Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 326–336
https://doi.org/10.22364/bjmc.2022.10.3.06

Estimating the Amount of Lithuanian Text Indexed by Global Search Engines

Virginijus DADURKEVIČIUS, Andrius UTKA

Vytautas Magnus University, Kaunas LT-44243, Lithuania

virginijus.dadurkevicius@vdu.lt, andrius.utka@vdu.lt

Abstract. The aim of the paper is the estimate of the amount of words in Lithuanian texts indexed by the selected Global Search Engines (GSE), namely Google (by Alphabet Inc.), Bing (by Microsoft Corporation), and Yandex (by ООО «Яндекс», Russia). For this purpose, a special list of 100 rare Lithuanian words (pivot words) with specific characteristics was compiled. Low frequency of pivot words is crucial to consider the count of document matches reported by GSE as an indicator of the word count. Statistical analysis has shown the following amounts of Lithuanian words as of April 2022: 56 billion words by Google, 29 billion words by Bing and 41 billion words by Yandex. Comparative results for neighbouring Belarusian (∼0.31×LT), Estonian (∼1.45×LT), Finnish (∼2.4×LT), Latvian (∼0.95×LT), Polish (∼11×LT), and Russian (∼49×LT) languages have also been assessed.

Keywords: global search engines, Google, Bing, Yandex, Lithuanian language, webometrics, corpus, pivot words

1 Introduction

Global search engines (GSE) became everyday tools that help us to open a window to the vast realm of information on the web. The usage of GSE has become so common that some people even confuse them with the internet itself.

The information on the web is highly multimodal and multilingual consisting of textual, audio, and visual material. However, when we are looking for information with the help of GSE, we can only discover and access that part of information which has been previously indexed by GSE and presented to a user. It must also be said that GSE index only a tiny part of all information that is accessible on the web.

All different modalities for a particular language constitute the digital presence of that language on the web. A larger or smaller digital presence of a language may signify its vitality and importance in the global community. In addition, it can indirectly speak

about the language community's economic development or even the level of adaptability to the modern world. The digital presence may also have a geopolitical importance, as it may have an impact on decision making processes of human societies.

There is no easy way to assess the size of the digital presence by using data of commercial search engines, as the main purpose of commercial tools is to generate profit and not to reflect the objective picture of the digital world. Many researchers warned about potential pitfalls when analysing commercial search engines (see e.g. Kilgarriff (2007); van den Bosch et al. (2016)). It needs to be acknowledged that the current research focuses only on that part of digital presence, which is represented as text, while recognizing that there exists a huge realm of audio-visual information, which cannot be assessed by our methods.

Presently, there are four major commercial global search engines with their own indices: Google, Bing, Yandex, and Baidu[1]. In this paper, however, we will focus on the first three, as our test queries have shown that the Chinese Baidu applies inappropriate segmentation methods of Lithuanian words, which adversely affects the results.

Prior to presenting the research we deem necessary to define the key terms of "word" and "token" in our analysis, as their definitions and treatment vary. In this paper we use the following definitions:

- **token**: the smallest unit in a text corpus. A token normally refers to a word form, a punctuation, a digit, an abbreviation, a product name, and anything else between spaces;
- **word**: any token if not a punctuation. Such a definition of word is also used by GSE, but not by corpus linguists. Corpus linguists tend to define a word as a token, which begins with a letter of the alphabet and consists solely from letters, thus ignoring numbers or any mixed alpha-numeric constructions. For this reason, the size in words of the same corpus may differ depending on the method of calculation.

It should be noted that we do not seek to estimate (in terabytes) the total size of indices operated by GSE nor to determine the total number of indexed URLs/documents, rather we seek to estimate the total amount of the text (in words) indexed by GSE. Also we do not consider the cleanness (deduplication) of corpora or GSE indexed texts as a factor to be accounted for. We can only speculate about the deduplication policy used by a particular GSE or corpus creators, so we seek to estimate the whole amount of text regardless of duplication.

