https://doi.org/10.36575/2353-2912/1(7)2019.009 *Forum Filologiczne Ateneum* 1(7)2019
pp. 9–25

Readability and lexical sophistication of colon cancer websites – a corpus-assisted assessment of online educational materials for patients

Czytelność i trudność leksykalna internetowych stron dotyczących raka jelita grubego – korpusowa analiza materiałów edukacyjnych dla pacjentów

Anna BĄCZKOWSKA[1]
Uniwersytet Mikołaja Kopernika w Toruniu

Abstract
The aim of this paper is to check whether the information for colon cancer patients available on top websites devoted to this disease is comprehensible for the readers or whether, due to high saturation with special medical terms, it is beyond the recommended readability level of an average internet user. Two main criteria of analysis were involved in the study: readability and lexical sophistication. The methods used in the study include 8 readability tests (ARI, Colemen-Liau, New Dale-Chall, Flesch-Kincid, Fry, Gunning Fog, Raygor Estimate, and SMOG), TAALES software and Lexical Complexity Analyser used to examine syntactic and lexical parameters of texts, and a corpus-assisted web-based tool used for lexical sophistication called VocabProfile. The study has shown that none of the 30 websites under scrutiny meets the demand of the recommended readability level, and that higher lexical sophistication involves a lower readability level.

Keywords: colon cancer, health literacy, readability, plain language, lexical sophistication, patient online materials, vocabulary profile, corpus-assisted study, medical discourse

[1] https://orcid.org/0000-0002-0147-2718.

Streszczenie
Celem artykułu jest sprawdzenie czy informacje na temat raka jelita grubego dostępne na najczęściej odwiedzanych stronach internetowych są zrozumiałe dla czytelników czy też, z uwagi na duże nasycenie terminami specjalistycznymi, informacje te są poza poziomem czytelności rekomendowanym dla przeciętnego użytkownika internetu. W analizie zastosowano dwa kryteria opisu: czytelność i trudność leksykalną. Metody badawcze polegały na wykorzystaniu 8 testów na czytelność (ARI, Colemen-Liau, New Dale-Chall, Flesch-Kincid, Fry, GunningFog, RaygorEstimate, SMOG) oraz kilku programów komputerowych, w tym TAALES i LexicalComplexityAnalyser do analizy syntaktycznych i leksykalnych cech tekstu oraz program VocabProfile, dostępny w internecie program oparty na analizie korpusowej pozwalający przeanalizować trudność leksykalną. Badanie wykazało, że wysoka wartość trudności leksykalnej skorelowana jest z poziomem czytelności.

Słowa kluczowe: rak jelita grubego, *healthliteracy*, czytelność, prosty język, trudność leksykalna, materiały online dla pacjentów, profil słownictwa, analiza korpusowa, dyskurs medyczny

1. Introduction

Colon cancer is the fourth most common type of cancer diagnosed in Great Britain [1], the third in the US [2], and the third in the world [3]. After lung and breast cancer, it is one of the three cancers responsible for both top incidence and mortality worldwide [3]. Over the past few years, it can be observed that patients increasingly seek information about healthcare and treatment disseminated on the internet (Shultz and Young 2017), where they can easily address their concerns, as the Internet is a cheap and broadly accessible source of information. One in three adult Americans tries to find information about their medical conditions online (Fox and Duggan 2013). They can learn and educate themselves as well as exchange opinions and experiences, and, still, remain anonymous. Forty-one percent of "online diagnosers" claim that their diagnosis was confirmed by physicians (Fox and Duggan 2013).

However, health information consumers often lack relevant education in health-related issues and thus they have problems with understanding the information found online. The average reading level in the US is estimated to be between the 7th and 8th grade, which is insufficient to comprehend health care information (Nielson-Bohlman, Panzer and Kindig 2004). However, there are also suggestions to prepare health educational materials for patients at the level of the 5th to 6th grade level "thus ensuring readability by the majority of adults", or even at the 3rd to 5th grade level to secure "limited literacy of a "high percentage of patients" (Weiss 2007: 38). It is claimed that 36%

of Americans, nearly 50% of Europeans, and about 60% of Canadians and Australians have inadequate health literacy (Wittlink and Oosterhaven 2018) and thus their understanding of patient materials published on the web is unsatisfactory.

