DEPTH$^+$: An Enhanced Depth Metric for Wikipedia Corpora Quality

Saied Alshahrani **Norah Alshahrani** **Jeanna Matthews**
Department of Computer Science
Clarkson University, Potsdam, NY, USA
{alshahsf,alshahnf,jnm}@clarkson.edu

Abstract

Wikipedia articles are a common source of training data for Natural Language Processing (NLP) research, especially as a source for corpora in languages other than English. However, research has shown that not all Wikipedia editions are produced organically by native speakers, and there are substantial levels of automation and translation activities in the Wikipedia project that could negatively impact the degree to which they truly represent the language and the culture of native speakers. To encourage transparency in the Wikipedia project, Wikimedia Foundation introduced the depth metric as an indication of the degree of collaboration or how frequently users edit a Wikipedia edition's articles. While a promising start, this depth metric suffers from a few serious problems, like a lack of adequate handling of inflation of edits metric and a lack of full utilization of users-related metrics. In this paper, we propose the DEPTH$^+$ metric, provide its mathematical definitions, and describe how it reflects a better representation of the depth of human collaborativeness. We also quantify the bot activities in Wikipedia and offer a bot-free depth metric after the removal of the bot-created articles and the bot-made edits on the Wikipedia articles.

1 Introduction

The Wikipedia project is a free online encyclopedia that aims to enable and involve people all over the globe in creating and disseminating knowledge. Wikipedia articles, i.e., content pages of Wikipedia, are also a common source of training data for Natural Language Processing (NLP) research, especially as a source for corpora in languages other than English. In particular, Wikipedia articles are used to train many Large Language Models (LLMs), such as ELMo (Embeddings from Language Models), which has been trained on the English Wikipedia and news crawl data (Peters et al., 2018); BERT (Bidirectional Encoder Representations from Transformers) has been trained on books with a crawl of the English Wikipedia articles (Devlin et al., 2018; Petroni et al., 2019); GPT-3 (Generative Pre-trained Transformer) has also been trained on five large datasets including the English Wikipedia (Brown et al., 2020); LaMDA (Language Model for Dialogue Applications) and PaLM (Pathways Language Model) were trained on a huge mixed dataset that includes Wikipedia articles, news articles, source code, and social media conversations (Thoppilan et al., 2022; Chowdhery et al., 2022); and LLaMA (Large Language Model Meta AI) was also pre-trained on the multilingual articles of Wikipedia from June to August 2022, covering 20 languages with a percentage of 4.5% of its overall training dataset size (Touvron et al., 2023).

Wikipedia corpora (editions) exist for more than 300 of the over 7,000 languages spoken worldwide.

Global Rank	Wikipedia Language	Depth Metric Rank	Depth Metric Value	Depth$^+$ Metric Rank	Depth$^+$ Metric Value	Depth$^+$ vs. Depth
1st	English	3rd ▼	1178	1st =	377	▲
2nd	Cebuano	300th ▼	2	49th ▼	0.64	▲
3rd	German	72nd ▼	93	2nd ▲	41	▲
4th	Swedish	216th ▼	17	12st ▼	6	▲
5th	French	22nd ▼	257	3th ▲	37	▲
6th	Dutch	210th ▼	18	18th ▼	3	▲
7th	Russian	46th ▼	153	6th ▲	12	▲
8th	Spanish	30th ▼	201	10th ▼	7	▲
9th	Italian	35th ▼	183	4th ▲	20	▲
10th	Egyptian	316th ▼	0.30	315th ▼	0.003	▲

Figure 1: Changes in the global rank for the top ten Wikipedia editions regarding the number of articles[1]. The arrows in the 3rd and 4th columns indicate the changes in the rankings of editions when depth and DEPTH$^+$ are compared to the global rank, and the arrows in the 5th column indicate the changes in rankings when DEPTH$^+$ and depth are compared head-to-head.

[1]The global rank of Wikipedia editions is calculated using the total number of articles (content pages) (Wikipedia, 2023a). See Appendix A for the full list.

Language (code)	Articles	Non-Articles	Total Pages	Edits	Users	Active Users	Admins	Depth (filtered)*	Depth (unfiltered)**	Depth+
Cree (cr)	161	2,027	2,188	38,220	17,790	16	2	– –	2,768.85	0.37
Greenlandic (kl)	242	2,023	2,265	74,746	12,796	12	3	– –	2,306.11	0.70
English (en)	6,642,196	51,299,727	57,941,923	1,144,555,884	45,353,848	127,885	908	1,178.29	1,178.29	376.77
Dzongkha (dz)	237	2,384	2,621	30,174	9,788	13	1	– –	1,164.88	0.10
Ripuarian (ksh)	2,940	7,644	10,584	1,607,356	22,054	17	3	1,026.62	1,026.62	0.87
Tigrinya (ti)	256	2,514	2,770	24,152	8,957	10	2	– –	840.86	0.15
Serbo-Croatian (sh)	457,985	4,189,557	4,647,542	41,404,769	184,125	201	8	745.52	745.52	0.99
Vietnamese (vi)	1,282,386	18,132,725	19,415,111	69,812,540	905,163	2,010	19	718.92	718.92	3.87
Bihari (Bhojpuri) (bh)	8,311	63,893	72,204	744,087	31,956	59	2	609.06	609.06	0.35
Inuktitut (iu)	449	2,563	3,012	46,139	18,216	32	2	– –	499.13	0.19

Table 1: Metrics for ten Wikipedia editions, including the number of articles, non-articles, total pages, edits, users, active users, and administrators (admins). These are the top ten languages ordered by the unfiltered depth metric** values. As we will discuss in more detail in this paper, the Wikipedia project uses a filtered depth metric*, replacing the depth values with "– –" for languages when the number of articles $< 100{,}000$, and the depth metric value > 300.

However, these corpora vary substantially in size and quality, and the Wikipedia project provides a rich set of metadata and metrics to help users compare the different corpora. Table 1 includes examples of some of these metrics across ten languages, including the number of articles, the number of non-articles (e.g., user pages, redirects, images, project pages, templates, and support pages), the total number of pages (articles and non-articles), the total number of edits, the number of users, the number of active users, and the number of admins. The difference between users and active users is that users refer to the number of user accounts regardless of current activity, whereas active users refer to registered users who have made at least one edit in the last thirty days (Wikipedia, 2023a).

In this paper, we will use the 320 open Wikipedia corpora available today, as listed in the appendices. We will not include the 13 closed Wikipedia editions (Afar, Northern Luri, Marshallese, Ndonga, Choctaw, Kwanyama, Herero, Hiri Motu, Kanuri, Muscogee, Sichuan Yi, Akan, and Nauruan). Closed editions are read-only, meaning registered users can no longer edit any content pages (Wikipedia, 2023a; Wikimedia Commons, 2023; Wikimedia Meta-Wiki, 2023). Since articles in closed editions can no longer be edited, the active users metric drops to zero because it only counts users active in the last 30 days[2]. The last three columns of Table 1 contain filtered depth metric (as the Wikipedia project does it), unfiltered depth metric (as we used to sort the table), and the new Depth$^+$ metric we are proposing in this paper. The current general formula of the depth metric used by Wikipedia is defined as the following:

$$Depth = \frac{Edits \cdot NonArticles}{Articles^2} \cdot \left(1 - \frac{Articles}{Total}\right) \quad (1)$$

[2]We would love to see the Wikimedia Foundation, in its Wikipedia project, maintain and report a count of the number of users who have ever made an edit in corpora (edition) rather than only reporting on the last 30 days. Such a metric would continue to be relevant even for closed editions.

The Wikimedia Foundation introduced the depth metric as an indicator of Wikipedia's collaborative quality to show how frequently a Wikipedia edition's articles are edited or updated by the users and is intended to indicate the depth of collaboration among contributors to corpora. The first variant of depth metric was added to the Wikipedia project in 2006, using only the first factor, the total number of edits divided by the number of articles. After that, the Wikipedia project added an additional factor of non-articles divided by articles. In 2007, the depth metric was again updated to add the third factor, the stub ratio, or one minus the articles divided by the total pages (Wikimedia Foundation, 2023c).

In this paper, we aim to explore the limitations of the depth metric and propose a new enhanced depth metric, Depth$^+$, to address these limitations. Figure 1 previews a comparison of Wikipedia's unfiltered depth metric and our Depth$^+$ metric for the top ten Wikipedia editions based on global rank (i.e., the total number of articles).

We observe that not all Wikipedia editions are produced organically by native speakers, and a substantial level of automation and translation is often used, which can negatively affect the integrity and trustworthiness of these articles. For example, Alshahrani et al. (2022) studied the Arabic Wikipedia editions (Modern Standard Arabic, Egyptian Arabic, and Moroccan Arabic) and found that more than one million articles have been shallowly translated from English using either direct translation or template-based translation (by one registered user) in the Egyptian Arabic Wikipedia edition. Unsurprisingly, some of these top ten Wikipedia editions, in Table 1, are mostly bot-generated, auto-translated, or even small enough not to be considered a common Wikipedia edition (Wikipedia, 2023a; Wikimedia Foundation, 2023a). We found that in the Vietnamese and Serbo-Croatian Wikipedia editions more than 58% and 55% of their articles are bot-created, respec-

tively (Wikipedia, 2023a; Wikimedia Foundation, 2019, 2023d). While automation and translation activities are not always problematic, we argue that metrics like the depth that do not distinguish between organic content generated by native speakers and bot-generated content can be a misleading indicator of the collaboration and richness in a dataset.

Section 2 examines the current depth metric used in Wikipedia, rewrites its mathematical representations, and underscores its limitations. In Section 3, the paper quantifies the bot activities within the Wikipedia project. Section 4 introduces a new metric called DEPTH$^+$, presents its mathematical definitions, and highlights its features. We shed light on the limitations of our work in Section 5. Lastly, Sections 6 and 7 briefly discuss related work, provide a concise conclusion, and offer a few future research ideas.

2 Depth Metric

The Wikipedia depth metric is currently reported in two places: *List of Wikipedias* (Wikipedia, 2023a) and *Wikipedia Article Depth* (Wikimedia Foundation, 2023c). Notably, the Wikipedia project filters the calculations of this depth metric and reports depth values only for the Wikipedia editions with more than 100,000 articles. If a Wikipedia edition has a depth value $>$ 300 and the total number of articles $<$ 100,000, then the depth metric value is arbitrarily replaced by "– –". This has the side effect of placing the English Wikipedia edition at the top of Wikipedia's ranking by depth metric. To better understand how the depth metric behaves, we manually calculate and report unfiltered depth metric values of all Wikipedia editions.

Returning to Table 1, the set of languages displayed shows the top ten Wikipedia editions ordered by the depth metric without filtering. We can see that most of the listed Wikipedia editions are small corpora. It is notable that English, the largest and oldest of the editions, is widely believed to have the most collaborative editing, but it only comes in third. Notably, only half of these ten editions (English, Ripuarian, Serbo-Croatian, Vietnamese, and Bihari) would remain after Wikipedia's filtering. The other half would have been given high depth values without filtering using ad-hoc limits, suggesting that the current depth metric may not truly reflect the collaborative quality of corpora. To expand on Table 1, we plotted the highest 50 Wikipedia editions ordered by the depth metric values in Figure 2. Once again, most Wikipedia editions in the highest ranks are counterintuitively small or uncommon languages, while large corpora, such as French (`fr`), Spanish (`es`), and Italian (`it`), all widely believed to have substantial collaborative editing, appear late in the ranking. Overall, this observation motivated our quest for an improved depth metric that would not require ad-hoc filtering.

In the following subsections, we discuss the formulas of the depth metric, rewrite its mathematical representations, and explain some of its limitations.

2.1 Formulas of Depth Metric

The Wikimedia Foundation, in its Wikipedia project, introduces two mathematical formulas for the depth metric that are written in high-level quantitative terms (Wikimedia Foundation, 2023c). In this work, we rewrite these mathematical definitions of the depth metric in detailed formal mathematical representations.

We have already seen one formula for the depth metric in Equation 1. That version emphasizes the three factors added by the Wikipedia project over time. After some simple algebraic transformations, there is an alternate version, Equation 2. It may not be immediately obvious that Equation 2 is equivalent to Equation 1, but for reference, we have provided the full derivation of Equation 2 in Supplementary Section 8.

$$Depth = \frac{Edits}{Total} \cdot \left(\frac{NonArticles}{Articles}\right)^2 \tag{2}$$

Let $\mathcal{W}_i$ represent all Wikipedia editions where i = {1, 2, 3, . . . , 320} (As noted earlier, we are not including the 13 closed editions). Let the total number of edits of $\mathcal{W}_i$ be $\mathcal{E}_{W_i}$ where e = {1, 2, 3, . . . , n}, let the total number of articles of $\mathcal{W}_i$ be $\mathcal{A}_{W_i}$ where a = {1, 2, 3, . . . , n}, let the total number of non-articles of $\mathcal{W}_i$ be $\mathcal{R}_{W_i}$ where r = {1, 2, 3, . . . , n}, and lastly, let the total number of pages of $\mathcal{W}_i$ be $\mathcal{T}_{W_i}$ where $\mathcal{T}_{W_i} = \mathcal{A}_{W_i} + \mathcal{R}_{W_i}$.

Therefore, our rewrite, using the mathematical representations, of the general mathematical definition of the depth metric of $\mathcal{W}_i$ is described as follows:

$$\mathcal{D}_{W_i} = \frac{\mathcal{E}_{W_i} \cdot \mathcal{R}_{W_i}}{\mathcal{A}^2_{W_i}} \cdot \left(1 - \frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right) \tag{3}$$

[3] We changed a few Wikipedia language codes for the sake of data visualization in some figures and tables, such as:
▷ Tarantino: (`roa-tara`) → (`tar`).
▷ Aromanian: (`roa-rup`) → (`roa`).
▷ Southern Min: (`zh-min-nan`) → (`zhm`).
▷ Classical Chinese: (`zh-classical`) → (`zhc`).