2 Related research

The field of research that focuses on assessing the quantity and quality of information on the web is called webometrics (Björneborn and Ingwersen, 2004). The first research papers on webometrics have been published some thirty years ago (Almind and Ingwersen, 1997) and since then many aspects of the web have been analysed, for instance,

[1] Note that there are a number of other popular search engines, but they do not have their own indices (e.g. Yahoo or DuckDuckGo). Their popularity is based not on different indexed information, but on different regional marketing preferences, functionalities, ranking algorithms or privacy policies. Therefore, they are not included into the present study.

assessing the index sizes of different search engines and different domains (e.g. Bharat and Broder (1998)), link structure of the web (e.g. Hirate et al. (2006)), bias of search results (e.g. Gezici et al. (2021)), evaluation of ranking algorithms (e.g. Canca (2022)) and others.

There are two research papers, namely Kilgarriff (2007) and van den Bosch et al. (2016), that are closely related to the present analysis, as in both cases the sizes of indices of search engines for specific languages were estimated by extrapolating query frequency results from known corpora against GSE search results.

Kilgarriff (2007) presented the analysis for German and Italian languages. Kilgarriff's main idea was to look at texts indexed by Google as a "black box" corpus that can only be studied by queries. The queries are based on a selected list of words, which can be referred to as *pivot words*. Then comparing the results of the same queries made on this "black box" with frequencies from a known reference corpus (RC), it is possible to infer the size of the "black box" corpus based on the average of count ratios for each tested word.

One of the recent attempts to estimate the size of Dutch and English indices was published by van den Bosch et al. (2016). The study presents a longitudinal observation of the size of Google and Bing indices based on frequencies of 28 pivot words. The unique feature of the study is its longitudinal aspect, as authors set up a system, which has been daily monitoring Google and Bing indices since 2006 and it is still ongoing[2].

In many ways, we followed the ideas in these two works, albeit with a very different approach to the selection of pivot words, doing more consistent calculations and neglecting the factor of repetitive documents.

3 Methodology

Our main interest is the estimates for the Lithuanian language. All the efforts, knowledge and sample sizes are adjusted for this purpose. However, for the sake of comparison we have performed a limited scope analysis with less precise estimates (due to smaller test samples) for the neighboring Latvian, Polish, Belarusian, Russian, Estonian, and Finnish languages examining only queries by Google.

For this research we have used the 2nd version of the Corpus of Contemporary Lithuanian Language CCLL2 (Utka et al., 2017) by Vytautas Magnus University (VMU) and various corpora of TenTen family by Sketch Engine (Jakubíček et al., 2013). The details of the corpora are provided in Table 1. As a reference corpus (RC) for Lithuanian we have chosen Sketch Engine ltTenTen14 because of its size, quality and more or less same origin as GSE text. The CCLL2 corpus (5 times smaller than ltTenTen14 and of different build policy) has been involved in this research for selecting pivot words and accomplishing the "proof of concept" when estimating the size of ltTenTen14. For similar reasons we also have chosen the corpora gathered by Sketch Engine for other languages.

[2] `http://www.worldwidewebsize.com/`

Table 1: Details of corpora used in this research (tokens and words are in millions)

Language	Source	Corpus	Tokens	Words	Docs
Belarusian	Sketch Engine	beTenTen16	80	65	166,079
English	BNC Consortium	BNC	112	98	4,054
English	Brown University	Brown	1.18	1.02	500
Estonian	Sketch Engine	etTenTen19	623	524	2,535,829
Finnish	Sketch Engine	fiTenTen14	1,697	1,434	3,610,670
Latvian	Sketch Engine	lvTenTen14	658	543	1,585,626
Lithuanian	VMU	CCLL2	208	166	8,098
Lithuanian	Sketch Engine	ltTenTen14	982	800	2,215,963
Polish	Sketch Engine	plTenTen19	5,216	4,387	13,145,670
Russian	Sketch Engine	ruTenTen11	18,280	14,938	36,946,344

3.1 Criteria for the list of pivot words

The most important part of this research is the selection of a list of pivot words to be used to query GSE and a reference corpus in parallel. Unfortunately, GSE's queries are only reporting the approximate number of documents found and not the word matches, so in order to compare apples to apples, we should also count documents and not words in a reference corpus.