The understanding of patient educational materials depends on two factors: the readers (their general education level and health-related skills and knowledge), i.e. on health literacy, and the quality of the information available to the general public, i.e. on text readability. Health literacy is a term that describes an individual's competencies in reading, understanding and using information on health-related topics, so that patients can participate in decision-making regarding their treatment (Weiss 2007). Decision-making shared between a clinician and a patient, responsive to the needs and preferences of the patient, is a strong tendency in today's healthcare, which witnesses a turn from the authoritarian doctor-centred to empathy-based patient-centred clinical communication. Patient-centeredness optimises the relation between the physician and the patient, and empowers patients by allowing them to share the responsibility for their own health (Silverman, Draper and Kurtz 2013). The patient is no longer seen as a passive receiver of arbitrary decisions made by medical professionals, but as an active participant in decision-making and an open and better educated citizen. Thus, a patient is provided with easy-to-understand information regarding treatment risks and benefits so that a final choice of treatment can be fully understood by a patient and can actually be taken by him/herself with full awareness and in line with his/her personal preferences and values. There is growing evidence that shared decision making does not only enhance doctor-patient communication but also brings better outcomes in patients already receiving curative treatment (Charles 1977).

The objective measurement of the ease with which a mainstream internet user can grasp the meaning of a text is known as text readability (Nielsen-Bohlman, Panzer and Kindig 2004). Readability depends on text characteristics, reader characteristics and the method of readability measurement (Leroy et al. 2013). In order to enhance the readability of health-related materials for patients, the information should be written in an accessible and understandable way. The US Department of Health and Human Services (USDHHS) recommends that the level of patient materials on the internet be at the 6th grade. This decision stems from other studies that show that one fourth of adult Americans are at the literacy level of the sixth grade or lower (Ley and Florio 1996).

The low level of readability and health education of patients and the low readability of health-related websites, together with the tendency to allow patients to participate in decision-making regarding their treatment, puts a huge burden on quality control and assessment of health-related websites. To safeguard against inaccurate, misleading or advertisement-biased information provision, which is a health jeopardising practise, special information quality labels have been introduced, such as Information Standard (IS) and HonCode certification in the UK, and HonCode certification, Standards of Excellence Certification Program (SE) and URAC (U) in the USA. Their aim is to provide standards of accountability and quality information on health on the web designed for patients and health professionals. They are optional yet their presence helps eliminate cases of evident non-compliance with health information standards and thus they confirm the quality of information and permit greater reliability of and confidence in the web content.

2. Corpus data

Three largest search engines (Google, Bing, and Yahoo) were used to retrieve relevant data and to create a colon-cancer corpus. The dominant search engine in the world is unquestionably Google (81% of market share), followed by Bing (below 10%) and Yahoo (3.9%) [4]. They are thus responsible for nearly 95% of all internet searches [5]. The search on colorectal carcinoma was performed using the terms "colon cancer" and "bowel cancer". The second term was used to secure a reliable data retrieval methodology. The search retrieved websites both about colon and bowel cancer as well as colorectal and rectal cancer. Those devoted to colorectal cancer were included in the analysis, while those describing solely rectal cancer were excluded.

The searches were performed from one location in a Poland-based IP address using the Chrome web browser in March 2019. Prior research has shown that 90% of Internet users do not go beyond the first page of search returns, and most consumers looking for health-related information on the internet tend to view the top five returns from search engines (Eysenchach and Kohler 2002). Therefore, ten top ranked websites were considered for further analysis in the present study. In cases where the websites overlapped with already retrieved ones by other search engines, up to the top 20 websites (with overall 200 texts) were selected for further analysis.

The whole corpus of data is ca. 260,000 words in size. The average text length per website is 8,500 words. Website texts were stored in folders, which contained documents with web pages gleaned from each website. Affiliation of websites registered with other than org, gov or edu domains was confirmed by the WHOis.net database. Altogether, the top 30 unique websites were retrieved and addressed for readability. Up to 10 documents were downloaded manually from each of the websites from three search engines, totalling 200 texts (webpages). Each text was stored in a separate plain text file in the form exactly as it appeared on the website; however, each document was edited to remove metadata and non-medical information, and these included references to other pages and clinics or doctors, repetitions of same text on one web page, figures, tables, images, in-text advertisements, references, tables of contents, disclaimers, acknowledgements, web navigation, addresses and telephone numbers to clinics, and readers' comments. Duplicated websites and websites in languages other than English were excluded. The coding system used in the study contains information about the version of English (British, American, Canadian or Australian), assessed on the basis of web domain and contact address of the institution, the quality of information (certified or non-certified information content), and the type of language used (plain English or non-verified in terms of language simplicity). This coding allows further, more detailed analysis based on correlation studies.

Of the 30 websites used in this study, 13 (43.33%) are certified websites and 17 (56.66%) are non-certified. Twenty (66.66%) websites are not-for-profit and ten (33.33%) are commercial ones. Ten (33.33%) of them are run by British institutions, 14 (46.66%) by American, two (6.66%) by Canadian and three (10%) by Australian. The structure of the data is tabulated in table 1.

