[2]. The former was created in 2003 and aims at providing easy to read content. It thus constitutes a good source of reference. Simple Wikipedia has over 85k articles written in easy English. On the other hand, Britannica contains over 90k articles and has been frequently used for comparison purposes with Wikipedia in terms of content quality and accuracy [8]. Besides comparing the three encyclopedias, we also measure readability levels of selected categories in Wikipedia.

In order to evaluate comprehensibility from diverse aspects we employ statistical analysis of text features and use standard readability metrics. We also propose three additional measures related to content familiarity: one based on word popularity in language corpus, one based on word popularity across different topics and one based on word popularity across different document genres.

Although many works were devoted to the problem of Wikipedia's quality (e.g., [8]) few focused particularly on Wikipedia's readability. For example, Ehmann et al. [5] demonstrated variability in article quality across diverse disciplines and on a relationship between talk page discussion and article editing activity. One of the used criteria was language complexity estimated by Flesch–Kincaid Grade Level and Flesh Reading Ease metrics. Blumenstock et al. [1] used word count as a proxy of article quality. In contrast, our study analyzes the

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

[1] http://simple.wikipedia.org

[2] http://www.britannica.com

readability of Wikipedia in greater detail as previous works as well as it provides comparison with other encyclopedias.

2. READABILITY METRICS

In this section we describe readability indices that we are going to use. All the measures were normalized using min-max normalization to fit into [0,1] interval. Also, as some of the measures originally indicate a schooling grade level necessary to comprehend texts while others output the raw ease score, we inverted the former so that all the used metrics estimate how easy is to read target documents. High values indicate easy text while low ones characterize difficult texts. We note that even though more complex readability approaches have been recently proposed (e.g., [3]), for large scale data analysis, easy measures perform best. Thus we limit the analysis to relatively simple but effective metrics, which in most cases have been frequently used for readability measurements until now.

2.1 Syntactical Readability Measures

2.1.1 Flesch Reading Ease

Flesch Reading Ease (*FRE*) [7] is a measure of how easy it is to read a text. It is defined as: 206.876 -1.015*ASL* – 84.6*ASW* where *ASL* is the average number of syllables per word and *ASW* is the average number of words per sentence.

2.1.2 Automated Readability Index

Automated Readability Index (*ARI*) produces an approximate representation of the U.S. grade level needed to comprehend any text. The formula is: 4.71(*C*/*W*) + 0.5(*W*/*S*) – 21.43, where *C* is the number of characters, *W* is the number of words and *S* denotes the number of sentences.

2.1.3 Coleman Liau Index

Coleman Liau Index (*CLI*) [2] like ARI approximates the U.S. grade level thought to be necessary for comprehending any text. The index is calculated as: 0.0588*L* – 0.296*S* – 15.8 where *L* stands for the average number of letters per 100 words and *S* denotes the average number of sentences per 100 words.

2.2 Familiarity-based Approaches

Intuitively, word popularity is related to document's readability. Prior studies [6] suggest that document vocabulary is a good predictor of document readability. Simple texts are more likely to use basic words as opposed to complicated, domain-specific texts. Moreover, the average college level has been found to be correlated to the amount of words ones knows. We then use several readability metrics based on word popularity.

2.2.1 New Dale-Chall Formula

New Dale-Chall (*NDC*) measure is an extension to an old formula proposed by Dale and Chall in 1949 [4]. This measure is a combination of syntactic and familiarity-based formulas. It considers both the average sentence length and the number of difficult words, where difficult words are the ones which are not on the prefixed list of 3,000 common words.

2.2.2 Popularity-based Familiarity

The problem with the NDC formula is that it is based on the fixed list of only 3,000 common words, while many words that are familiar to average users are treated as being equally difficult. We use the Corpus of Contemporary American English[3] (COCA) as an indication of word frequency. It consists of about 500k words compiled over 160k documents that were equally divided among spoken language, fiction, popular magazines, newspapers, and academic texts. The proposed measure is calculated as follows:

$$\frac{1}{|D|}\sum_{t\in D}\ln\big(cf(t)\big) \quad (1)$$

where *cf(t)* is the frequency of term *t* in COCA and D is a target document. We call this measure the *Popularity-based Familiarity* (*PF*).

2.2.3 Topic-based Familiarity

In addition to the frequency-based approach, we provide a topic-based measure. We collected articles related to distinct topical categories that are common to everyday life. These include: nation, world, business, entertainment, sports, health, science and technology. We used topic feeds from Google News[4] as a source of category-related content and gathered in total over 107k news articles. We next calculated the probability distribution of each word across the corresponding categories. The entropy over this distribution indicates whether a given word is equally common to all the categories or is rather specific to few ones. Then, the average entropy score of words in a target document is used as a measure of document's readability. We call this measure the *Topic-based Familiarity* (*TF*).