The estimation of size ratio of the two corpora on the docs-to-docs basis (instead of words-to-words) is an indirect measurement. Such a dependency may be highly susceptible to text chunking policy of a particular corpus and as a result it can be nonlinear, e.g. a double increase in docs count may not mean a double increase in corpus size. Let's consider a small example regarding the word *the* in British National Corpus (BNC)[3] and Brown corpus (Francis and Kucera (1979)). In BNC a word-count is 6,054,939 and docs-count is 4,050, while in the Brown Corpus a word-count is 69,971, while a docs-count is 500 (i.e. in both cases every document contains *the*). The actual corpus size ratio is 95.9 (i.e. BNS is 95.5 times bigger than the Brown corpus). Thus, the words-to-words ratio of 86.5 gives us a much more realistic estimate of the actual corpus size ratio compared to the docs-to-docs ratio of 8.1.

Since our main interest is the number of words, and we can only measure the number of documents, we should keep the resemblance between them as close as possible. This factor raises one important requirement for pivot words: they should be *hapax legomena* in all documents where a queried word was found (occurring no more than once per document). That means, we should use infrequent words with low counts through the corpus while ensuring $\{$*word frequency count*$\}\approx\{$*number of documents with the word in*$\}$. Adherence to this principle also avoids some of the subjective peculiarities inherent in low-frequency words: they tend to cluster in certain documents, possibly because of the inclination by some authors to "invent" and use them for very specific purposes.

On the other hand, extremely low frequency counts are statistically prone to greater sampling errors. Therefore, it is essential to select pivot words from within the range of low and high frequency counts. In order to assess this issue, we have evaluated the

[3] `https://www.english-corpora.org/bnc/`

estimation ratio for the two corpora: CCLL2 and ltTenTen14. 5,000 test words were filtered out from CCLL2 having frequency counts ranging from 1 to extremely high 50,000. The sample of the test words has been divided into 30 intervals and individual docs-to-docs ratios as well as means and medians per interval were calculated. The results, presented in Fig. 1 confirm our reasoning about the unsuitability of high frequency words, as well as those below 10. So for pivot words, we decided to choose the words with frequency counts between 10 and 100 in CCLL2. Words with these frequencies in CCLL2 showed the most appropriate prediction of the size ratio of the two corpora with relative error of 12% (5.4 estimate versus 4.8 actual) suggesting that ltTenTen14 versus GSE comparison will also be feasible.