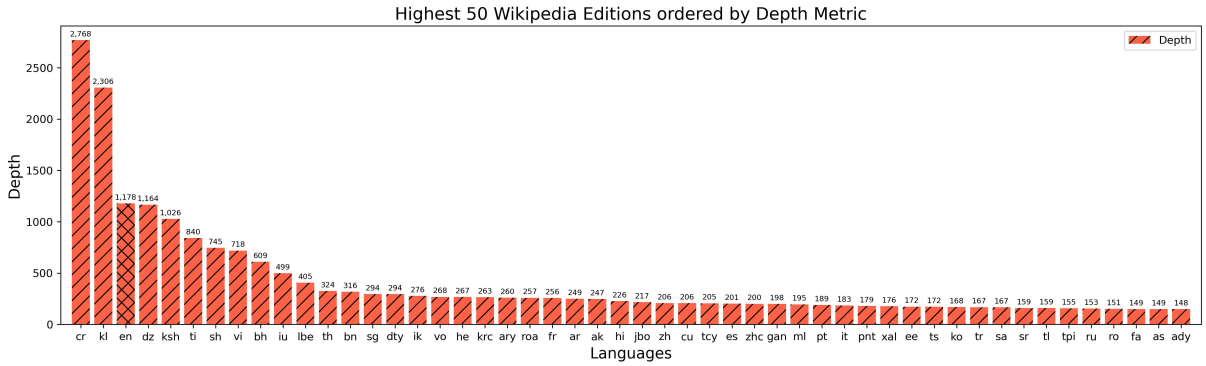

Figure 2: The highest 50 Wikipedia editions ordered by the unfiltered depth metric values[3]. We highlighted English Wikipedia since it is the largest Wikipedia edition. We can see that most languages in the highest ranks are either small or uncommon. See Appendix B for the full list.

The rewrite of the simplified mathematical definition of the depth metric of $\mathcal{W}_i$ is also described using the mathematical representations as follows:

$$\mathcal{D}_{W_i} = \frac{\mathcal{E}_{W_i}}{\mathcal{T}_{W_i}} \cdot \left(\frac{\mathcal{R}_{W_i}}{\mathcal{A}_{W_i}}\right)^2 \tag{4}$$

2.2 Problems of Depth Metric

Having presented the detailed formulas for the depth metric, in this section, we now discuss its key limitations.

2.2.1 Depth Metric is Bot-influenced

The current depth metric is misleading because it measures the total activity on the Wikipedia project, which includes bot and automation activities, instead of solely measuring the human activities, interactions, and collaborations on the project. While not all automated activities are problematic, they provide a misleading sense of the level of collaboration which is one of the stated functions of the depth metric. As an example, the bot-made edits undoubtedly maximize the measurements of the edits metric, causing incorrect calculations of the depth metric. For instance, we found that in the Serbo-Croatian and Inuktitut Wikipedia editions more than 41% and 39% of the total edits on their articles are bot-made, respectively (Wikipedia, 2023a; Wikimedia Foundation, 2019, 2023d).

Furthermore, the current depth metric considers the non-articles in Equations 1 and 2, mostly user pages, redirects, project pages, templates, and discussion pages that are not directly correlated to human activities on Wikipedia articles. Although the users or admins could discuss the contents of articles on their pages (forums), these discussions are not included in the content pages and are not counted toward human activities on those pages.

2.2.2 Depth Metric is Easy-inflatable

The depth metric uses the edits metric as one of the fundamental metrics on which the depth measurements rely. Yet, editing wars in the Wikipedia project inflate this metric of edits, causing inaccurate measurements of the depth metric, even though editing wars are a normal part of Wikipedia's life that is sometimes hard to control (Wikimedia Foundation, 2023b). As an example of the editing wars, in late July 2022, the Wikipedia project locked the English Wikipedia page about the "*recession*" and set restrictions on who could edit this page. The freeze was set after a lot of editors made a series of revisions to the definition of "*recession*" (National Public Radio (NPR), 2022).

2.2.3 Depth Metric Misses User Activity

The depth metric only utilizes a few already calculated metrics by the Wikipedia project, such as articles, non-articles, total pages, and edits, but it does not take advantage of any other metrics related to users of any type, like users, admins, and active users. These user-related metrics already exist and have been calculated by the Wikipedia project for almost all editions (Wikipedia, 2023a). We believe utilizing more metrics could give us insights into the collaborative quality of the Wikipedia editions.

3 Quantification of Bot Activities

The Wikimedia Foundation, in its Wikipedia project, permits users or editors to use bots (software programs) to automate repetitive and everyday tasks in many Wikipedia editions (Wikipedia, 2023d, 2022). The only advantage of Wikipedia bots is to make edits rapidly, yet they can disrupt the Wikipedia project if they are incorrectly designed or operated without approval. For these

reasons, Wikipedia bot policy has been developed and enforced (Wikipedia, 2023c). However, these Wikipedia bots in the past years noticeably are not used only to commit edits but also to create articles on the Wikipedia project, which often produces unrepresentative, inorganic content that does not echo the complex structure of the human languages, does not express the views of the native speakers of those languages, and does not represent the cultural richness and historical heritage of those languages and their people (Alshahrani et al., 2022). As an example of Wikipedia bots, the "Lsjbot" bot is responsible for creating more than 6 million articles (99.61%) in the Cebuano Wikipedia edition, one million articles (90%) in the Waray Wikipedia edition, and one million articles (68%) in the Swedish Wikipedia edition (Popular Science, 2014; Wikimedia Foundation, 2019; Wikipedia, 2023b).

We discuss the quantification and clear labeling of bot-generated Wikipedia articles and bot-made edits on these articles in different Wikipedia editions. If bot-generated content was clearly labeled, it could be included where helpful or ignored when it is not. For instance, if an NLP task involves measuring the opinions or biases of native speakers, including content that has been translated from another language is likely to reflect the opinions or biases of the authors of the original text from which it was translated.

3.1 Bot-generated Articles

To quantify the bot-generated articles in all Wikipedia editions, we used the online Wikimedia Statistics[4] service (https://stats.wikimedia.org) to collect the total number of bot-created articles. Specifically, we collected the statistics of the new content pages (articles) that are created by both group-bots (logged-in registered users who are part of a bot group) and name-bots (logged-in registered users whose name contains 'bot') (Wikimedia Foundation, 2023d). Next, we summed these totals of the bot-generated articles for each Wikipedia edition and subtracted them from the already calculated metrics: articles and total pages by the Wikipedia project to ultimately have a bot-free depth metric.

Table 2 shows the top ten Wikipedia editions that have the most bot-created articles in the Wikipedia project, ordered by the percentage of how much bot automation each Wikipedia edition has. We can see that the Cebuano Wikipedia edition—the second Wikipedia edition in the globe rank in terms of the total number of articles has 99.61% of its total number of articles are bot-generated.

[4]We took a data snapshot of all Wikipedia editions' statistics on the 31st of March, 2023, using the online Wikimedia Statistics service (Wikimedia Foundation, 2023d).

LANGUAGE (CODE)	ARTICLES	BOT-ARTICLES	PERCENTAGE
Cebuano (ceb)	**6,123,587**	**6,099,406**	**99.61%**
Pali (pi)	2,548	2,532	99.37%
Southern Min (zh-min-nan)	432,436	401,203	92.78%
Bishnupriya Manipuri (bpy)	25,087	22,935	91.42%
Waray (war)	1,266,100	1,142,993	90.28%
Malagasy (mg)	95,465	85,574	89.64%
Newar (new)	72,348	63,459	87.71%
Tatar (tt)	499,963	431,558	86.32%
Chechen (ce)	599,686	504,686	84.16%
Tarantino (roa-tara)	9,317	7,521	80.72%

Table 2: The top ten Wikipedia editions that have the most bot-created articles, ordered by the percentage of how much bot automation each Wikipedia edition has. We highlighted the Cebuano Wikipedia edition since it comes second in the global rank and has the highest number of bot-generated articles (content pages). See Appendix C for the full list.

3.2 Bot-made Edits on Articles

With the same aim as above, we want to quantify and eliminate the bot-made edits on Wikipedia articles in all Wikipedia editions. We used the online Wikimedia Statistics service to collect the total number of bot-made edits on articles (content pages). Particularly, we collected the statistics of the made edits on the articles that were done by both group-bots and name-bots (Wikimedia Foundation, 2023d). After that, we summed these totals of the bot-made edits for each Wikipedia edition and subtracted them from the existing edits metric by the Wikipedia project to eventually have a bot-free depth metric.

Table 3 shows the top ten Wikipedia editions with the most bot-made edits on their articles in the Wikipedia project, ordered by the percentage of bot automation each Wikipedia edition has. It is clear the Cebuano Wikipedia edition—the second Wikipedia edition in the globe rank in terms of the total number of articles has 94.05% of its total number of edits on its articles (content pages) are bot-made edits.

4 DEPTH$^+$ Metric

The depth metric is a useful indicator of Wikipedia's collaborative quality, which reflects how frequently a Wikipedia edition's articles are edited or updated by users (Wikimedia Foundation, 2023c). However, we believe the depth metric must be enhanced to solve some of the limitations spotlighted in this study.

LANGUAGE (CODE)	EDITS	BOT-EDITS	PERCENTAGE
Cebuano (ceb)	**34,900,283**	**32,822,497**	**94.05%**
Welsh (cy)	11,743,296	10,113,230	86.12%
Pali (pi)	101,934	85,498	83.88%
Norman (nrm)	219,464	172,629	78.66%
Waray (war)	6,420,883	4,962,642	77.29%
Buginese (bug)	202,056	154,684	76.56%
Chechen (ce)	9,638,638	7,375,144	76.52%
Minangkabau (min)	2,505,093	1,851,865	73.92%
Piedmontese (pms)	864,648	631,724	73.06%
Neapolitan (nap)	666,293	471,852	70.82%

Table 3: The top ten Wikipedia editions that have the most bot-made edits on their articles, ordered by the percentage of how much bot automation each Wikipedia edition has. We highlighted the Cebuano Wikipedia edition since it comes second in the global rank and has the highest bot-made edits on its articles (content pages). See Appendix D for the full list.

In the following subsections, we revise the original depth definitions after quantifying and removing bot activities, propose the DEPTH$^+$ metric as an enhanced depth metric for Wikipedia corpora quality, mathematically define its definitions, and highlight its key features.

4.1 Revision of Depth Definitions

To better reflect true collaborative activities in the DEPTH$^+$ metric, we will first remove the bot-created Wikipedia articles and the bot-made edits on the Wikipedia articles from the depth metric. We revisit the mathematical definitions of the depth metric and redefine the related metrics: edits, articles, and total pages accordingly.

Let all Wikipedia editions be $\mathcal{W}_i$, let the total number of edits of $\mathcal{W}_i$ be $\mathcal{E}_{W_i}$, let the total number of bot-made edits of $\mathcal{W}_i$ be $\mathcal{E}^b_{W_i}$ where $e^b = \{1, 2, 3, \ldots, n\}$, let the total number of articles of $\mathcal{W}_i$ be $\mathcal{A}_{W_i}$, let the total number of bot-created articles of $\mathcal{W}_i$ be $\mathcal{A}^b_{W_i}$ where $a^b = \{1, 2, 3, \ldots, n\}$, let the total number of non-articles of $\mathcal{W}_i$ be $\mathcal{R}_{W_i}$, and lastly, let the total number of pages of $\mathcal{W}_i$ be $\mathcal{T}_{W_i}$.

Therefore, the updated mathematical definitions of these metrics: edits, articles, and total pages of $\mathcal{W}_i$ using the mathematical representations after removing the bot activities are defined as follows:

$$\mathcal{E}_{W_i} = \mathcal{E}_{W_i} - \mathcal{E}^b_{W_i} \tag{5}$$

$$\mathcal{A}_{W_i} = \mathcal{A}_{W_i} - \mathcal{A}^b_{W_i} \tag{6}$$

$$\mathcal{T}_{W_i} = (\mathcal{A}_{W_i} - \mathcal{A}^b_{W_i}) + \mathcal{R}_{W_i} \tag{7}$$

4.2 Formulas of DEPTH$^+$ Metric

We understand that $(\frac{NonArticles}{Articles})$ from Equations 1 and 2 are to emphasize that the article count of a Wikipedia edition is just the tip of the iceberg, and other metrics, such as user pages, project pages, and discussion pages, are crucial indicators of "*Wikipedianness*" and the $(\frac{Edits}{Articles})$ from Equations 1 and 2 are also to emphasize that some Wikipedia editions might only include some copied and pasted articles or articles written by only one single registered user (which does not necessarily mean they are biased, but surely means they are not collaboratively edited, i.e., "*Wikipedian*") (Wikimedia Foundation, 2023b).

However, we propose a few significant additions to the depth metric's formulas. We first add a few available user-related metrics, like users, admins, and active users, to the DEPTH$^+$ metric and call them the "editors" metric. The difference between users and active users is that users refer to the number of user accounts regardless of current activity, whereas active users refer to registered users who have made at least one edit in the last thirty days (Wikipedia, 2023a). We add the active users over the users to normalize the measurements of the DEPTH$^+$ metric and add the admins as a constraint that gives the large Wikipedia editions higher priority, assuming that the larger the Wikipedia edition, the greater the number of admins.

The formula of the "editors" metric is defined as:

$$Editors = Admins \cdot \frac{ActiveUsers}{Users} \tag{8}$$

Secondly, we propose a few meaningful modifications to the depth metric's formulas, where we eliminate the square power of the depth simplified equation (in bold), Equation 2, $(\frac{NonArticles}{Articles})^{\mathbf{2}}$, because the square power will double the depth metric measurements, and we prefer to keep the DEPTH$^+$ metric values relatively small. We also eliminate the subtraction part of the stub ratio (in bold) from Equation 1, $(\mathbf{1}-\frac{Articles}{Total})$, because it was added to decrease the results of the stub ratio in 2007 (Wikimedia Foundation, 2023a), but now, it is irrelevant since we added the active users over the users to normalize the measurements of the DEPTH$^+$ metric and added the admins metric as a constraint to give large Wikipedia editions higher priority.