2.2.4 Genre-based Familiarity

We also provide document genre-based measure which uses the American National Corpus[5] (ANC) as a source of diverse document genres. ANC is a corpus covering American English and contains 8,832 documents (total of 11 million words) belonging to genres such as government documents, technical documents, travel documents, letters, non-fiction and journals. In a similar way to the Topic-based Familiarity we calculated word distribution across different genres and then measured its entropy. The average entropy score of words in a target document is then used as a measure of its readability. We call this measure the *Genre-based Familiarity* (*GF*).

3. EXPERIMENTS

3.1 Simple Wikipedia vs. Wikipedia

In order to compare Simple Wikipedia and English Wikipedia we collected all the articles in the former and found their corresponding articles in Wikipedia. The datasets were created on June 2012. Next, from both the datasets we removed articles having less than 50 words, disambiguation pages and redirection pages. We also removed outlier articles based on their readability levels by applying Modified Thompson Tau technique. We then kept only those article pairs for which the two versions, the one in Simple Wikipedia and in Wikipedia, remained after the filtering process. In total, 25,970 article pairs were obtained (51,940 articles). For each article we removed markup tags and extracted its core content. The resulting dataset was then used for describing the characteristics of Simple Wikipedia and Wikipedia. Note that we did not perform random selection of articles from Wikipedia as we simply collected the ones which have their corresponding articles in Simple Wikipedia. Nevertheless, we can still treat them as a representative part of Wikipedia since there is no concrete policy for editors to choose article topics in Simple Wikipedia.

First, we wished to know how much different both encyclopedias are in terms of the information size provided in their articles. A crude approach for comparing the amount of information in two documents is simply to measure their sizes, as, intuitively, long

[3] http://corpus.byu.edu/coca

[4] http://news.google.com

[5] http://americannationalcorpus.org/OANC/index.html

documents tend to have more information. In Table 1 we show Wikipedia's and Simple Wikipedia's comparison in terms of basic text features. We observe that the average article length of Wikipedia as expressed by either word or sentence length is several times higher than the one in Simple Wikipedia. The average lengths of words and sentences are also larger in Wikipedia than in Simple Wikipedia. Note that most readability indexes such as Flesh-Reading Ease, Coleman-Liau or Dale-Chall include both the word and sentence lengths as key factors.

To find the vocabulary variation we calculated *Type to Token Ratio* (*TTR*) which is the rate of unique words to the total number of words. It is higher in Simple Wikipedia (0.51) than in Wikipedia (0.32). We attribute this fact to the much larger average size of Wikipedia's articles. It has been actually found that TTR to document size correlation is nearly -0.8 on large datasets such as LIWC2001[6]. We use thus the *Standardized TTR* (*STTR*) measure which calculates TTR on a fixed number of 100 top words in each document, provided the document length is over 100 words. STTR value is 0.61 for Simple Wikipedia and 0.63 for Wikipedia thus both encyclopedias have nearly same STTR.

Table 1 Comparison of average text features of Simple Wikipedia (SW) and Wikipedia (W).

	#words	#unique words	#sent	sent. length	word length
SW	428	160	22.19	20.32	4.15
W	3550	893	138	25.54	4.34

Next we compared readability of both the encyclopedias. The results are shown in Figure 1. We can see that Simple Wikipedia is easier according to all the metrics. The percentage difference of readability is relatively stable for most of the used measures (average difference of 26%) apart for the case of FRE (66% difference).

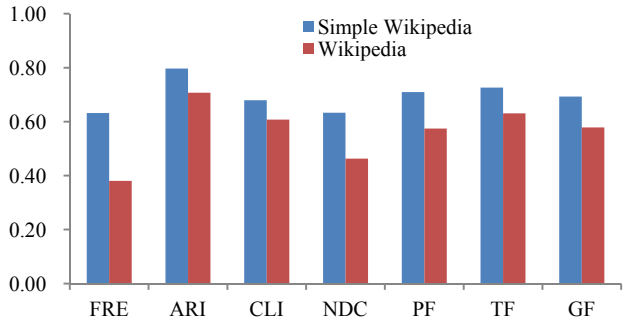

Figure 1 Comparison of readability measures of Simple Wikipedia and Wikipedia.

Not surprisingly, Simple Wikipedia is more readable than Wikipedia, however, this gain is offset by low information coverage of the former as approximated by the average document size and low topical coverage (approximated by the article count) when compared to the ones of Wikipedia.