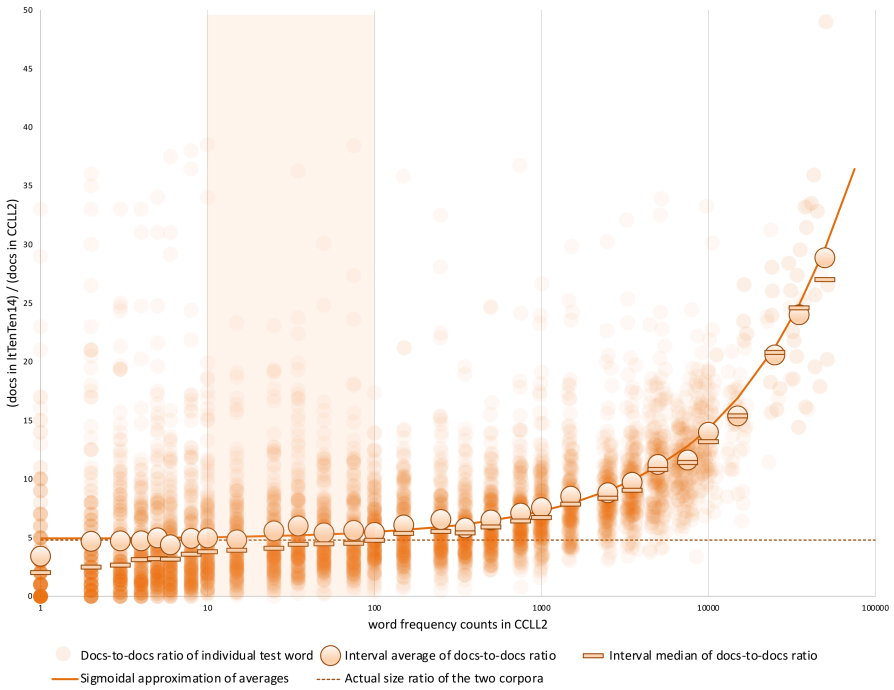

Fig. 1: Docs-to-docs ratio as a function of test word frequency. Corpora under investigation – CCLL2 and ltTenTen14. Shadowed zone (frequency counts between 10 and 100) chosen as a best compromise between statistical errors inherent to low counts and apparently biased ratio estimate at high counts.

Other important requirements to the pivot words are language specific. Pivot words should be able to slice the corpus of particular GSE precisely to subcorpus of documents in specified language (e.g. Lithuanian). Pivot words cannot coincide with regular words of another language. For example the word ”imam” is of no use for the examining Lithuanian-only content because it is regular for English, French, Italian an other lan-

guages too. Moreover, we should avoid words which have the following characteristics:

- shorter than 7 letters;
- international origin;
- foreign loanwords;
- proper names of any kind;
- headword forms;
- having accented characters;
- specific for particular domain or time period;
- normalized (diacritics removed) variants of other words (e.g. Lithuanian *sukuosi* and *šukuosi*);
- common misspellings in target or any other language (as Lithuanian *permetant* and French *permettant*).

3.2 Querying the GSE

In order to ensure comparative results between GSE and languages, we have adhered to specific criteria. All GSE and languages should be tested at the same time and in the same way. All queries have been performed by using "exact word form search" functionality by means of double quotes surrounding the word to be searched. Query process has been performed manually, in order to avoid anti-robot functionalities behind the scenes that are used by some GSE. No special linguistic, date, type or any other option than can narrow the search scope should be set. When analyzing search results, "Documents found" number should be recorded regardless of any further circumstances (reports of possible duplicates, copyright issues etc.).

3.3 Querying reference corpora

All the reference corpora (RC) from Sketch Engine can by queried using special built-in functionality of Sketch Engine's "Wordlist" advanced features allowing batch processing of all the list of test words. The query returns word-counts and document-counts for each test word. Date of the query is not important because the content of Sketch Engine corpora does not change.

3.4 Statistical calculations

Following Kilgarriff (2007) we have used docs-counts ratio as an estimate of size ratio of the two corpora (or ordinary corpus versus GSE as a corpus). As it has been explained earlier we have been used thoroughly selected test words to avoid biased estimations. So the i-th estimate of size in words of the indexed Lithuanian text by particular GSE:

$$N_i = N_{RC}\frac{y_i}{x_i} \tag{1}$$

where:

x_i is document-count for i-th test word in Lithuanian RC,
y_i is document-count for i-th test word in particular GSE,
N_{RC} is size in words of the RC.

Having the set of estimates $\{N_i\}$ we can calculate mean, median, and outliers.

4 Results

Main results of this research will be published on CLARIN-LT repository[4]. Here we present the most important part of the results and some samples of test words for all languages. It should be noted that the list of 100 Lithuanian pivot words was prepared with great care and in accordance with the criteria laid out in Section 3.1. Due to the lack of deep and specific knowledge of the other languages, corresponding lists of test words are substantially shorter, may have inconsistencies with the principles listed in Section 3.1 and the results for these language may be less precise.