Table 1. Website addresses with details concerning the number of words and web pages, the country of the institution that runs each website, their commercial or not-for-profit status and their certification status

Website address	Code	No of words	No of web pages	Country/ language version	Commercial (C) and not-for-profit (N) websites	Certified (Cr)
www.mayoclinic.org	MAC	6362	6	B	N	Cr (H, IS, P)
www.fascrs.org	Fas	8001	6	A	N	NC
www.cancerresearchuk.org	CR	8440	10	B	N	Cr (IS, P)
www.bupa.co.uk	Bu	11312	10	B	C	Cr (H, IS)
www.nhs.uk	NHS	6503	6	B	N	
www.bowercancerresearch.org	BCR	2021	6	B	N	
www.nice.org.uk	Ni	10235	8	B	N	
www.wcrf.org	WC	1011	1	A	N	

www.cancer.org	ACS	9321	10	A	N	
www.preventcancer.org	PCF	4741	10	A	N	Cr (SE)
www.bowelcanceruk.org.uk	BC	5292	10	B	N	Cr (H, IS)
www.bowercanceraustralia.org	BCA	8086	9	B	N	
www.betterhealth.vic.gov.au	BH	2943	2	Au	N	
www.macmillan.org.uk	Mc	1728	10	B	N	Cr (H, IS)
www.cancer.org.au	C	5979	10	Au	N	
www.royalmarsden.nhs.uk	Ro	1932	2	B	N	Q
www.nhsinform.scot	NH	4084	5	B	N	
www.medicinenet.com	MDC	13151	3	A	C	Cr (H)
www.medicalnewstoday.com	MNT	8660	10	A	C	Cr (H)
www.verywellheath.com	VWH	11060	5	A	C	Cr (H)
www.cancer.ca	CCa	6399	10	Ca	N	Cr (H)
www.cancerfightstrategies.com	CFS	77314	1	Ca	C	
www.webmd.com	WM	6772	10	A	C	Cr (H, U)
www.healthline.com	HL	8446	8	A	C	Cr (H)
www.nmihi.com	NM	2646	1	S	C	
www.medlineplus.gov	Med	14103	5	A	N	
www.cdc.gov	CD	1339	6	A	N	
www.ehealthiq.com	EH	2217	6	A	C	
www.emedicinehealth.com	EM	3523	7	A	C	Cr (H)
www.mskcc.org	MS	3523	7	A	N	

3. Research methods

3.1. Readability tests

The research method used to assess readability involved measuring readability levels assessed by eight readability formulae (Oleander Studio software, v. 2015) and calculating the correlation among a number of parameters of all the texts using statistical tests. All statistical calculations were conducted using Statistica software (v. 10), and significance was set to $p<.05$.

The following readability tests were used: Automated Readability Index (ARI), Colemen-Liau, New Dale-Chall, Flesch-Kincaid, Fry, Gunning Fog, Raygor Estimate, and SMOG. The outcome of each of these tests is the grade level in the US schooling system. ARI (Smith and Senter 1967) is a test that calculates the comprehensibility of text by indicating the grade level based on character count and sentence length. Similarly, the Colemen-Liau (Colemen and Liau 1975) test relies on the same data as ARI, yet these two tests use different formulae. New Dale-Chall (1995) builds on sentence length and the number of unfamiliar words. Flesch-Kincaid (Kincaid et al. 1975) test predicates on sentence length and syllable count. Fry (1977) is a graphical test that calculates the average number of sentences and syllables per 100 words.

Gunning Fog (1968) and Raygor Estimate (1977) are also graphical tests. The former scores the number of sentences and complex words, whilst the latter counts the average number of sentences and long words per 100 words. Finally, SMOG (Simple Measure of Gobbledygook) hinges on complex word density (McLoughlin 1969).

3.2. Lexical sophistication

Lexical sophistication (LS) was a measure originally developed to assess language proficiency in second language learners (Laufer and Nation 1995). Comparisons of words used in students' 300-word-long essays and a list of 1,000 most common words (derived from the General Service List; West 1953) allowed the authors to find correlations between students' vocabulary knowledge and their language proficiency. LS was further used by other scholars to design language courses in terms of learners' vocabulary needs or was further developed by other authors (e.g. Morris and Cobb 2004, Brezina and Gablasova 2013), even though LS had to confront some criticism based *inter alia* on the apparently insufficient length of the texts on the basis of which LS was calculated and ignorance of multi-word units (Meara and Bell 2001, Meara 2005). Both GSL, and later the new-GSL (Brezina and Gablasova 2013), as well as the Academic Word List (AWL; Coxhead 1998) and the New AWL (Coxhead 2000) were used by some scholars in subsequent years as an improvement on the existing lists (e.g. Higginbotham and Reid 2018).

The computer software currently widely used, and also employed in this study, which calculates LS is the VocabProfiler (VP). It is an open-source system used to assess lexical sophistication in English and French made available through the Lextutor website (Cobb 2012). The VP relies on the two classical wordlists, i.e. the GSL and the AWL, their two new versions, the NGSL and the NAWL, as well as two language corpora in the case of English: the British National Corpus (BNC) and the Contemporary Corpus of American English (COCA, Davis 2008).