In the next step, we performed POS tag analysis using POS tagger available in CLIPS Pattern Library[7]. Our objective was to compare the grammatical complexity of the encyclopedias. POS tags are used as the basis for lexical density measure [10] which estimates the rate of content words (e.g., nouns, adjectives, verbs, etc.) to the total number of words in a document. It is considered that documents with high lexical density tend to contain much information (e.g., academic papers) and may be poorly understandable by readers. In Figure 2 we show probability distributions of POS tags following the Penn Treebank II tag set[8] for Simple Wikipedia, Wikipedia and Britannica. We can observe that Wikipedia contains on average more proper nouns than Britannica. This may be due to more relaxed editorial rules for the former, while the editors of the Britannica may not be allowed to mention too many names of objects, persons or companies unless they are necessary for explaining target topics. We also see that Simple Wikipedia articles contain more nouns (NN) and proper nouns (NNP) than the other encyclopedias, while having less determinants and conjunctions, adjectives and prepositions. This may be due to the need to convey much information in relatively short content.

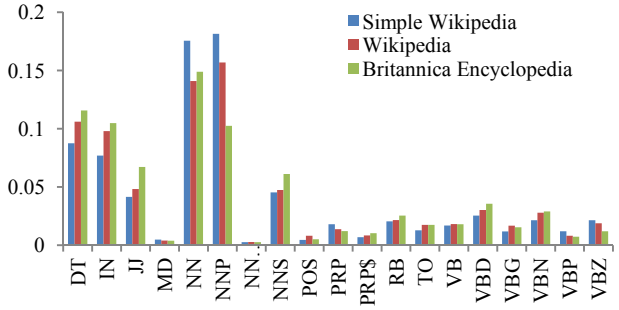

Figure 2 Distribution of POS tags in Simple Wikipedia, Wikipedia and Britannica.

3.2 Readability of Wikipedia Categories

In this section we analyze the readability of Wikipedia articles according to selected categories to investigate how readability varies across different domains in Wikipedia. We have chosen 8 categories using the DBpedia ontology[9] and randomly picked up to 500 articles from within each category. The categories are: Biology, Chemistry, Computing, Economics, History, Literature, Mathematics, and Philosophy. Figure 3 shows the average readability scores of pages in all the chosen categories. Looking at the results we can observe that different categories have sometimes different readability levels and the familiarity-based metrics tend to produce more varying results than the syntactical readability measures. We can also notice that the articles in the Computing category are the most readable using both the syntactical and familiarity measures. This could be partly explained by the ubiquitous character of IT technology. On the other hand, the articles in Biology and Chemistry categories seem to be the most difficult. This could be attributed to high number of rare technical words used in these domains, which often tend to be relatively long (e.g., the names of species and chemical compounds). Another observation is that the articles in History category have relatively high syntactical readability while quite low the familiarity-based one. We think that while the writing style of those articles is not too difficult, many historical entities mentioned in their content (e.g., names of historical places or persons) may not be frequently used in everyday life.

[6] http://www.liwc.net/liwcdescription.php

[7] http://www.clips.ua.ac.be/pages/pattern

[8] http://www.clips.ua.ac.be/pages/mbsp-tags

[9] http://dbpedia.org

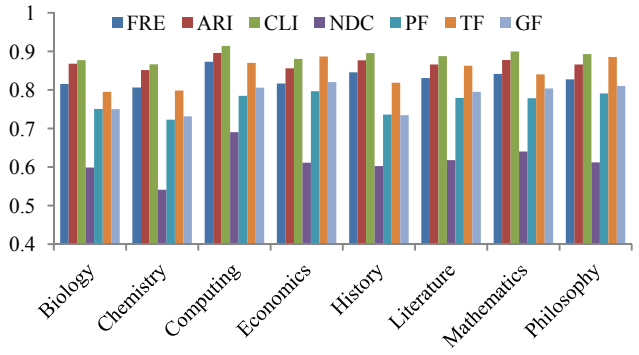

Figure 3 Readability of selected Wikipedia categories.

3.3 Readability of Britannica

Britannica's articles belong to three categories: the Encyclopedia, Intermediate and Introductory category according to their reading ease and the knowledge levels of expected readers. We collected 93,659 articles in the Encyclopedia, 16,250 in Intermediate and 2,777 articles in Introductory category. We then removed markup tags and extracted core content of each article.

First, we show in Table 2 the same basic statistics of the articles as in the case of Wikipedia and Simple Wikipedia (see Section 3.1.). We can see that the Introductory category's articles have shorter sentences, while the average article length is rather similar across the three categories. The average word length is the shortest for the Introductory category. The TTR and STTR values are as follows: Introductory (0.38, 0.62), Intermediate (0.37, 0.66) and Encyclopedia (0.34, 0.68). STTR values tend to mildly increase for more difficult categories. We see that Britannica's Encyclopedia category has a little higher STTR than the one for Wikipedia (0.63).