The DEPTH$^+$ metric is finally defined by combining the above modifications on Equations 1 and 2 with Equation 8 of the "editors" metric and inserting the revised mathematical definitions of metrics: edits, articles, and total pages from Equations 5, 6, and 7 to exclude the bot activities, as the following:

$$\text{DEPTH}^+ = Editors \cdot \frac{Edits \cdot NonArticles}{Articles^2} \cdot \frac{Articles}{Total} \tag{9}$$

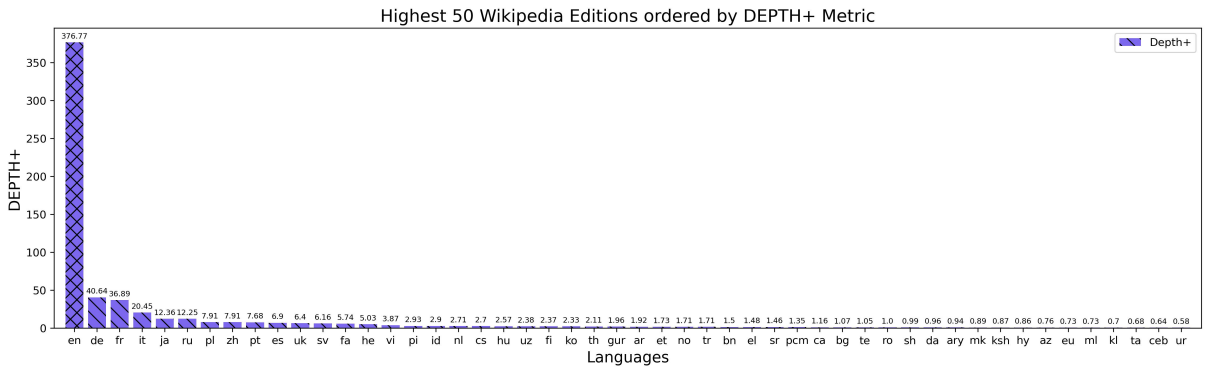

Figure 3: The highest 50 Wikipedia editions ordered by the DEPTH$^+$ metric values (all bot activities removed). We highlighted English Wikipedia since it is the largest Wikipedia edition. We can see that most languages in the highest ranks are either large or common Wikipedia editions. See Appendix E for the full list.

The DEPTH$^+$ metric can rearrange to a simplified equivalent formula as the following:

$$\text{DEPTH}^+ = Editors \cdot \frac{Edits}{Total} \cdot \frac{NonArticles}{Articles} \tag{10}$$

Let all Wikipedia editions be $\mathcal{W}_i$ where $i = \{1, 2, 3, \ldots, 320\}$ for the 320 open editions, let the total number of admins of $\mathcal{W}_i$ be $\mathcal{M}_{W_i}$ where $m = \{1, 2, 3, \ldots, n\}$, let the total number of active users of $\mathcal{W}_i$ be $\mathcal{V}_{W_i}$ where $v = \{1, 2, 3, \ldots, n\}$, let the total number of users of $\mathcal{W}_i$ be $\mathcal{U}_{W_i}$ where $u = \{1, 2, 3, \ldots, n\}$, and lastly, let the "editors" of $\mathcal{W}_i$ be $\mathcal{O}_{W_i}$.

Therefore, the mathematical definition of the "editors" metric of $\mathcal{W}_i$ using the mathematical representations is described as the following:

$$\mathcal{O}_{W_i} = \mathcal{M}_{W_i} \cdot \frac{\mathcal{V}_{W_i}}{\mathcal{U}_{W_i}} \tag{11}$$

Let the total number of edits of $\mathcal{W}_i$ be $\mathcal{E}_{W_i}$ where $e = \{1, 2, 3, \ldots, n\}$ (Equation 5), let the total number of articles of $\mathcal{W}_i$ be $\mathcal{A}_{W_i}$ where $a = \{1, 2, 3, \ldots, n\}$ (Equation 6), let the total number of non-articles of $\mathcal{W}_i$ be $\mathcal{R}_{W_i}$ where $r = \{1, 2, 3, \ldots, n\}$, and let the total number of pages of $\mathcal{W}_i$ be $\mathcal{T}_{W_i}$ where $\mathcal{T}_{W_i} = (\mathcal{A}_{W_i} - \mathcal{A}^b_{W_i}) + \mathcal{R}_{W_i}$ (Equation 7).

Therefore, the general mathematical definition of the DEPTH$^+$ metric of $\mathcal{W}_i$ using the mathematical representations is described as the following:

$$\mathcal{D}^+_{\mathcal{W}i} = \mathcal{O}_{W_i} \cdot \frac{\mathcal{E}_{W_i} \cdot \mathcal{R}_{W_i}}{\mathcal{A}^2_{W_i}} \cdot \frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}} \tag{12}$$

Lastly, the simplified mathematical definition of the DEPTH$^+$ metric of $\mathcal{W}_i$ using the mathematical representations is described as the following:

$$\mathcal{D}^+_{\mathcal{W}i} = \mathcal{O}_{W_i} \cdot \frac{\mathcal{E}_{W_i}}{\mathcal{T}_{W_i}} \cdot \frac{\mathcal{R}_{W_i}}{\mathcal{A}_{W_i}} \tag{13}$$

4.3 Features of DEPTH$^+$ Metric

The DEPTH$^+$ metric overcomes some of the drawbacks of the depth metric, employs Wikipedia's users-related metrics, and offers bot-free Wikipedia editions. Revisiting Figure 1, we see that the changes in the global rank for the top ten languages (editions) regarding the number of articles on the Wikipedia project when both metrics (depth and DEPTH$^+$) are applied, illustrating that the DEPTH$^+$ metric successfully prioritizes the large and most common Wikipedia editions.

Figure 3 shows the highest 50 Wikipedia editions ordered by the DEPTH$^+$ metric values after eliminating all bot activities (bot-generated articles and bot-made edits). Unlike the depth metric, we no longer use a somewhat arbitrary filtering step to disadvantage lower-resource languages. It makes sense that older, larger editions like English may have richer collaboration and depth, but using a filtering step to remove small languages does not seem fair. Small languages could have rich collaboration and depth as well. With the DEPTH$^+$ metric, we see that the English Wikipedia edition is at the top of the rank without filtering, followed by very large editions like German (`de`), French (`fr`), Italian (`it`), and Japanese (`ja`), but smaller languages still have the potential to score high on the DEPTH$^+$ ranking. For example, the Greenlandic Wikipedia edition was filtered in the depth metric, but with the DEPTH$^+$ metric, it is now among the top 50 Wikipedia editions. The DEPTH$^+$ metric successfully removes the bot-generated Wikipedia editions from the top of the rankings.

The original depth metric did not include any user-related metrics offered by the Wikipedia project, only focusing on the edits activities of the

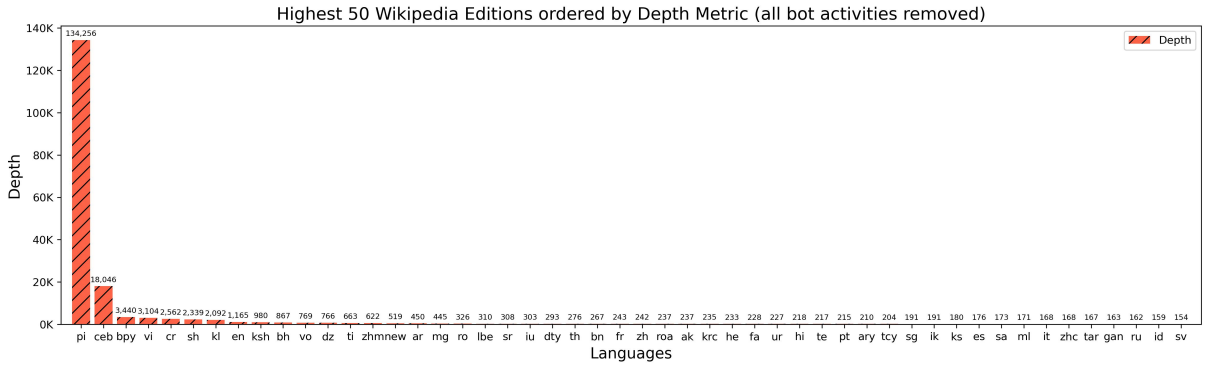

Figure 4: The highest 50 Wikipedia editions ordered by the depth metric after removing all bot activities. Even after removing the bot-activity, we can see that Wikipedia editions like Pali (`pi`), Cebuano (`ceb`), Bishnupriya (`bpy`), and Vietnamese (`vi`) still have unintuitively high depth values.

different types of pages (articles and non-articles) and neglecting the activity of the different types of users (users, admins, and active users) who contributed to these edits in the first place. The DEPTH$^+$ metric introduced the "editors" metric (see Equation 8), which utilizes these metrics: admins, users, and active users and actively puts the users at the heart of the DEPTH$^+$ metric.

We quantified the bot activities of creating articles (content pages) and the bot activities of editing those articles. We also successfully integrated our quantifications of the bot activities into the DEPTH$^+$ metric. We also found that the DEPTH$^+$ metric is more robust than the original depth metric when we remove all bot activities and apply the two metrics. The DEPTH$^+$ metric returns mostly identical calculations when we include or remove all bot activities from the metric, whereas the original depth metric returns completely different questionable calculations, as shown in Figure 4.

5 Limitations

The DEPTH$^+$ metric resolves the problem of bot-generated Wikipedia editions that have many bot-created articles and bot-made edits on their articles. Yet, the DEPTH$^+$ metric does not fix the problem of automatically translated Wikipedia editions in the Wikipedia project that their articles have been largely translated by poor direct translation or shallow template-based translation. The quantifications of these automatically translated Wikipedia editions in the Wikipedia project cannot be carried out as systematically as the bot-generated Wikipedia editions, and examining each Wikipedia edition separately is the only way to accomplish such quantification. Another limitation of the DEPTH$^+$ metric is depending on the active users metric, which dynamically decreases the DEPTH$^+$ metric values if there are no editing activities on the articles in the last thirty days. We preferred to use the total unique users who made at least one edit but do not have that figure, so we are approximating it with the already calculated active users metric by the Wikipedia project.

6 Related Work

Due to the widespread use of Wikipedia articles as training corpora for many NLP toolchains, especially for low-resource languages, many researchers have addressed the importance of transparency in the Wikipedia project, encouraged the transparency values in the project, and proposed improvements on accountability and social transparency through visualizations. For example, Suh et al. (2008) presented a social dynamic analysis tool called "WikiDashboard" to improve the social transparency and accountability of Wikipedia articles. This tool aims to enhance the interpretation, communication, and trustworthiness of Wikipedia articles by visualizing the social dynamics and editing patterns of every article and editor in the Wikipedia project.

Biuk-Aghai et al. (2014) also studied the visualization of large-scale human collaboration on the Wikipedia project, analyzed the co-authoring across the entire Wikipedia editions in various languages (English, German, Chinese, Swedish, and Danish), and found it to follow a geometric distribution in all the investigated language editions. To better understand the geometric distribution of co-author counts across different topics on the Wikipedia project, they aggregated Wikipedia

content by category and visualized it in a form resembling a geographic map. These geographically looking map visualizations show significant differences in co-author counts across different topics in all the visualized Wikipedia language editions.

At the intersection of transparency and underrepresentation in the Wikipedia project, Wali et al. (2020) discussed the available Wikipedia corpora for eight languages: English, Chinese, Arabic, Urdu, Farsi, French, Spanish, and Wolof. They closely examined the typical NLP pipeline and highlighted that significant limitations persist even when a language is technically supported, hindering full participation. They specifically compared the number of language speakers to the number of articles in the respective Wikipedia edition, using the "Articles/1000 Speakers" metric. Despite the dedicated efforts of numerous Wikipedia contributors who have invested substantially in compiling a vast multilingual dataset, not all language speakers have equal opportunities to contribute to the Wikipedia project.

7 Conclusion and Future Work

We have discussed Wikipedia's current depth metric in detail, rewritten its mathematical representations, and underlined the limitations of its representation of the depth of collaboration in Wikipedia corpora. We also quantified the bot activities in the Wikipedia project and excluded the bot-created articles and the bot-made edits on Wikipedia articles. We lastly proposed the DEPTH$^+$ metric, defined its formal definitions, and highlighted its features, including a better representation of the depth of collaborativeness, a user-centered depth metric, and bot-free Wikipedia editions after the removal of the bot-generated articles and the bot-made edits on those Wikipedia editions' articles.

We hypothesize that a metric that is a better measure of authentic human collaborativeness will be a better measure of the degree to which corpora authentically represents the language and the culture of native speakers. One key aspect of our future work is to find ways to test this hypothesis. Specifically, we aim to examine the performance and societal implications of training LLMs on unrepresentative and inorganic corpora, particularly on the bot-generated Wikipedia articles.

Reproducibility

Data collection, implementation of the DEPTH$^+$ metric, and an expanded technical report can be found on GitHub at https://github.com/SaiedAlshahrani/DEPTHplus.

References

Saied Alshahrani, Esma Wali, and Jeanna Matthews. 2022. Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 361–371, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Robert P. Biuk-Aghai, Cheong-Iao Pang, and Yain-Whar Si. 2014. Visualizing Large-scale Human Collaboration in Wikipedia. *Future Generation Computer Systems*, 31:120–133. Special Section: Advances in Computer Supported Collaboration: Systems and Technologies.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

National Public Radio (NPR). 2022. What Is a Recession? Wikipedia Can't Decide. Last accessed on 2023-6-1.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases? *CoRR*, abs/1909.01066.

Popular Science. 2014. This Bot Has Written More Wikipedia Articles Than Anybody. Last accessed on 2023-6-1.

Bongwon Suh, Ed H. Chi, Aniket Kittur, and Bryan A. Pendleton. 2008. Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with Wikidashboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1037–1040, New York, NY, USA. Association for Computing Machinery.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. 2020. Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages. *arXiv preprint arXiv:2007.05872*.

Wikimedia Commons. 2023. Data: Wikipedia Statistics/meta.tab. Last accessed on 2023-6-1.

Wikimedia Foundation. 2019. Wikipedia Statistics v1: Bot Article Creations Only. Last accessed on 2023-6-1.

Wikimedia Foundation. 2023a. Depth 2.0. Last accessed on 2023-6-1.

Wikimedia Foundation. 2023b. Please Delete or Redefine "Depth". Last accessed on 2023-6-1.

Wikimedia Foundation. 2023c. Wikipedia Article Depth. Last accessed on 2023-6-1.

Wikimedia Foundation. 2023d. Wikipedia Statistics v2. Last accessed on 2023-6-1.

Wikimedia Meta-Wiki. 2023. List of Wikipedias: Closed and Read-only. Last accessed on 2023-6-1.

Wikipedia. 2022. Wikipedia: History of Wikipedia Bots. Last accessed on 2023-6-1.

Wikipedia. 2023a. List of Wikipedias. Last accessed on 2023-6-1.

Wikipedia. 2023b. Lsjbot. Last accessed on 2023-6-1.

Wikipedia. 2023c. Wikipedia: Bot Policy. Last accessed on 2023-6-1.

Wikipedia. 2023d. Wikipedia: Bots. Last accessed on 2023-6-1.