4.1 Results for Lithuanian

The GSE measurements for Lithuanian were performed twice with an interval of approximately six months – for the first time on September 27, 2021 and for the second time on April 11, 2022. The counts for sample test words are presented in Table 2. Statistical analysis of all the test words is presented in Figure 2 and Table 3.

Table 2: Sample of the Lithuanian pivot words and their counts

	Test word	RC words	RC docs	Google 04/22	Google 09/21	Bing 04/22	Bing 09/21	Yandex 04/22	Yandex 10/21
1	kraujuodamas	21	21	2250	2690	1300	754	1000	1000
2	kapstydamasis	23	22	938	1130	3610	99	2000	1000
3	aplipusiomis	24	22	1250	1290	1140	736	571	416
4	giedotojais	23	23	886	981	1880	754	410	4000
5	titnaginius	25	25	2210	2280	1590	1470	2000	1000
...									
48	pamestinukai	102	96	6030	6110	7060	2810	4000	4000
49	gimdytojais	101	98	3500	3930	4040	2230	2000	4000
50	gamtosaugininko	116	105	2570	3430	3210	2880	2000	2000
51	apaugusiame	110	107	11200	12900	3350	4600	6000	4000
52	margaspalvius	110	109	3540	5980	1900	2580	4000	3000
...									
96	vieninteliais	397	386	22700	20900	4440	3620	18000	4000
97	nevertinamas	414	405	40900	33100	4770	4650	20000	3000
98	susirinkau	521	512	28900	16600	5390	3840	24000	3000
99	nepritaikytas	608	592	58400	40400	4320	4240	21000	3000
100	nuotaikingos	799	784	57700	64300	5410	5450	18000	3000

Not surprisingly, the biggest number of 56 billion Lithuanian words was indexed by Google, followed by Russian Yandex (41bn) and Microsoft Bing (29bn). It should be noted however, that Yandex's scores raise reasonable doubts, as a significant portion of the reported "number of documents found" is heavily rounded and appears to be suspiciously repetitive.

[4] `https://clarin.vdu.lt/xmlui/`

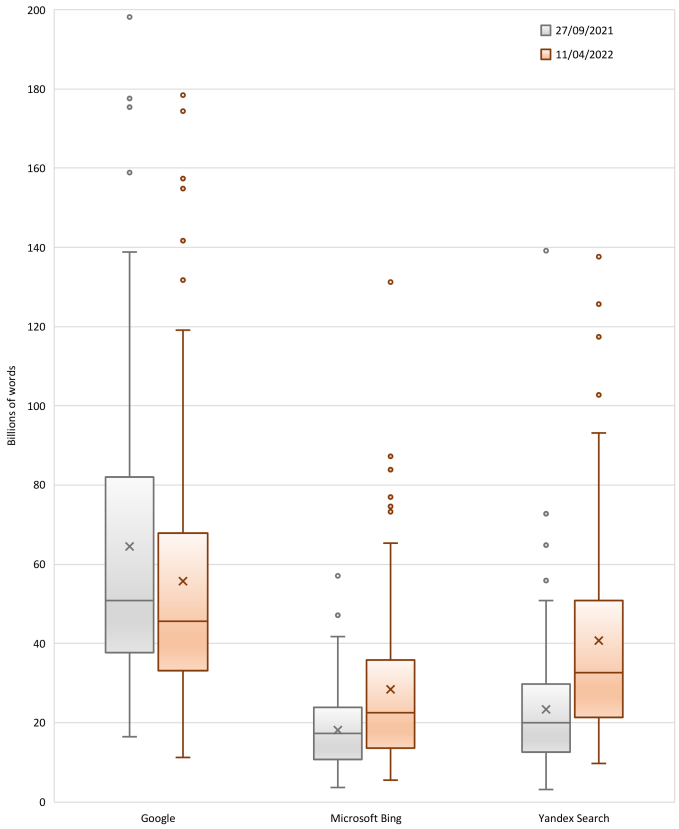

Fig. 2: Estimated amount of Lithuanian text indexed by Global Search Engines as of September 2021 and April 2022