Table 2 Average text features of Britannica categories.

	#words	#unique words	#sent	sent. length	word length
Introd.	661	252	42.72	15.37	4.13
Interm.	609	222	31.06	29.49	4.23
Encycl.	640	214	25.21	23.98	4.32

In the next step we calculated readability values of Britannica categories as shown in Figure 4. Comparing Figure 4 with Figure 1 we see that, on average, the articles in the Britannica seem to be easier than the ones in Wikipedia (21% average difference over all the metrics when using the Encyclopedia category of Britannica). Also, we see that the Introductory category's articles are the easiest and are on average more readable than the articles of Simple Wikipedia (11% average difference across all the metrics).

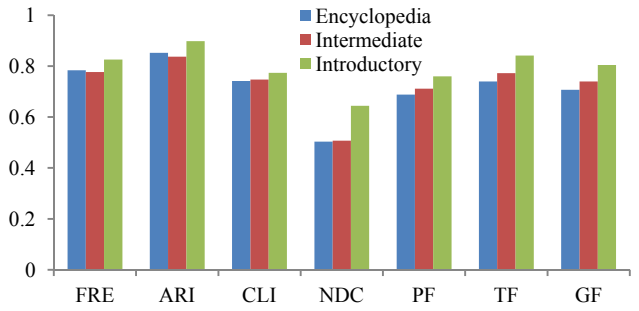

Figure 4 Readability of Britannica categories.

Lastly, we show POS tag analysis within the article categories in Britannica in Figure 5. The comparison of Britannica's and Wikipedia's POS tag distributions has been already shown in Figure 2. From Figure 5, we can see that there are certain differences between different categories, although they seem to be less pronounced than the ones for the case of Simple Wikipedia vs. Wikipedia comparison.

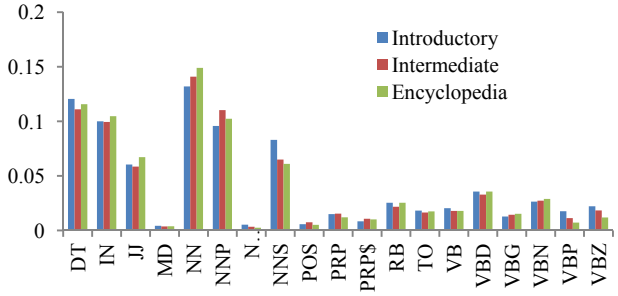

Figure 5 Distribution of POS tags in Britannica.

4. CONCLUSIONS AND FUTURE WORK

Since the accuracy and coverage are the main goals in the process of editing Wikipedia articles, articles' comprehensibility may be low. We study this problem by comparing reading ease of Wikipedia content to the one in Simple Wikipedia and Britannica. We use common readability metrics and also propose new readability measures based on word familiarity. In conclusion, Wikipedia seems to lag behind the other encyclopedias in terms of readability and comprehensibility of its content. Modifying editorial guidelines or automatically flagging poorly comprehensible content for revision may be thus needed to improve this situation. In future, we plan to conduct user studies to investigate comprehensibility issues of Wikipedia in detail.

5. ACKNOWLEDGMENTS

This research was supported in part by MEXT Grant-in-Aid for Scientific Research (#24240013).

6. REFERENCES

[1] J. Blumenstock. Automatically Assessing the Quality of Wikipedia Articles, Recent works, School of Information, UC Berkeley, 2008

[2] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring, *Journal of Applied Psychology*, Vol. 60, pp. 283–284, 1975.

[3] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL 2004*

[4] E. Dale and J.S. Chall. The concept of readability, *Elementary English*, 26(23), 1949

[5] K. Ehmann, A. Large, and J. Beheshti. Collaboration in Context: Comparing Article Evolution among Subject Disciplines in Wikipedia, First Monday, Volume 13 Number 10, 2008.

[6] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively Motivated Features for Readability Assessment. In *ECCL, 2009*

[7] R. Flesch. —A new readability yardstick, *Journal of Applied Psychology*, 1948, 32(3), pp. 221-233.

[8] J. Giles. Internet encyclopaedias go head to head, Nature 438, 900-901, 2005

[9] D.I. Shalowitz and S. Wolf. Shared decision-making and the lower literate patient. Journal of law, medicine & ethics, 32, 759-64, 2004

[10] J. Ure. *Lexical density and register differentiation*, London: Cambridge University Press, 443–452, 1971