Full Derivation of Depth's Formulas

Let $\mathcal{W}_i$ represent Wikipedia editions, let the number of edits be $\mathcal{E}_{W_i}$, let the number of articles be $\mathcal{A}_{W_i}$, let the number of non-articles be $\mathcal{R}_{W_i}$, and lastly, let the number of pages be $\mathcal{T}_{W_i}$. We, next, show the full derivation of the depth's formulas.

$$\mathcal{D}_{W_i} = \frac{\mathcal{E}_{W_i}}{\mathcal{A}_{W_i}} \cdot \frac{\mathcal{R}_{W_i}}{\mathcal{A}_{W_i}} \cdot \left(1 - \frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right) \text{ Original Equation (1)}$$

First, we transform the third factor (stub ratio), $\left(1 - \frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right)$, into $\left(\frac{\mathcal{R}_{W_i}}{\mathcal{T}_{W_i}}\right)$:

$$\Rightarrow \left(1 - \frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right) \Rightarrow \left(\frac{\mathcal{T}_{W_i}}{\mathcal{T}_{W_i}}\right) - \left(\frac{\mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right)$$

Since $\mathcal{T}_{W_i} = \mathcal{A}_{W_i} + \mathcal{R}_{W_i}$, then, $\mathcal{R}_{W_i} = \mathcal{T}_{W_i} - \mathcal{A}_{W_i}$

$$\Rightarrow \left(\frac{\mathcal{T}_{W_i} - \mathcal{A}_{W_i}}{\mathcal{T}_{W_i}}\right) \Rightarrow \left(\frac{\mathcal{R}_{W_i}}{\mathcal{T}_{W_i}}\right)$$

Second, we insert $\left(\frac{\mathcal{R}_{W_i}}{\mathcal{T}_{W_i}}\right)$ in the original depth's formula (Equation 1) to get the simplified formula:

$$\mathcal{D}_{W_i} = \left(\frac{\mathcal{E}_{W_i}}{\mathcal{A}_{W_i}}\right) \cdot \left(\frac{\mathcal{R}_{W_i}}{\mathcal{A}_{W_i}}\right) \cdot \left(\frac{\mathcal{R}_{W_i}}{\mathcal{T}_{W_i}}\right) \text{ By Rearranging}$$

$$\Rightarrow \frac{\mathcal{E}_{W_i}}{\mathcal{T}_{W_i}} \cdot \left(\frac{\mathcal{R}_{W_i}}{\mathcal{A}_{W_i}}\right)^2 \text{ The Simplified Equation (2)}$$

■

Appendix A: Global Rank of Wikipedia Editions

Rank	Language	Code	Articles	Rank	Language	Code	Articles	Rank	Language	Code	Articles
1	English	(en)	6,642,196	111	Gujarati	(gu)	30,117	221	Konkani (Goan Konkani)	(gom)	3,570
2	Cebuano	(ceb)	6,123,587	112	Interlingua	(ia)	29,924	222	Permyak	(koi)	3,443
3	German	(de)	2,790,340	113	Kannada	(kn)	29,882	223	Extremaduran	(ext)	3,415
4	Swedish	(sv)	2,561,243	114	Alemannic German	(als)	29,750	224	Tuvan	(tyv)	3,395
5	French	(fr)	2,512,610	115	Kotava	(avk)	27,029	225	Lower Sorbian	(dsb)	3,336
6	Dutch	(nl)	2,120,283	116	Bavarian	(bar)	26,901	226	Avar	(av)	3,334
7	Russian	(ru)	1,907,471	117	Sicilian	(scn)	26,240	227	Lingala	(ln)	3,326
8	Spanish	(es)	1,853,145	118	Bishnupriya Manipuri	(bpy)	25,087	228	Doteli	(dty)	3,324
9	Italian	(it)	1,806,143	119	Hausa	(ha)	24,383	229	Karakalpak	(kaa)	3,243
10	Egyptian Arabic	(arz)	1,617,246	120	Crimean Tatar	(crh)	23,938	230	Papiamento	(pap)	3,148
11	Polish	(pl)	1,563,797	121	Quechua (Southern Quechua)	(qu)	23,383	231	Chavacano (Zamboanga)	(cbk-zam)	3,128
12	Japanese	(ja)	1,369,714	122	Navajo	(nv)	22,069	232	Maldivian	(dv)	3,024
13	Chinese	(zh)	1,345,918	123	Mongolian	(mn)	21,999	233	Moksha	(mdf)	2,963
14	Vietnamese	(vi)	1,282,386	124	Mingrelian	(xmf)	19,999	234	Ripuarian	(ksh)	2,940
15	Waray	(war)	1,266,100	125	Sinhala	(si)	18,556	235	Twi	(tw)	2,896
16	Ukrainian	(uk)	1,257,759	126	Balinese	(ban)	18,342	236	Gagauz	(gag)	2,803
17	Arabic	(ar)	1,204,339	127	Pashto	(ps)	17,408	237	Kashmiri	(ks)	2,777
18	Portuguese	(pt)	1,101,393	128	North Frisian	(frr)	17,155	238	Buryat (Russia Buriat)	(bxr)	2,772
19	Persian	(fa)	958,816	129	Samogitian	(bat-smg)	17,147	239	Palatine German	(pfl)	2,741
20	Catalan	(ca)	724,808	130	Ossetian	(os)	16,962	240	Luganda	(lg)	2,689
21	Serbian	(sr)	669,768	131	Odia	(or)	16,611	241	Zhuang (Standard Zhuang)	(za)	2,568
22	Indonesian	(id)	643,081	132	Yakut	(sah)	16,377	242	Pali	(pi)	2,548
23	Korean	(ko)	630,546	133	Eastern Min	(cdo)	15,927	243	Pangasinan	(pag)	2,504
24	Norwegian (Bokmål)	(no)	608,985	134	Scottish Gaelic	(gd)	15,920	244	Sakizaya	(szy)	2,502
25	Chechen	(ce)	599,686	135	Buginese	(bug)	15,823	245	Hawaiian	(haw)	2,494
26	Finnish	(fi)	550,503	136	Yiddish	(yi)	15,502	246	Awadhi	(awa)	2,436
27	Hungarian	(hu)	523,645	137	Sindhi	(sd)	15,379	247	Atayal	(tay)	2,421
28	Czech	(cs)	522,302	138	Ilocano	(ilo)	15,375	248	Pa'O	(blk)	2,295
29	Turkish	(tr)	517,602	139	Amharic	(am)	15,189	249	Ingush	(inh)	2,166
30	Tatar	(tt)	499,963	140	Neapolitan	(nap)	14,778	250	Karachay-Balkar	(krc)	2,065
31	Serbo-Croatian	(sh)	457,985	141	Mazanderani	(mzn)	14,428	251	Kalmyk Oirat	(xal)	2,048
32	Romanian	(ro)	437,712	142	Limburgish	(li)	14,276	252	Pennsylvania Dutch	(pdc)	2,003
33	Southern Min	(zh-min-nan)	432,436	143	Gorontalo	(gor)	13,894	253	Tongan	(to)	1,955
34	Basque	(eu)	409,627	144	Upper Sorbian	(hsb)	13,891	254	Atikamekw	(atj)	1,949
35	Malay	(ms)	364,205	145	Faroese	(fo)	13,889	255	Aramaic (Syriac)	(arc)	1,887
36	Esperanto	(eo)	334,673	146	Banyumasan	(map-bms)	13,845	256	Tulu	(tcy)	1,855
37	Hebrew	(he)	332,783	147	Igbo	(ig)	13,781	257	Mon	(mnw)	1,763
38	Armenian	(hy)	296,647	148	Maithili	(mai)	13,731	258	Jamaican Patois	(jam)	1,705
39	Danish	(da)	290,726	149	Central Bikol	(bcl)	13,522	259	Kabiye	(kbp)	1,697
40	Bulgarian	(bg)	289,861	150	Emilian-Romagnol	(eml)	13,029	260	Nauruan	(na)	1,670
41	Welsh	(cy)	278,635	151	Shan	(shn)	12,743	261	Wolof	(wo)	1,650
42	Slovak	(sk)	244,334	152	Acehnese	(ace)	12,725	262	Kabardian	(kbd)	1,597
43	South Azerbaijani	(azb)	242,972	153	Classical Chinese	(zh-classical)	12,294	263	Nias	(nia)	1,569
44	Estonian	(et)	235,273	154	Sanskrit	(sa)	11,974	264	Novial	(nov)	1,530
45	Kazakh	(kk)	233,210	155	Walloon	(wa)	11,755	265	Shilha	(shi)	1,522
46	Belarusian	(be)	230,170	156	Assamese	(as)	11,572	266	Kikuyu	(ki)	1,505
47	Simple English	(simple)	228,588	157	Interlingue	(ie)	11,560	267	N'Ko	(nqo)	1,465
48	Minangkabau	(min)	226,589	158	Ligurian	(lij)	11,122	268	Bislama	(bi)	1,408
49	Uzbek	(uz)	224,124	159	Zulu	(zu)	10,909	269	Tok Pisin	(tpi)	1,359
50	Greek	(el)	219,052	160	Meadow Mari	(mhr)	10,758	270	Tetum	(tet)	1,347
51	Croatian	(hr)	214,365	161	Western Armenian	(hyw)	10,623	271	Lojban	(jbo)	1,325
52	Lithuanian	(lt)	209,617	162	Fiji Hindi	(hif)	10,483	272	Aromanian	(roa-rup)	1,302
53	Galician	(gl)	195,667	163	Hill Mari	(mrj)	10,430	273	Xhosa	(xh)	1,289
54	Azerbaijani	(az)	193,432	164	Shona	(sn)	10,417	274	Fijian	(fj)	1,277
55	Urdu	(ur)	188,660	165	Banjarese	(bjn)	10,280	275	Lak	(lbe)	1,264
56	Slovene	(sl)	180,603	166	Meitei	(mni)	10,220	276	Kongo (Kituba)	(kg)	1,264
57	Georgian	(ka)	166,967	167	Khmer	(km)	10,077	277	Oromo	(om)	1,258
58	Norwegian (Nynorsk)	(nn)	164,952	168	Hakka Chinese	(hak)	10,043	278	Tahitian	(ty)	1,202
59	Hindi	(hi)	156,119	169	Tumbuka	(tum)	9,950	279	Gun	(guw)	1,199
60	Thai	(th)	155,115	170	Tarantino	(roa-tara)	9,317	280	Old Church Slavonic	(cu)	1,192
61	Tamil	(ta)	153,462	171	Somali	(so)	9,226	281	Seediq	(trv)	1,130
62	Latin	(la)	137,710	172	Kapampangan	(pam)	8,882	282	Sranan Tongo	(srn)	1,117
63	Bengali	(bn)	137,028	173	Rusyn	(rue)	8,631	283	Samoan	(sm)	1,073
64	Macedonian	(mk)	135,485	174	Northern Sotho	(nso)	8,546	284	Southern Altai	(alt)	1,063
65	Asturian	(ast)	132,057	175	Bihari (Bhojpuri)	(bh)	8,311	285	French Guianese Creole	(gcr)	1,059
66	Cantonese	(zh-yue)	130,956	176	Santali	(sat)	8,210	286	Cherokee	(chr)	1,052
67	Ladin	(lld)	130,202	177	Northern Sámi	(se)	7,841	287	Latgalian	(ltg)	1,040
68	Latvian	(lv)	119,331	178	Erzya	(myv)	7,797	288	Tswana	(tn)	1,027
69	Tajik	(tg)	109,497	179	Māori	(mi)	7,787	289	Chewa	(ny)	1,021
70	Afrikaans	(af)	107,494	180	West Flemish	(vls)	7,773	290	Madurese	(mad)	1,015
71	Burmese	(my)	106,322	181	Dutch Low Saxon	(nds-nl)	7,640	291	Sotho	(st)	912
72	Malagasy	(mg)	95,465	182	Nahuatl	(nah)	7,566	292	Norfuk	(pih)	895
73	Bosnian	(bs)	91,729	183	Sardinian	(sc)	7,384	293	Gothic	(got)	872
74	Marathi	(mr)	91,214	184	Cornish	(kw)	7,238	294	Ewe	(ee)	822
75	Albanian	(sq)	89,168	185	Gilaki	(glk)	6,810	295	Amis	(ami)	816
76	Occitan	(oc)	88,515	186	Veps	(vep)	6,780	296	Romani (Vlax Romani)	(rmy)	814
77	Low German	(nds)	84,178	187	Kabyle	(kab)	6,691	297	Bambara	(bm)	785
78	Malayalam	(ml)	83,364	188	Turkmen	(tk)	6,678	298	Fula	(ff)	763
79	Belarusian (Taraškievica)	(be-tarask)	82,176	189	Gan Chinese	(gan)	6,596	299	Venda	(ve)	753
80	Telugu	(te)	81,962	190	Moroccan Arabic	(ary)	6,593	300	Tsonga	(ts)	732
81	Kyrgyz	(ky)	80,368	191	Corsican	(co)	6,533	301	Cheyenne	(chy)	697
82	Breton	(br)	79,098	192	Dagbani	(dag)	6,489	302	Swazi	(ss)	637
83	Swahili	(sw)	76,736	193	Võro	(fiu-vro)	6,451	303	Kirundi	(rn)	627
84	Javanese	(jv)	72,462	194	Lhasa Tibetan	(bo)	6,395	304	Tyap	(kcg)	626
85	Newar	(new)	72,348	195	Abkhaz	(ab)	6,045	305	Nigerian Pidgin	(pcm)	614
86	Venetian	(vec)	69,152	196	Manx	(gv)	5,875	306	Chamorro	(ch)	546
87	Haitian Creole	(ht)	68,387	197	Saraiki	(skr)	5,710	307	Iñupiaq	(ik)	503
88	Western Punjabi	(pnb)	68,353	198	Zeelandic	(zea)	5,672	308	Pontic Greek	(pnt)	486
89	Piedmontese	(pms)	67,867	199	Franco-Provençal	(frp)	5,670	309	Wayuu	(guc)	467
90	Bashkir	(ba)	62,498	200	Uyghur	(ug)	5,655	310	Adyghe	(ady)	464
91	Luxembourgish	(lb)	61,650	201	Kinyarwanda	(rw)	5,607	311	Inuktitut	(iu)	449
92	Sundanese	(su)	61,417	202	Udmurt	(udm)	5,536	312	Akan	(ak)	417
93	Kurdish (Kurmanji)	(ku)	59,045	203	Picard	(pcd)	5,517	313	Paiwan	(pwn)	325
94	Irish	(ga)	58,411	204	Komi	(kv)	5,501	314	Sango	(sg)	314
95	Lombard	(lmo)	57,550	205	Kashubian	(csb)	5,450	315	Dinka	(din)	308
96	Silesian	(szl)	56,862	206	Maltese	(mt)	5,276	316	Tigrinya	(ti)	256
97	Icelandic	(is)	56,288	207	Guarani	(gn)	5,192	317	Greenlandic	(kl)	242
98	West Frisian	(fy)	51,147	208	Inari Sámi	(smn)	5,062	318	Dzongkha	(dz)	237
99	Chuvash	(cv)	50,963	209	Aymara	(ay)	5,034	319	Frafra	(gur)	216
100	Kurdish (Sorani)	(ckb)	49,046	210	Norman	(nrm)	4,834	320	Cree	(cr)	161
101	Punjabi	(pa)	46,000	211	Lezgian	(lez)	4,318				
102	Tagalog	(tl)	44,438	212	Lingua Franca Nova	(lfn)	4,196				
103	Aragonese	(an)	43,635	213	Livvi-Karelian	(olo)	4,100				
104	Wu Chinese	(wuu)	42,796	214	Saterland Frisian	(stq)	4,095				
105	Zaza	(diq)	40,348	215	Mirandese	(mwl)	3,982				
106	Ido	(io)	37,346	216	Lao	(lo)	3,969				
107	Scots	(sco)	36,127	217	Old English	(ang)	3,919				
108	Volapük	(vo)	33,272	218	Friulian	(fur)	3,841				
109	Yoruba	(yo)	32,285	219	Romansh	(rm)	3,757				
110	Nepali	(ne)	31,407	220	Judaeo-Spanish	(lad)	3,625				