Table 3: Estimated amount of Lithuanian text in billions of words indexed by Global Search Engines as of September 2021 and April 2022

	Google, bn	Microsoft Bing, bn	Yandex Search, bn
Sep-2021	64.5	18.3	26.0
Apr-2022	55.7	28.7	41.0
Change	-13.6%	56.8%	57.6%

Another interesting observation relates the indexing by Google: during the past six months, the volume of indexed Lithuanian text there has decreased. This could be explained by the recent Google's policy of "cleaning up" its indices from junk, duplicates or intentionally misleading content. This policy by Google has also been mentioned by Indig (2020). Rather large fluctuations of Google's index size were also reported by

van den Bosch et al. (2016) in their longitudinal analysis. On the contrary our counts show that the size of Lithuanian text by Microsoft Bing and Yandex have increased by more that 50% during the past six months, which raises some interesting questions. Such a large increase could be caused by many different reasons, for instance by the technical advancement of web crawling algorithms, by the proliferation of AI-generated texts, by fluctuations similar to those observed by Google, or simply by the tendency to increase indexes. Further investigation is needed to answer these questions.

4.2 Results for other languages

A comparative assessment of the amount of indexed text in other neighboring languages was performed on April 2022 only with Google and only with a limited set of pivot words. Sample pivot words and corresponding docs-counts are presented in Table 4 through Table 9.

Interlingual assessment results are presented in Table 10, including calculated results *per capita*.

Table 4: Sample **Belarusian** pivot words

	Pivot word	RC	Google
1	марнеюць	9	3430
2	улагоджвалі	6	261
...			
8	сховішчамі	10	6050
9	апраўдваючыся	28	5900
10	неспадзяваныя	20	3180

Table 5: Sample **Estonian** pivot words

	Pivot word	RC	Google
1	saamahimuliste	8	220
2	vaatlusnimekirja	9	4230
...			
14	kirikukellade	128	17900
15	kurjategijateks	128	9580
16	kujunemisloost	128	22100

Table 6: Sample **Finnish** pivot words

	Pivot word	RC	Google
1	jahtaamiselta	8	513
2	laadullisuudesta	10	524
...			
36	junamatkan	1945	383000
37	tiedostavat	1984	158000
38	viitisentoista	1994	131000

Table 7: Sample **Latvian** pivot words

	Pivot word	RC	Google
1	stiepjamu	9	309
2	ielenktajiem	18	1250
...			
5	aizaugumu	27	2750
6	kopsavilkumos	44	4480
7	aizvainojumam	109	8720

Table 8: Sample **Polish** pivot words

	Pivot word	RC	Google
1	utopionych	432	48700
2	przytulonych	217	36100
...			
8	skrawkiem	825	126000
9	niedostosowanemu	8	853
10	najkorzystniejszemu	13	1530

Table 9: Sample **Russian** pivot words

	Pivot word	RC	Google
1	увядаем	63	12900
2	отплачивая	83	11000
...			
18	прохожими	5030	542000
19	безграмотные	5273	1370000
20	оправдываясь	6280	300000

5 Conclusions

Given the current significance of GSE in everyday decision making, it is important to track the changes in the volume of indexed texts, as this may signal important political, technological or social processes within the corresponding societies. On the other hand, the changes could be just a technological or marketing decision of GSE. In any case, the amount of accessible information influences our daily lives and shows the extent of digital presence of the language that we speak.

Is it possible to imagine the size of 56 billion words of Lithuanian text indexed by Google? Is this a really large number? This could be compared to books. As one book usually contains about 100 thousand words, Google's "assets" are comparable to 0.5m books, which roughly corresponds to the amount of unique books published (keeping the current production rate) in Lithuania in about 100 years! Even though the exact amount of unique texts is difficult to estimate both on the web and in the libraries due to duplicated material, we think that the calculated size is interesting for data scientists, as well as linguists in establishing the order of magnitude of accessible Lithuanian text on the web.