Appendix B: Calculations of Depth Metric of Wikipedia Editions

#	Language	Code	Depth	#	Language	Code	Depth	#	Language	Code	Depth
1	Cree	(cr)	2768.85	111	Kashmiri	(ks)	53.97	221	Pali	(pi)	14.97
2	Greenlandic	(kl)	2306.11	112	Cheyenne	(chy)	53.72	222	Latin	(la)	14.01
3	English	(en)	1178.29	113	Scots	(sco)	51.85	223	Kabyle	(kab)	13.71
4	Dzongkha	(dz)	1164.88	114	Kurdish (Sorani)	(ckb)	51.74	224	French Guianese Creole	(gcr)	13.64
5	Ripuarian	(ksh)	1026.62	115	Latgalian	(ltg)	51.66	225	Lombard	(lmo)	13.5
6	Tigrinya	(ti)	840.86	116	Oromo	(om)	50.25	226	North Frisian	(frr)	13.42
7	Serbo-Croatian	(sh)	745.52	117	Czech	(cs)	50.01	227	Kazakh	(kk)	12.89
8	Vietnamese	(vi)	718.92	118	Khmer	(km)	49.37	228	Basque	(eu)	12.86
9	Bihari (Bhojpuri)	(bh)	609.06	119	Armenian	(hy)	48.86	229	Emilian-Romagnol	(eml)	12.74
10	Inuktitut	(iu)	499.13	120	Frafra	(gur)	48.62	230	Kinyarwanda	(rw)	12.33
11	Lak	(lbe)	405.72	121	Dinka	(din)	48.59	231	Võro	(fiu-vro)	11.78
12	Thai	(th)	324.41	122	Norwegian (Bokmål)	(no)	46.68	232	Gun	(guw)	11.59
13	Bengali	(bn)	316.61	123	Yiddish	(yi)	45.87	233	Western Armenian	(hyw)	11.01
14	Sango	(sg)	294.83	124	Franco-Provençal	(frp)	45.82	234	Breton	(br)	10.87
15	Doteli	(dty)	294.58	125	West Flemish	(vls)	45.68	235	Malagasy	(mg)	10.68
16	Iñupiaq	(ik)	276.05	126	Dutch Low Saxon	(nds-nl)	45.16	236	Nias	(nia)	10.65
17	Volapük	(vo)	268.15	127	Corsican	(co)	44.43	237	Neapolitan	(nap)	10.5
18	Hebrew	(he)	267.24	128	Afrikaans	(af)	43.1	238	Cantonese	(zh-yue)	10.48
19	Karachay-Balkar	(krc)	263.63	129	Romansh	(rm)	41.49	239	Tajik	(tg)	10.36
20	Moroccan Arabic	(ary)	260.13	130	Mon	(mnw)	40.57	240	Asturian	(ast)	10.2
21	Aromanian	(roa-rup)	257.99	131	Lao	(lo)	40.5	241	Banjarese	(bjn)	10.05
22	French	(fr)	256.53	132	Northern Sámi	(se)	40.18	242	Luganda	(lg)	10.05
23	Arabic	(ar)	249.63	133	Marathi	(mr)	40.01	243	Banyumasan	(map-bms)	10.0
24	Akan	(ak)	247.48	134	Tahitian	(ty)	39.67	244	Santali	(sat)	9.88
25	Hindi	(hi)	226.48	135	Azerbaijani	(az)	39.63	245	Rusyn	(rue)	9.84
26	Lojban	(jbo)	217.06	136	Albanian	(sq)	39.63	246	Kurdish (Kurmanji)	(ku)	9.67
27	Chinese	(zh)	206.98	137	Abkhaz	(ab)	39.43	247	Swahili	(sw)	9.61
28	Old Church Slavonic	(cu)	206.93	138	Catalan	(ca)	38.99	248	Occitan	(oc)	9.35
29	Tulu	(tcy)	205.46	139	Finnish	(fi)	38.91	249	Livvi-Karelian	(olo)	9.32
30	Spanish	(es)	201.37	140	Xhosa	(xh)	37.71	250	Zeelandic	(zea)	9.22
31	Classical Chinese	(zh-classical)	200.51	141	Komi	(kv)	37.55	251	Newar	(new)	8.78
32	Gan Chinese	(gan)	198.39	142	Lingala	(ln)	36.88	252	Wayuu	(guc)	8.6
33	Malayalam	(ml)	195.14	143	Chavacano (Zamboanga)	(cbk-zam)	36.84	253	Welsh	(cy)	8.49
34	Portuguese	(pt)	189.63	144	Udmurt	(udm)	35.85	254	Venetian	(vec)	8.49
35	Italian	(it)	183.02	145	Tamil	(ta)	35.59	255	Pa'O	(blk)	8.42
36	Pontic Greek	(pnt)	179.19	146	Nigerian Pidgin	(pcm)	35.11	256	Mazanderani	(mzn)	8.27
37	Kalmyk Oirat	(xal)	176.18	147	Venda	(ve)	34.82	257	Chuvash	(cv)	7.91
38	Ewe	(ee)	172.62	148	Belarusian (Taraškievica)	(be-tarask)	34.77	258	Kashubian	(csb)	7.88
39	Tsonga	(ts)	172.03	149	Sakizaya	(szy)	34.38	259	Chechen	(ce)	7.81
40	Korean	(ko)	168.02	150	Somali	(so)	34.0	260	Māori	(mi)	7.63
41	Turkish	(tr)	167.64	151	Amharic	(am)	33.74	261	Kikuyu	(ki)	7.49
42	Sanskrit	(sa)	167.11	152	Faroese	(fo)	33.65	262	Zhuang (Standard Zhuang)	(za)	6.97
43	Serbian	(sr)	159.57	153	Erzya	(myv)	33.6	263	Paiwan	(pwn)	6.96
44	Tagalog	(tl)	159.36	154	Fijian	(fj)	33.55	264	Tarantino	(roa-tara)	6.87
45	Tok Pisin	(tpi)	155.44	155	Polish	(pl)	33.0	265	Acehnese	(ace)	6.58
46	Russian	(ru)	153.38	156	Friulian	(fur)	32.94	266	Awadhi	(awa)	6.54
47	Romanian	(ro)	151.6	157	Bishnupriya Manipuri	(bpy)	32.62	267	Samogitian	(bat-smg)	6.4
48	Persian	(fa)	149.61	158	Yakut	(sah)	32.54	268	Southern Min	(zh-min-nan)	6.4
49	Assamese	(as)	149.16	159	Shilha	(shi)	32.17	269	Zulu	(zu)	6.38
50	Adyghe	(ady)	148.89	160	Uzbek	(uz)	31.81	270	Picard	(pcd)	5.76
51	Novial	(nov)	144.49	161	Georgian	(ka)	31.6	271	Aymara	(ay)	5.71
52	Gothic	(got)	138.9	162	Lezgian	(lez)	31.41	272	Hakka Chinese	(hak)	5.63
53	Old English	(ang)	137.06	163	Icelandic	(is)	31.27	273	Irish	(ga)	5.55
54	Swazi	(ss)	135.82	164	Sindhi	(sd)	30.69	274	Low German	(nds)	5.52
55	Indonesian	(id)	129.0	165	Amis	(ami)	30.57	275	Yoruba	(yo)	5.48
56	Tyap	(kcg)	123.32	166	Turkmen	(tk)	30.53	276	Gilaki	(glk)	5.43
57	Manx	(gv)	122.74	167	Palatine German	(pfl)	30.41	277	South Azerbaijani	(azb)	5.03
58	Chamorro	(ch)	120.38	168	Sranan Tongo	(srn)	30.21	278	Kabiye	(kbp)	4.7
59	Ingush	(inh)	119.9	169	West Frisian	(fy)	30.04	279	Burmese	(my)	4.51
60	Bambara	(bm)	118.45	170	Saterland Frisian	(stq)	30.0	280	N'Ko	(nqo)	4.51
61	Chewa	(ny)	112.15	171	Slovene	(sl)	29.96	281	Jamaican Patois	(jam)	4.32
62	Romani (Vlax Romani)	(rmy)	112.05	172	Galician	(gl)	29.78	282	Ido	(io)	4.2
63	Maltese	(mt)	109.66	173	Pangasinan	(pag)	29.75	283	Madurese	(mad)	4.15
64	Judaeo-Spanish	(lad)	108.38	174	Uyghur	(ug)	29.38	284	Balinese	(ban)	4.01
65	Kannada	(kn)	105.45	175	Permyak	(koi)	28.87	285	Shona	(sn)	3.87
66	Urdu	(ur)	103.66	176	Alemannic German	(als)	28.61	286	Mingrelian	(xmf)	3.83
67	Telugu	(te)	102.06	177	Pashto	(ps)	28.52	287	Hill Mari	(mrj)	3.76
68	Wolof	(wo)	99.44	178	Lithuanian	(lt)	28.38	288	Navajo	(nv)	3.6
69	Cherokee	(chr)	99.1	179	Extremaduran	(ext)	28.1	289	Waray	(war)	3.59
70	Norfuk	(pih)	98.77	180	Kapampangan	(pam)	27.79	290	Shan	(shn)	3.57
71	Sotho	(st)	97.42	181	Norman	(nrm)	27.77	291	Crimean Tatar	(crh)	3.5
72	German	(de)	92.97	182	Bulgarian	(bg)	27.58	292	Interlingua	(ia)	3.46
73	Limburgish	(li)	91.92	183	Hawaiian	(haw)	26.92	293	Minangkabau	(min)	3.26
74	Mongolian	(mn)	90.23	184	Walloon	(wa)	26.34	294	Eastern Min	(cdo)	3.25
75	Japanese	(ja)	88.54	185	Upper Sorbian	(hsb)	26.0	295	Hausa	(ha)	3.07
76	Maldivian	(dv)	88.53	186	Kongo (Kituba)	(kg)	25.93	296	Western Punjabi	(pnb)	2.88
77	Fula	(ff)	88.13	187	Seediq	(trv)	25.89	297	Sundanese	(su)	2.34
78	Aramaic (Syriac)	(arc)	87.78	188	Inari Sámi	(smn)	25.76	298	Piedmontese	(pms)	2.29
79	Bosnian	(bs)	87.24	189	Lhasa Tibetan	(bo)	25.7	299	Zaza	(diq)	2.24
80	Southern Altai	(alt)	86.38	190	Luxembourgish	(lb)	25.13	300	Cebuano	(ceb)	2.16
81	Kirundi	(rn)	82.29	191	Mirandese	(mwl)	24.57	301	Tumbuka	(tum)	2.14
82	Sinhala	(si)	82.14	192	Tongan	(to)	24.5	302	Lingua Franca Nova	(lfn)	2.07
83	Macedonian	(mk)	81.12	193	Belarusian	(be)	24.03	303	Dagbani	(dag)	2.05
84	Odia	(or)	77.43	194	Bislama	(bi)	23.77	304	Atikamekw	(atj)	1.62
85	Avar	(av)	77.36	195	Punjabi	(pa)	23.42	305	Igbo	(ig)	1.44
86	Tswana	(tn)	75.6	196	Quechua (Southern Quechua)	(qu)	23.15	306	Interlingue	(ie)	1.14
87	Latvian	(lv)	73.82	197	Bashkir	(ba)	22.99	307	Meitei	(mni)	1.13
88	Kabardian	(kbd)	72.83	198	Tuvan	(tyv)	22.71	308	Tatar	(tt)	0.86
89	Ilocano	(ilo)	72.3	199	Slovak	(sk)	22.2	309	Gorontalo	(gor)	0.78
90	Lower Sorbian	(dsb)	71.74	200	Twi	(tw)	22.18	310	Buginese	(bug)	0.73
91	Bavarian	(bar)	71.71	201	Maithili	(mai)	21.62	311	Atayal	(tay)	0.68
92	Nahuatl	(nah)	71.59	202	Central Bikol	(bcl)	21.37	312	Wu Chinese	(wuu)	0.58
93	Veps	(vep)	70.71	203	Estonian	(et)	21.27	313	Kyrgyz	(ky)	0.57
94	Moksha	(mdf)	68.77	204	Javanese	(jv)	20.41	314	Haitian Creole	(ht)	0.49
95	Nauruan	(na)	68.56	205	Karakalpak	(kaa)	19.92	315	Northern Sotho	(nso)	0.34
96	Pennsylvania Dutch	(pdc)	68.07	206	Malay	(ms)	19.71	316	Egyptian Arabic	(arz)	0.3
97	Fiji Hindi	(hif)	67.54	207	Guarani	(gn)	19.03	317	Silesian	(szl)	0.29
98	Gujarati	(gu)	64.5	208	Gagauz	(gag)	19.03	318	Kotava	(avk)	0.2
99	Ossetian	(os)	62.19	209	Scottish Gaelic	(gd)	18.61	319	Saraiki	(skr)	0.06
100	Aragonese	(an)	61.1	210	Dutch	(nl)	18.19	320	Ladin	(lld)	0.0
101	Hungarian	(hu)	59.95	211	Ligurian	(lij)	17.78				
102	Nepali	(ne)	59.94	212	Croatian	(hr)	17.67				
103	Simple English	(simple)	59.52	213	Meadow Mari	(mhr)	17.65				
104	Greek	(el)	59.26	214	Papiamento	(pap)	17.14				
105	Tetum	(tet)	58.81	215	Sardinian	(sc)	17.11				
106	Danish	(da)	58.53	216	Swedish	(sv)	16.79				
107	Samoan	(sm)	57.51	217	Sicilian	(scn)	16.76				
108	Buryat (Russia Buriat)	(bxr)	57.05	218	Cornish	(kw)	16.24				
109	Konkani (Goan Konkani)	(gom)	57.0	219	Esperanto	(eo)	16.19				
110	Ukrainian	(uk)	54.57	220	Norwegian (Nynorsk)	(nn)	15.36				

Appendix C: Quantifications of Bot-generated Wikipedia Articles

#	Language	Code	Percentage	#	Language	Code	Percentage	#	Language	Code	Percentage
1	Cebuano	(ceb)	99.61%	111	English	(en)	2.52%	221	French Guianese Creole	(gcr)	0.0%
2	Pali	(pi)	99.37%	112	Simple English	(simple)	2.23%	222	Swazi	(ss)	0.0%
3	Southern Min	(zh-min-nan)	92.78%	113	Mingrelian	(xmf)	2.02%	223	Southern Altai	(alt)	0.0%
4	Bishnupriya Manipuri	(bpy)	91.42%	114	Fijian	(fj)	1.64%	224	Iñupiaq	(ik)	0.0%
5	Waray	(war)	90.28%	115	Lithuanian	(lt)	1.64%	225	Aromanian	(roa-rup)	0.0%
6	Malagasy	(mg)	89.64%	116	Finnish	(fi)	1.58%	226	Venda	(ve)	0.0%
7	Newar	(new)	87.71%	117	Norwegian (Bokmål)	(no)	1.39%	227	Kongo (Kituba)	(kg)	0.0%
8	Tatar	(tt)	86.32%	118	Kurdish (Kurmanji)	(ku)	1.33%	228	Chamorro	(ch)	0.0%
9	Chechen	(ce)	84.16%	119	Low German	(nds)	1.3%	229	Nigerian Pidgin	(pcm)	0.0%
10	Tarantino	(roa-tara)	80.72%	120	Mongolian	(mn)	1.21%	230	Tyap	(kcg)	0.0%
11	South Azerbaijani	(azb)	77.94%	121	Azerbaijani	(az)	1.17%	231	Oromo	(om)	0.0%
12	Silesian	(szl)	76.17%	122	Norwegian (Nynorsk)	(nn)	1.07%	232	Tahitian	(ty)	0.0%
13	Asturian	(ast)	71.83%	123	Interlingua	(ia)	0.99%	233	Gun	(guw)	0.0%
14	Piedmontese	(pms)	71.67%	124	Hebrew	(he)	0.58%	234	Seediq	(trv)	0.0%
15	Swedish	(sv)	68.14%	125	Czech	(cs)	0.54%	235	Kirundi	(rn)	0.0%
16	Welsh	(cy)	66.1%	126	Slovene	(sl)	0.5%	236	Sango	(sg)	0.0%
17	Burmese	(my)	64.1%	127	Minangkabau	(min)	0.49%	237	Frafra	(gur)	0.0%
18	Māori	(mi)	63.55%	128	Lao	(lo)	0.43%	238	Samoan	(sm)	0.0%
19	Kyrgyz	(ky)	62.31%	129	Belarusian (Taraškievica)	(be-tarask)	0.42%	239	Sranan Tongo	(srn)	0.0%
20	Vietnamese	(vi)	58.22%	130	Sindhi	(sd)	0.36%	240	Western Armenian	(hyw)	0.0%
21	Eastern Min	(cdo)	55.77%	131	Estonian	(et)	0.35%	241	Luganda	(lg)	0.0%
22	Serbo-Croatian	(sh)	55.52%	132	Greek	(el)	0.3%	242	Buryat (Russia Buriat)	(bxr)	0.0%
23	Neapolitan	(nap)	54.51%	133	Bavarian	(bar)	0.3%	243	Central Bikol	(bcl)	0.0%
24	Venetian	(vec)	53.56%	134	Ripuarian	(ksh)	0.24%	244	Emilian-Romagnol	(eml)	0.0%
25	Mazanderani	(mzn)	53.39%	135	Xhosa	(xh)	0.23%	245	Shan	(shn)	0.0%
26	Uzbek	(uz)	52.41%	136	Lojban	(jbo)	0.23%	246	Acehnese	(ace)	0.0%
27	Kazakh	(kk)	51.69%	137	Tagalog	(tl)	0.22%	247	Classical Chinese	(zh-classical)	0.0%
28	Lombard	(lmo)	51.66%	138	Scots	(sco)	0.13%	248	Walloon	(wa)	0.0%
29	Banyumasan	(map-bms)	50.57%	139	Swahili	(sw)	0.13%	249	Assamese	(as)	0.0%
30	Basque	(eu)	49.46%	140	Lower Sorbian	(dsb)	0.12%	250	Interlingue	(ie)	0.0%
31	Serbian	(sr)	48.75%	141	Spanish	(es)	0.11%	251	Ligurian	(lij)	0.0%
32	Urdu	(ur)	46.07%	142	Pennsylvania Dutch	(pdc)	0.1%	252	Zulu	(zu)	0.0%
33	Volapük	(vo)	45.22%	143	Old Church Slavonic	(cu)	0.08%	253	Shona	(sn)	0.0%
34	Chuvash	(cv)	44.82%	144	Khmer	(km)	0.06%	254	Banjarese	(bjn)	0.0%
35	Bashkir	(ba)	44.78%	145	German	(de)	0.06%	255	Meitei	(mni)	0.0%
36	Kashmiri	(ks)	44.72%	146	Thai	(th)	0.04%	256	Hakka Chinese	(hak)	0.0%
37	Romanian	(ro)	42.22%	147	Palatine German	(pfl)	0.04%	257	Tumbuka	(tum)	0.0%
38	Occitan	(oc)	42.11%	148	Uyghur	(ug)	0.04%	258	Kapampangan	(pam)	0.0%
39	Dutch	(nl)	40.04%	149	Limburgish	(li)	0.03%	259	Northern Sotho	(nso)	0.0%
40	Arabic	(ar)	39.87%	150	Saterland Frisian	(stq)	0.02%	260	Igbo	(ig)	0.0%
41	Telugu	(te)	34.76%	151	Japanese	(ja)	0.02%	261	Faroese	(fo)	0.0%
42	Slovak	(sk)	34.67%	152	Icelandic	(is)	0.02%	262	Upper Sorbian	(hsb)	0.0%
43	Sundanese	(su)	32.19%	153	Guarani	(gn)	0.02%	263	Sicilian	(scn)	0.0%
44	Afrikaans	(af)	32.15%	154	Scottish Gaelic	(gd)	0.02%	264	Ladin	(lld)	0.0%
45	Tetum	(tet)	31.85%	155	Balinese	(ban)	0.02%	265	Haitian Creole	(ht)	0.0%
46	Persian	(fa)	30.63%	156	Corsican	(co)	0.02%	266	Western Punjabi	(pnb)	0.0%
47	Zeelandic	(zea)	30.5%	157	Turkmen	(tk)	0.01%	267	Punjabi	(pa)	0.0%
48	Tajik	(tg)	25.58%	158	Maithili	(mai)	0.01%	268	Ido	(io)	0.0%
49	Kurdish (Sorani)	(ckb)	25.4%	159	Nahuatl	(nah)	0.01%	269	Kannada	(kn)	0.0%
50	Indonesian	(id)	24.71%	160	North Frisian	(frr)	0.01%	270	Kotava	(avk)	0.0%
51	Armenian	(hy)	23.09%	161	Somali	(so)	0.01%	271	Hausa	(ha)	0.0%
52	Belarusian	(be)	21.86%	162	Latvian	(lv)	0.01%	272	Gorontalo	(gor)	0.0%
53	Ukrainian	(uk)	21.26%	163	Yoruba	(yo)	0.01%	273	Navajo	(nv)	0.0%
54	Gagauz	(gag)	20.3%	164	Malayalam	(ml)	0.0%	274	Sinhala	(si)	0.0%
55	Hill Mari	(mrj)	19.31%	165	Gujarati	(gu)	0.0%	275	Samogitian	(bat-smg)	0.0%
56	Odia	(or)	18.94%	166	Cantonese	(zh-yue)	0.0%	276	Yakut	(sah)	0.0%
57	Fiji Hindi	(hif)	18.9%	167	Breton	(br)	0.0%	277	Buginese	(bug)	0.0%
58	Northern Sámi	(se)	18.51%	168	Zaza	(diq)	0.0%	278	Yiddish	(yi)	0.0%
59	Karachay-Balkar	(krc)	18.21%	169	West Frisian	(fy)	0.0%	279	Ilocano	(ilo)	0.0%
60	Bihari (Bhojpuri)	(bh)	17.19%	170	Egyptian Arabic	(arz)	0.0%	280	Santali	(sat)	0.0%
61	Meadow Mari	(mhr)	16.83%	171	Shilha	(shi)	0.0%	281	West Flemish	(vls)	0.0%
62	Malay	(ms)	16.15%	172	Kabiye	(kbp)	0.0%	282	Sardinian	(sc)	0.0%
63	Bosnian	(bs)	15.53%	173	Paiwan	(pwn)	0.0%	283	Tuvan	(tyv)	0.0%
64	Tamil	(ta)	15.44%	174	Dinka	(din)	0.0%	284	Mirandese	(mwl)	0.0%
65	Sanskrit	(sa)	15.37%	175	Pangasinan	(pag)	0.0%	285	Old English	(ang)	0.0%
66	Hungarian	(hu)	15.15%	176	Nias	(nia)	0.0%	286	Romansh	(rm)	0.0%
67	Ossetian	(os)	14.61%	177	Kikuyu	(ki)	0.0%	287	Judaeo-Spanish	(lad)	0.0%
68	Macedonian	(mk)	13.42%	178	Akan	(ak)	0.0%	288	Konkani (Goan Konkani)	(gom)	0.0%
69	Amharic	(am)	13.09%	179	Kabardian	(kbd)	0.0%	289	Permyak	(koi)	0.0%
70	Quechua (Southern Quechua)	(qu)	12.28%	180	Wolof	(wo)	0.0%	290	Extremaduran	(ext)	0.0%
71	Bulgarian	(bg)	12.28%	181	Nauruan	(na)	0.0%	291	Lingala	(ln)	0.0%
72	Portuguese	(pt)	12.0%	182	N'Ko	(nqo)	0.0%	292	Lingua Franca Nova	(lfn)	0.0%
73	Polish	(pl)	11.89%	183	Pa'O	(blk)	0.0%	293	Doteli	(dty)	0.0%
74	Chinese	(zh)	11.86%	184	Hawaiian	(haw)	0.0%	294	Karakalpak	(kaa)	0.0%
75	Irish	(ga)	11.49%	185	Sakizaya	(szy)	0.0%	295	Papiamento	(pap)	0.0%
76	Moroccan Arabic	(ary)	11.25%	186	Ingush	(inh)	0.0%	296	Chavacano (Zamboanga)	(cbk-zam)	0.0%
77	Esperanto	(eo)	10.45%	187	Awadhi	(awa)	0.0%	297	Maldivian	(dv)	0.0%
78	Albanian	(sq)	10.2%	188	Jamaican Patois	(jam)	0.0%	298	Moksha	(mdf)	0.0%
79	Gan Chinese	(gan)	10.19%	189	Wayuu	(guc)	0.0%	299	Twi	(tw)	0.0%
80	Catalan	(ca)	10.14%	190	Mon	(mnw)	0.0%	300	Livvi-Karelian	(olo)	0.0%
81	Aragonese	(an)	9.7%	191	Bislama	(bi)	0.0%	301	Lezgian	(lez)	0.0%
82	Hindi	(hi)	9.19%	192	Tulu	(tcy)	0.0%	302	Cornish	(kw)	0.0%
83	Erzya	(myv)	8.64%	193	Aramaic (Syriac)	(arc)	0.0%	303	Manx	(gv)	0.0%
84	Crimean Tatar	(crh)	8.51%	194	Atikamekw	(atj)	0.0%	304	Gilaki	(glk)	0.0%
85	Russian	(ru)	7.89%	195	Tongan	(to)	0.0%	305	Veps	(vep)	0.0%
86	Croatian	(hr)	6.69%	196	Zhuang (Standard Zhuang)	(za)	0.0%	306	Kabyle	(kab)	0.0%
87	Kashubian	(csb)	6.64%	197	Kalmyk Oirat	(xal)	0.0%	307	Dagbani	(dag)	0.0%
88	Dutch Low Saxon	(nds-nl)	6.44%	198	Inuktitut	(iu)	0.0%	308	Võro	(fiu-vro)	0.0%
89	Italian	(it)	6.38%	199	Atayal	(tay)	0.0%	309	Lhasa Tibetan	(bo)	0.0%
90	Pashto	(ps)	5.78%	200	Adyghe	(ady)	0.0%	310	Abkhaz	(ab)	0.0%
91	Danish	(da)	5.53%	201	Tigrinya	(ti)	0.0%	311	Saraiki	(skr)	0.0%
92	Korean	(ko)	4.81%	202	Tok Pisin	(tpi)	0.0%	312	Norman	(nrm)	0.0%
93	Avar	(av)	4.62%	203	Sotho	(st)	0.0%	313	Franco-Provençal	(frp)	0.0%
94	Novial	(nov)	4.25%	204	Cheyenne	(chy)	0.0%	314	Kinyarwanda	(rw)	0.0%
95	Galician	(gl)	4.05%	205	Latgalian	(ltg)	0.0%	315	Picard	(pcd)	0.0%
96	Lak	(lbe)	3.96%	206	Tswana	(tn)	0.0%	316	Komi	(kv)	0.0%
97	Latin	(la)	3.92%	207	Chewa	(ny)	0.0%	317	Maltese	(mt)	0.0%
98	Alemannic German	(als)	3.8%	208	Greenlandic	(kl)	0.0%	318	Inari Sámi	(smn)	0.0%
99	Wu Chinese	(wuu)	3.72%	209	Tsonga	(ts)	0.0%	319	Aymara	(ay)	0.0%
100	Javanese	(jv)	3.33%	210	Madurese	(mad)	0.0%	320	Cree	(cr)	0.0%
101	Bengali	(bn)	3.31%	211	Norfuk	(pih)	0.0%				
102	Turkish	(tr)	3.27%	212	Pontic Greek	(pnt)	0.0%				
103	Georgian	(ka)	3.24%	213	Gothic	(got)	0.0%				
104	Friulian	(fur)	3.15%	214	Ewe	(ee)	0.0%				
105	Marathi	(mr)	3.13%	215	Dzongkha	(dz)	0.0%				
106	French	(fr)	3.08%	216	Amis	(ami)	0.0%				
107	Rusyn	(rue)	2.83%	217	Romani (Vlax Romani)	(rmy)	0.0%				
108	Udmurt	(udm)	2.75%	218	Bambara	(bm)	0.0%				
109	Luxembourgish	(lb)	2.65%	219	Fula	(ff)	0.0%				
110	Nepali	(ne)	2.59%	220	Cherokee	(chr)	0.0%				