Such an amount of Lithuanian text operated by GSE shows the language's vitality and allows us to expect rather good results of the search queries. Of course, they may be affected by deliberate text filtering, e.g. for a justified reason of personal data protection, harmful or false information. Besides, the access to the presented information is also influenced by hit ranking algorithms.

Among our future plans is setting up a monitoring system that is similar to the one designed by van den Bosch et al. (2016), first of all, for monitoring the change of the volume size of Lithuanian indexed text and, perhaps, eventually for monitoring other languages. It is also important to continue working on testing and validation of the lists of pivot words for other languages with linguists of these languages, in order to ensure comparable results.

Table 10: Comparative interlingual estimate of text in billions of words indexed by Google as of April 2022

Language	Words	Native speakers*	Words per capita
	bn	m	k
Belarusian	17	5.1	3
Estonian	81	1.1	73
Finnish	134	5.8	23
Latvian	53	1.75	30
Lithuanian	56	3	19
Polish	629	40	16
Russian	2,716	154	18

* The numbers of native speakers were taken from Wikipedia.

6 Acknowledgements

We would like to thank Dr. Kristina Vaisvalavičienė for her valuable advice concerning the list of Latvian pivot words and Kalok Man on his consultation regarding Chinese Baidu.

References

Almind, T., Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics', *Journal of Documentation* **53**, 404–426.

Bharat, K., Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines, *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, Elsevier Science Publishers B. V., NLD, p. 379–388.

Björneborn, L., Ingwersen, P. (2004). Toward a basic framework for webometrics, *J. Assoc. Inf. Sci. Technol.* **55**(14), 1216–1227.
https://doi.org/10.1002/asi.20077

Canca, C. (2022). Did you find it on the internet? ethical complexities of search engine rankings, *in* Werthner, H., Prem, E., Lee, E. A., Ghezzi, C. (eds), *Perspectives on Digital Humanism*, Springer, pp. 135–144.

Francis, W. N., Kucera, H. (1979). *The Brown Corpus: A Standard Corpus of Present-Day Edited American English*, Department of Linguistics, Brown University.

Gezici, G., Lipani, A., Saygin, Y., Yilmaz, E. (2021). Evaluation metrics for measuring bias in search engine results, *Information Retrieval Journal* **24**, 85–113.

Hirate, Y., Kato, S., Yamana, H. (2006). Web structure in 2005, *in* Aiello, W., Broder, A. Z., Janssen, J. C. M., Milios, E. E. (eds), *Algorithms and Models for the Web-Graph, Fourth International Workshop, WAW 2006, Banff, Canada, November 30 - December 1, 2006. Revised Papers*, Vol. 4936 of *Lecture Notes in Computer Science*, Springer, pp. 36–46.
http://dx.doi.org/10.1007/978-3-540-78808-9_4

Indig, K. (2020). Google's index is smaller than we think - and might not grow at all.
https://www.kevin-indig.com/googles-index-is-smaller-than-we-think-and-might-not-grow-at-all/

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. (2013). The TenTen corpus family, *7th International Corpus Linguistics Conference CL 2013*, Lancaster, pp. 125–127.
https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf

Kilgarriff, A. (2007). Googleology is bad science, *Computational Linguistics* **33**, 147–151.

Utka, A., Rimkutė, E., Kovalevskaitė, J., Bielinskienė, A., Petkevičius, M., Petrauskaitė, R., Mikelionienė, J. (2017). Corpus of the contemporary lithuanian language. CLARIN-LT digital library in the Republic of Lithuania.
http://hdl.handle.net/20.500.11821/16

van den Bosch, A., Bogers, T., de Kunder, M. (2016). Estimating search engine index size variability: a 9-year longitudinal study, *Scientometrics* **107**, 839 – 856.

Received August 11, 2022 , accepted August 11, 2022