Appendix D: Quantifications of Bot-made Edits on Wikipedia articles

#	Language	Code	Percentage	#	Language	Code	Percentage	#	Language	Code	Percentage
1	Cebuano	(ceb)	94.05%	111	Nahuatl	(nah)	39.2%	221	Lao	(lo)	20.53%
2	Welsh	(cy)	86.12%	112	Novial	(nov)	38.75%	222	Kirundi	(rn)	20.48%
3	Pali	(pi)	83.88%	113	Arabic	(ar)	38.7%	223	Italian	(it)	20.31%
4	Norman	(nrm)	78.66%	114	Kazakh	(kk)	38.55%	224	Azerbaijani	(az)	19.98%
5	Waray	(war)	77.29%	115	Papiamento	(pap)	38.53%	225	Cantonese	(zh-yue)	19.95%
6	Buginese	(bug)	76.56%	116	Acehnese	(ace)	38.07%	226	Macedonian	(mk)	19.87%
7	Chechen	(ce)	76.52%	117	Meadow Mari	(mhr)	38.03%	227	Northern Sotho	(nso)	19.79%
8	Minangkabau	(min)	73.92%	118	Permyak	(koi)	37.99%	228	Hungarian	(hu)	19.37%
9	Piedmontese	(pms)	73.06%	119	Romansh	(rm)	37.68%	229	Venda	(ve)	19.11%
10	Neapolitan	(nap)	70.82%	120	Latin	(la)	37.37%	230	Luganda	(lg)	18.89%
11	Malagasy	(mg)	70.79%	121	Yiddish	(yi)	37.15%	231	Simple English	(simple)	18.73%
12	Tatar	(tt)	70.36%	122	Armenian	(hy)	37.14%	232	Korean	(ko)	18.51%
13	Asturian	(ast)	69.91%	123	Moroccan Arabic	(ary)	37.06%	233	Gothic	(got)	18.21%
14	Haitian Creole	(ht)	69.37%	124	Lithuanian	(lt)	37.0%	234	Turkish	(tr)	17.89%
15	Southern Min	(zh-min-nan)	68.35%	125	West Flemish	(vls)	36.75%	235	Finnish	(fi)	17.79%
16	Friulian	(fur)	67.19%	126	Crimean Tatar	(crh)	36.68%	236	Telugu	(te)	17.27%
17	Kapampangan	(pam)	65.42%	127	Tagalog	(tl)	36.56%	237	Bavarian	(bar)	17.18%
18	Banyumasan	(map-bms)	62.84%	128	Bosnian	(bs)	36.27%	238	Tswana	(tn)	16.82%
19	Sicilian	(scn)	62.19%	129	Sardinian	(sc)	36.24%	239	Kannada	(kn)	16.81%
20	Kashubian	(csb)	60.49%	130	Marathi	(mr)	36.01%	240	Nepali	(ne)	16.11%
21	Ido	(io)	60.35%	131	Belarusian	(be)	35.94%	241	Classical Chinese	(zh-classical)	15.97%
22	Franco-Provençal	(frp)	60.04%	132	Nauruan	(na)	35.69%	242	Zulu	(zu)	15.84%
23	Lojban	(jbo)	59.38%	133	Breton	(br)	35.58%	243	North Frisian	(frr)	15.82%
24	Māori	(mi)	59.21%	134	Sango	(sg)	34.89%	244	Mingrelian	(xmf)	15.55%
25	Aramaic (Syriac)	(arc)	59.1%	135	Slovak	(sk)	34.82%	245	Greek	(el)	15.28%
26	Tahitian	(ty)	59.04%	136	Gan Chinese	(gan)	34.74%	246	Thai	(th)	14.9%
27	Võro	(fiu-vro)	58.79%	137	Zhuang (Standard Zhuang)	(za)	34.69%	247	Portuguese	(pt)	14.21%
28	Kongo (Kituba)	(kg)	58.63%	138	Ilocano	(ilo)	34.68%	248	Fijian	(fj)	14.01%
29	Samogitian	(bat-smg)	58.63%	139	Ewe	(ee)	34.3%	249	Hebrew	(he)	13.62%
30	Amharic	(am)	57.02%	140	Dzongkha	(dz)	34.19%	250	Shona	(sn)	13.14%
31	Venetian	(vec)	56.97%	141	Albanian	(sq)	34.16%	251	Spanish	(es)	12.78%
32	Kalmyk Oirat	(xal)	56.33%	142	Mirandese	(mwl)	33.41%	252	Abkhaz	(ab)	12.42%
33	Scottish Gaelic	(gd)	56.24%	143	Indonesian	(id)	33.3%	253	Malayalam	(ml)	12.34%
34	Maldivian	(dv)	56.02%	144	Erzya	(myv)	33.22%	254	Russian	(ru)	12.02%
35	Egyptian Arabic	(arz)	55.92%	145	Swedish	(sv)	33.19%	255	Khmer	(km)	11.45%
36	Cornish	(kw)	55.42%	146	Icelandic	(is)	33.18%	256	French	(fr)	11.32%
37	Turkmen	(tk)	55.24%	147	Karakalpak	(kaa)	33.14%	257	Lezgian	(lez)	11.01%
38	Hawaiian	(haw)	55.14%	148	Romanian	(ro)	32.97%	258	Vèps	(vep)	10.99%
39	Chuvash	(cv)	55.11%	149	Fiji Hindi	(hif)	32.79%	259	Chinese	(zh)	10.94%
40	Sranan Tongo	(srn)	54.34%	150	Kabardian	(kbd)	32.4%	260	Buryat (Russia Buriat)	(bxr)	10.53%
41	Serbian	(sr)	53.4%	151	Kinyarwanda	(rw)	32.29%	261	Gilaki	(glk)	10.44%
42	Uzbek	(uz)	53.32%	152	West Frisian	(fy)	32.26%	262	Navajo	(nv)	10.29%
43	Sundanese	(su)	53.25%	153	Kyrgyz	(ky)	32.18%	263	Kashmiri	(ks)	10.02%
44	Lingala	(ln)	52.78%	154	Old Church Slavonic	(cu)	32.07%	264	Livvi-Karelian	(olo)	9.84%
45	Uyghur	(ug)	52.78%	155	Igbo	(ig)	31.77%	265	Greenlandic	(kl)	9.27%
46	Norfuk	(pih)	51.52%	156	Afrikaans	(af)	31.71%	266	Sindhi	(sd)	8.79%
47	Northern Sámi	(se)	51.34%	157	Cherokee	(chr)	31.58%	267	Sakizaya	(szy)	8.69%
48	Quechua (Southern Quechua)	(qu)	50.82%	158	Bulgarian	(bg)	31.58%	268	Sinhala	(si)	8.25%
49	Interlingua	(ia)	50.57%	159	Gujarati	(gu)	31.33%	269	Jamaican Patois	(jam)	8.16%
50	Bishnupriya Manipuri	(bpy)	50.53%	160	Galician	(gl)	31.28%	270	Aromanian	(roa-rup)	7.92%
51	Saterland Frisian	(stq)	50.35%	161	Slovene	(sl)	31.24%	271	Assamese	(as)	7.89%
52	Aragonese	(an)	50.18%	162	Lower Sorbian	(dsb)	31.21%	272	Chewa	(ny)	7.51%
53	Catalan	(ca)	49.9%	163	Burmese	(my)	31.03%	273	Cree	(cr)	7.46%
54	Yoruba	(yo)	49.38%	164	Malay	(ms)	30.85%	274	Tumbuka	(tum)	7.39%
55	South Azerbaijani	(azb)	49.12%	165	Corsican	(co)	30.81%	275	Shilha	(shi)	7.35%
56	Hill Mari	(mrj)	48.73%	166	Iñupiaq	(ik)	30.66%	276	Japanese	(ja)	6.91%
57	Interlingue	(ie)	48.73%	167	Upper Sorbian	(hsb)	30.61%	277	Hausa	(ha)	6.78%
58	Basque	(eu)	48.44%	168	Persian	(fa)	30.43%	278	English	(en)	6.24%
59	Javanese	(jv)	48.43%	169	Bashkir	(ba)	29.87%	279	German	(de)	5.7%
60	Occitan	(oc)	48.43%	170	Croatian	(hr)	29.84%	280	Lingua Franca Nova	(lfn)	5.33%
61	Lhasa Tibetan	(bo)	48.35%	171	Lak	(lbe)	29.69%	281	Amis	(ami)	5.17%
62	Wolof	(wo)	48.21%	172	Moksha	(mdf)	28.63%	282	Ripuarian	(ksh)	4.96%
63	Silesian	(szl)	48.05%	173	Tsonga	(ts)	28.59%	283	Twi	(tw)	4.41%
64	Tarantino	(roa-tara)	47.49%	174	Tongan	(to)	28.49%	284	Bihari (Bhojpuri)	(bh)	4.25%
65	Komi	(kv)	47.42%	175	Belarusian (Taraškievica)	(be-tarask)	28.37%	285	Akan	(ak)	4.06%
66	Hakka Chinese	(hak)	47.18%	176	Vietnamese	(vi)	27.54%	286	Adyghe	(ady)	3.43%
67	Guarani	(gn)	46.7%	177	Sanskrit	(sa)	27.46%	287	Paiwan	(pwn)	2.59%
68	Limburgish	(li)	46.65%	178	Latgalian	(ltg)	27.08%	288	Shan	(shn)	2.36%
69	Pennsylvania Dutch	(pdc)	46.55%	179	Danish	(da)	26.89%	289	Inguish	(inh)	2.24%
70	Western Punjabi	(pnb)	46.37%	180	Norwegian (Bokmål)	(no)	26.86%	290	Gun	(guw)	2.15%
71	Lombard	(lmo)	45.93%	181	Latvian	(lv)	26.84%	291	Tyap	(kcg)	1.82%
72	Extremaduran	(ext)	45.48%	182	Sotho	(st)	26.78%	292	Tuvan	(tyv)	1.67%
73	Ligurian	(lij)	45.21%	183	Banjarese	(bjn)	26.32%	293	Konkani (Goan Konkani)	(gom)	1.63%
74	Aymara	(ay)	44.82%	184	Cheyenne	(chy)	26.1%	294	French Guianese Creole	(gcr)	0.99%
75	Newar	(new)	44.71%	185	Scots	(sco)	25.98%	295	Dinka	(din)	0.97%
76	Tetum	(tet)	44.54%	186	Kurdish (Sorani)	(ckb)	25.75%	296	Southern Altai	(alt)	0.77%
77	Mazanderani	(mzn)	44.42%	187	Judaeo-Spanish	(lad)	25.62%	297	Maithili	(mai)	0.67%
78	Low German	(nds)	44.41%	188	Georgian	(ka)	25.24%	298	Tulu	(tcy)	0.63%
79	Pontic Greek	(pnt)	44.04%	189	Somali	(so)	25.24%	299	Kotava	(avk)	0.52%
80	Central Bikol	(bcl)	43.59%	190	Ukrainian	(uk)	25.09%	300	Mon	(mnw)	0.48%
81	Luxembourgish	(lb)	43.43%	191	Polish	(pl)	25.03%	301	Gorontalo	(gor)	0.46%
82	Tajik	(tg)	43.35%	192	Czech	(cs)	25.01%	302	Inari Sámi	(smn)	0.42%
83	Ossetian	(os)	43.29%	193	Chavacano (Zamboanga)	(cbk-zam)	24.63%	303	Madurese	(mad)	0.35%
84	Faroese	(fo)	43.14%	194	Avar	(av)	24.19%	304	Doteli	(dty)	0.33%
85	Manx	(gv)	42.58%	195	Dutch	(nl)	23.76%	305	Saraiki	(skr)	0.24%
86	Samoan	(sm)	42.52%	196	Chamorro	(ch)	23.73%	306	Kabiye	(kbp)	0.19%
87	Old English	(ang)	42.37%	197	Wu Chinese	(wuu)	23.55%	307	Atikamekw	(atj)	0.19%
88	Romani (Vlax Romani)	(rmy)	42.36%	198	Eastern Min	(cdo)	23.52%	308	Meitei	(mni)	0.17%
89	Kurdish (Kurmanji)	(ku)	42.33%	199	Palatine German	(pfl)	23.4%	309	Awadhi	(awa)	0.17%
90	Karachay-Balkar	(krc)	41.86%	200	Balinese	(ban)	22.89%	310	Seediq	(trv)	0.14%
91	Irish	(ga)	41.84%	201	Kabyle	(kab)	22.86%	311	Ladin	(lld)	0.12%
92	Rusyn	(rue)	41.79%	202	Alemannic German	(als)	22.69%	312	Dagbani	(dag)	0.12%
93	Dutch Low Saxon	(nds-nl)	41.76%	203	Tamil	(ta)	22.65%	313	N'Ko	(nqo)	0.11%
94	Swazi	(ss)	41.7%	204	Volapük	(vo)	22.54%	314	Atayal	(tay)	0.07%
95	Urdu	(ur)	41.61%	205	Western Armenian	(hyw)	22.37%	315	Nias	(nia)	0.01%
96	Gagauz	(gag)	41.57%	206	Maltese	(mt)	22.13%	316	Santali	(sat)	0.01%
97	Swahili	(sw)	41.49%	207	Kikuyu	(ki)	22.07%	317	Pa'O	(blk)	0.0%
98	Serbo-Croatian	(sh)	41.31%	208	Xhosa	(xh)	21.68%	318	Wayuu	(guc)	0.0%
99	Udmurt	(udm)	40.99%	209	Estonian	(et)	21.64%	319	Nigerian Pidgin	(pcm)	0.0%
100	Bambara	(bm)	40.77%	210	Emilian-Romagnol	(eml)	21.48%	320	Frafra	(gur)	0.0%
101	Tok Pisin	(tpi)	40.35%	211	Hindi	(hi)	21.35%				
102	Esperanto	(eo)	40.25%	212	Bengali	(bn)	21.17%				
103	Norwegian (Nynorsk)	(nn)	40.24%	213	Punjabi	(pa)	21.07%				
104	Bislama	(bi)	39.97%	214	Tigrinya	(ti)	21.07%				
105	Zeelandic	(zea)	39.68%	215	Zaza	(diq)	21.01%				
106	Yakut	(sah)	39.68%	216	Oromo	(om)	20.96%				
107	Walloon	(wa)	39.61%	217	Mongolian	(mn)	20.85%				
108	Picard	(pcd)	39.37%	218	Odia	(or)	20.82%				
109	Pangasinan	(pag)	39.31%	219	Fula	(ff)	20.68%				
110	Inuktitut	(iu)	39.24%	220	Pashto	(ps)	20.64%				

Appendix E: Calculations of DEPTH$^+$ Metric of Wikipedia Editions

#	LANGUAGE	CODE	DEPTH+	#	LANGUAGE	CODE	DEPTH+	#	LANGUAGE	CODE	DEPTH+
1	English	(en)	376.77	111	Georgian	(ka)	0.14	221	Fijian	(fj)	0.03
2	German	(de)	40.64	112	Alemannic German	(als)	0.14	222	Bislama	(bi)	0.03
3	French	(fr)	36.89	113	Hausa	(ha)	0.14	223	Latgalian	(ltg)	0.03
4	Italian	(it)	20.45	114	Novial	(nov)	0.14	224	Luganda	(lg)	0.03
5	Japanese	(ja)	12.36	115	Nias	(nia)	0.14	225	Māori	(mi)	0.03
6	Russian	(ru)	12.25	116	Latin	(la)	0.14	226	Dinka	(din)	0.03
7	Polish	(pl)	7.91	117	Ewe	(ee)	0.14	227	Pontic Greek	(pnt)	0.03
8	Chinese	(zh)	7.91	118	Limburgish	(li)	0.13	228	Tumbuka	(tum)	0.03
9	Portuguese	(pt)	7.68	119	West Frisian	(fy)	0.13	229	Udmurt	(udm)	0.03
10	Spanish	(es)	6.9	120	South Azerbaijani	(azb)	0.12	230	Gothic	(got)	0.03
11	Ukrainian	(uk)	6.4	121	Sanskrit	(sa)	0.12	231	Tok Pisin	(tpi)	0.03
12	Swedish	(sv)	6.16	122	Tsonga	(ts)	0.12	232	Lak	(lbe)	0.03
13	Persian	(fa)	5.74	123	Santali	(sat)	0.12	233	Nauruan	(na)	0.03
14	Hebrew	(he)	5.03	124	Paiwan	(pwn)	0.11	234	N'Ko	(nqo)	0.03
15	Vietnamese	(vi)	3.87	125	Norwegian (Nynorsk)	(nn)	0.11	235	Chuvash	(cv)	0.03
16	Pali	(pi)	2.93	126	Lombard	(lmo)	0.11	236	Central Bikol	(bcl)	0.03
17	Indonesian	(id)	2.9	127	Sakizaya	(szy)	0.11	237	Atayal	(tay)	0.03
18	Dutch	(nl)	2.71	128	Aragonese	(an)	0.11	238	Oromo	(om)	0.03
19	Czech	(cs)	2.7	129	Twi	(tw)	0.11	239	Chamorro	(ch)	0.03
20	Hungarian	(hu)	2.57	130	Balinese	(ban)	0.11	240	Xhosa	(xh)	0.03
21	Uzbek	(uz)	2.38	131	Chewa	(ny)	0.1	241	Kyrgyz	(ky)	0.03
22	Finnish	(fi)	2.37	132	Luxembourgish	(lb)	0.1	242	Cornish	(kw)	0.02
23	Korean	(ko)	2.33	133	Dzongkha	(dz)	0.1	243	Lower Sorbian	(dsb)	0.02
24	Thai	(th)	2.11	134	Occitan	(oc)	0.1	244	Mingrelian	(xmf)	0.02
25	Frafra	(gur)	1.96	135	Chechen	(ce)	0.1	245	Kabyle	(kab)	0.02
26	Arabic	(ar)	1.92	136	Madurese	(mad)	0.1	246	Norfuk	(pih)	0.02
27	Estonian	(et)	1.73	137	Lingala	(ln)	0.1	247	Mirandese	(mwl)	0.02
28	Norwegian (Bokmål)	(no)	1.71	138	Malagasy	(mg)	0.09	248	Kabiye	(kbp)	0.02
29	Turkish	(tr)	1.71	139	Sango	(sg)	0.09	249	Guarani	(gn)	0.02
30	Bengali	(bn)	1.5	140	Judaeo-Spanish	(lad)	0.09	250	Veps	(vep)	0.02
31	Greek	(el)	1.48	141	Cantonese	(zh-yue)	0.09	251	Quechua (Southern Quechua)	(qu)	0.02
32	Serbian	(sr)	1.46	142	Sinhala	(si)	0.09	252	Banyumasan	(map-bms)	0.02
33	Nigerian Pidgin	(pcm)	1.35	143	Mongolian	(mn)	0.09	253	Cheyenne	(chy)	0.02
34	Catalan	(ca)	1.16	144	Ingush	(inh)	0.09	254	Meitei	(mni)	0.02
35	Bulgarian	(bg)	1.07	145	Akan	(ak)	0.09	255	Atikamekw	(atj)	0.02
36	Telugu	(te)	1.05	146	French Guianese Creole	(gcr)	0.08	256	Ido	(io)	0.02
37	Romanian	(ro)	1.0	147	Tetum	(tet)	0.08	257	Hawaiian	(haw)	0.02
38	Serbo-Croatian	(sh)	0.99	148	Classical Chinese	(zh-classical)	0.08	258	Kinyarwanda	(rw)	0.02
39	Danish	(da)	0.96	149	Bambara	(bm)	0.08	259	Friulian	(fur)	0.02
40	Moroccan Arabic	(ary)	0.94	150	Wolof	(wo)	0.08	260	Gan Chinese	(gan)	0.02
41	Macedonian	(mk)	0.89	151	Dutch Low Saxon	(nds-nl)	0.08	261	Kalmyk Oirat	(xal)	0.02
42	Ripuarian	(ksh)	0.87	152	Fiji Hindi	(hif)	0.08	262	Gilaki	(glk)	0.02
43	Armenian	(hy)	0.86	153	Belarusian (Taraškievica)	(be-tarask)	0.08	263	Interlingua	(ia)	0.02
44	Azerbaijani	(az)	0.76	154	Sindhi	(sd)	0.07	264	Tahitian	(ty)	0.02
45	Basque	(eu)	0.73	155	Nahuatl	(nah)	0.07	265	Tongan	(to)	0.02
46	Malayalam	(ml)	0.73	156	Newar	(new)	0.07	266	Romani (Vlax Romani)	(rmy)	0.02
47	Greenlandic	(kl)	0.7	157	Tswana	(tn)	0.07	267	Aramaic (Syriac)	(arc)	0.02
48	Tamil	(ta)	0.68	158	Corsican	(co)	0.07	268	Buryat (Russia Buriat)	(bxr)	0.02
49	Cebuano	(ceb)	0.64	159	Palatine German	(pfl)	0.07	269	Emilian-Romagnol	(eml)	0.02
50	Urdu	(ur)	0.58	160	Tajik	(tg)	0.07	270	Kashubian	(csb)	0.02
51	Wayuu	(guc)	0.57	161	Manx	(gv)	0.07	271	Minangkabau	(min)	0.02
52	Latvian	(lv)	0.54	162	West Flemish	(vls)	0.07	272	Tuvan	(tyv)	0.02
53	Slovak	(sk)	0.52	163	Ligurian	(lij)	0.07	273	Livvi-Karelian	(olo)	0.02
54	Slovene	(sl)	0.52	164	Upper Sorbian	(hsb)	0.07	274	Chavacano (Zamboanga)	(cbk-zam)	0.02
55	Tulu	(tcy)	0.5	165	Erzya	(myv)	0.06	275	Kabardian	(kbd)	0.02
56	Inari Sámi	(smn)	0.5	166	Neapolitan	(nap)	0.06	276	Samoan	(sm)	0.02
57	Doteli	(dty)	0.48	167	Sotho	(st)	0.06	277	Pennsylvania Dutch	(pdc)	0.02
58	Kazakh	(kk)	0.47	168	Breton	(br)	0.06	278	Old English	(ang)	0.02
59	Assamese	(as)	0.47	169	Walloon	(wa)	0.06	279	Meadow Mari	(mhr)	0.02
60	Seediq	(trv)	0.46	170	Venetian	(vec)	0.06	280	Gagauz	(gag)	0.02
61	Gun	(guw)	0.45	171	Yakut	(sah)	0.06	281	Pashto	(ps)	0.01
62	Tyap	(kcg)	0.45	172	Old Church Slavonic	(cu)	0.06	282	Komi	(kv)	0.01
63	Kurdish (Sorani)	(ckb)	0.38	173	Irish	(ga)	0.06	283	Sranan Tongo	(srn)	0.01
64	Simple English	(simple)	0.38	174	Northern Sámi	(se)	0.06	284	Sicilian	(scn)	0.01
65	Mon	(mnw)	0.37	175	Venda	(ve)	0.06	285	Shan	(shn)	0.01
66	Icelandic	(is)	0.37	176	Bavarian	(bar)	0.06	286	Cherokee	(chr)	0.01
67	Maltese	(mt)	0.37	177	Javanese	(jv)	0.05	287	Norman	(nrm)	0.01
68	Cree	(cr)	0.37	178	Moksha	(mdf)	0.05	288	Zhuang (Standard Zhuang)	(za)	0.01
69	Amis	(ami)	0.35	179	Ossetian	(os)	0.05	289	Samogitian	(bat-smg)	0.01
70	Bihari (Bhojpuri)	(bh)	0.35	180	Yiddish	(yi)	0.05	290	Picard	(pcd)	0.01
71	Hindi	(hi)	0.34	181	Sardinian	(sc)	0.05	291	Permyak	(koi)	0.01
72	Bashkir	(ba)	0.34	182	Avar	(av)	0.05	292	Low German	(nds)	0.01
73	Southern Min	(zh-min-nan)	0.33	183	Piedmontese	(pms)	0.05	293	Amharic	(am)	0.01
74	Kannada	(kn)	0.31	184	Scottish Gaelic	(gd)	0.05	294	Acehnese	(ace)	0.01
75	Tagalog	(tl)	0.31	185	Burmese	(my)	0.05	295	Navajo	(nv)	0.01
76	Albanian	(sq)	0.3	186	Zeelandic	(zea)	0.05	296	Uyghur	(ug)	0.01
77	Fula	(ff)	0.3	187	Romansh	(rm)	0.05	297	Saraiki	(skr)	0.01
78	Welsh	(cy)	0.3	188	Pangasinan	(pag)	0.05	298	Kapampangan	(pam)	0.01
79	Pa'O	(blk)	0.3	189	Papiamento	(pap)	0.05	299	Zaza	(diq)	0.01
80	Malay	(ms)	0.29	190	Lezgian	(lez)	0.05	300	Zulu	(zu)	0.01
81	Karakalpak	(kaa)	0.29	191	Mazanderani	(mzn)	0.04	301	Crimean Tatar	(crh)	0.01
82	Lithuanian	(lt)	0.28	192	Maldivian	(dv)	0.04	302	Kongo (Kituba)	(kg)	0.01
83	Afrikaans	(af)	0.26	193	North Frisian	(frr)	0.04	303	Lhasa Tibetan	(bo)	0.01
84	Croatian	(hr)	0.26	194	Franco-Provençal	(frp)	0.04	304	Gorontalo	(gor)	0.01
85	Konkani (Goan Konkani)	(gom)	0.26	195	Extremaduran	(ext)	0.04	305	Jamaican Patois	(jam)	0.01
86	Shilha	(shi)	0.25	196	Dagbani	(dag)	0.04	306	Interlingue	(ie)	0.01
87	Odia	(or)	0.25	197	Turkmen	(tk)	0.04	307	Western Punjabi	(pnb)	0.01
88	Khmer	(km)	0.24	198	Igbo	(ig)	0.04	308	Shona	(sn)	0.01
89	Belarusian	(be)	0.24	199	Karachay-Balkar	(krc)	0.04	309	Lingua Franca Nova	(lfn)	0.01
90	Galician	(gl)	0.24	200	Somali	(so)	0.04	310	Kikuyu	(ki)	0.01
91	Bishnupriya Manipuri	(bpy)	0.24	201	Adyghe	(ady)	0.04	311	Aymara	(ay)	0.0
92	Southern Altai	(alt)	0.23	202	Võro	(fiu-vro)	0.04	312	Rusyn	(rue)	0.0
93	Volapük	(vo)	0.23	203	Waray	(war)	0.04	313	Hill Mari	(mrj)	0.0
94	Esperanto	(eo)	0.23	204	Scots	(sco)	0.04	314	Wu Chinese	(wuu)	0.0
95	Asturian	(ast)	0.22	205	Gujarati	(gu)	0.04	315	Egyptian Arabic	(arz)	0.0
96	Western Armenian	(hyw)	0.22	206	Saterland Frisian	(stq)	0.04	316	Haitian Creole	(ht)	0.0
97	Nepali	(ne)	0.22	207	Faroese	(fo)	0.04	317	Hakka Chinese	(hak)	0.0
98	Tarantino	(roa-tara)	0.22	208	Abkhaz	(ab)	0.03	318	Ladin	(lld)	0.0
99	Swahili	(sw)	0.21	209	Kotava	(avk)	0.03	319	Northern Sotho	(nso)	0.0
100	Bosnian	(bs)	0.21	210	Ilocano	(ilo)	0.03	320	Buginese	(bug)	0.0
101	Marathi	(mr)	0.2	211	Sundanese	(su)	0.03				
102	Punjabi	(pa)	0.19	212	Kirundi	(rn)	0.03				
103	Inuktitut	(iu)	0.19	213	Awadhi	(awa)	0.03				
104	Swazi	(ss)	0.18	214	Lojban	(jbo)	0.03				
105	Maithili	(mai)	0.17	215	Banjarese	(bjn)	0.03				
106	Tatar	(tt)	0.16	216	Yoruba	(yo)	0.03				
107	Kashmiri	(ks)	0.16	217	Eastern Min	(cdo)	0.03				
108	Tigrinya	(ti)	0.15	218	Lao	(lo)	0.03				
109	Iñupiaq	(ik)	0.15	219	Kurdish (Kurmanji)	(ku)	0.03				
110	Aromanian	(roa-rup)	0.14	220	Silesian	(szl)	0.03				