4. Results

4.1. Readability

Readability assessed by eight tests is shown in Figure 1. The highest readability scores are observed for the following websites: CR, BC, Mc and WM. The lowest readability scores were obtained for MDC, WC and Ni.

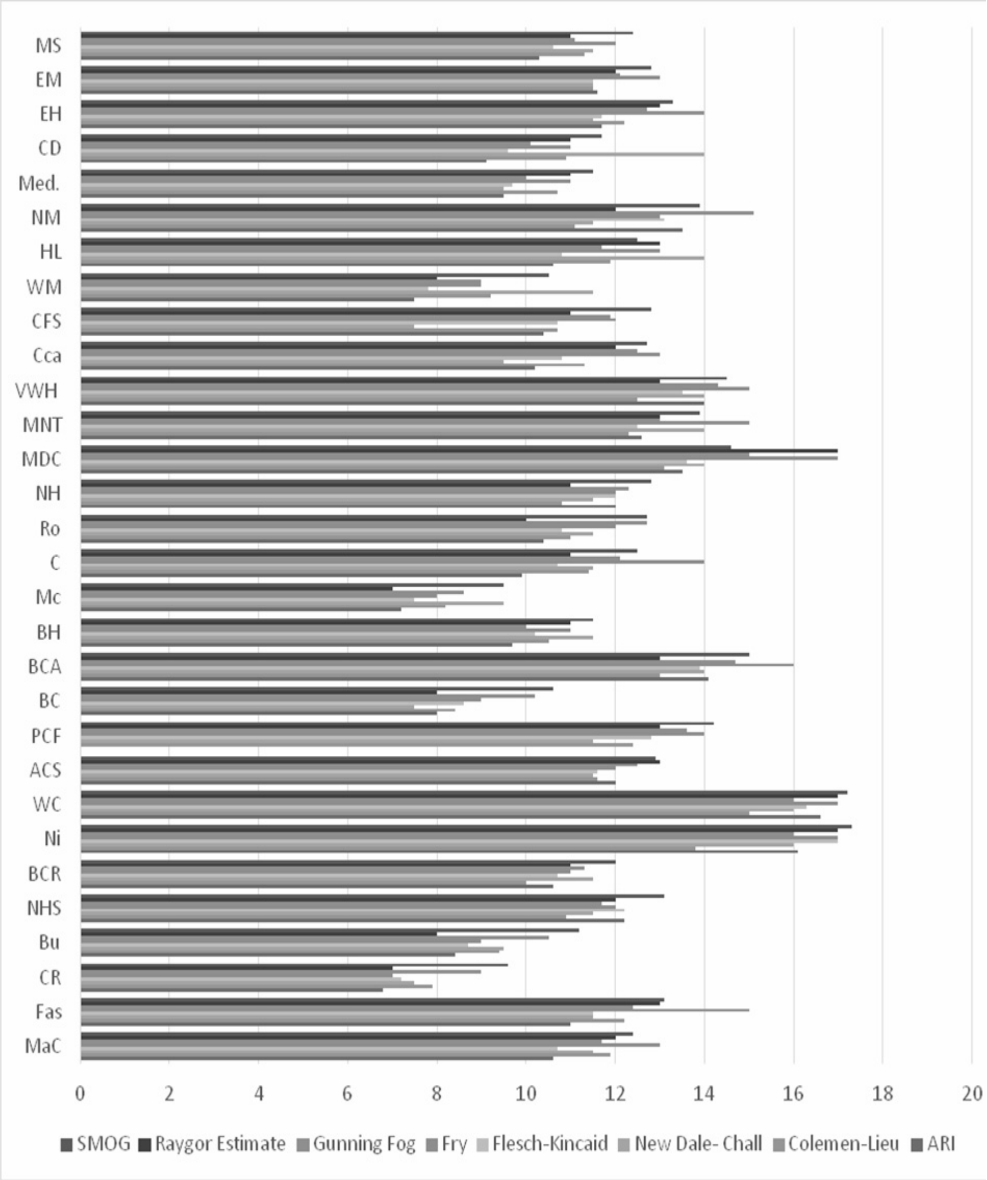

Fig. 1. Results of readability tests

Cronbach's alpha value was computed to check the reliability of the eight readability indices. The result of α=.983 shows excellent internal consistency of the test results (Bland and Altman 1997). Thus, the mean of all readability scores (grade level) was calculated for each of the 20 websites to get the average of readability grade (ARG), which equals 11.8 (95% CI 11.1 to 12.5). Figure 2 shows ARG with the trend line at the 12th grade and the USDDH recommended readability level for health-related website content at the 6th grade.

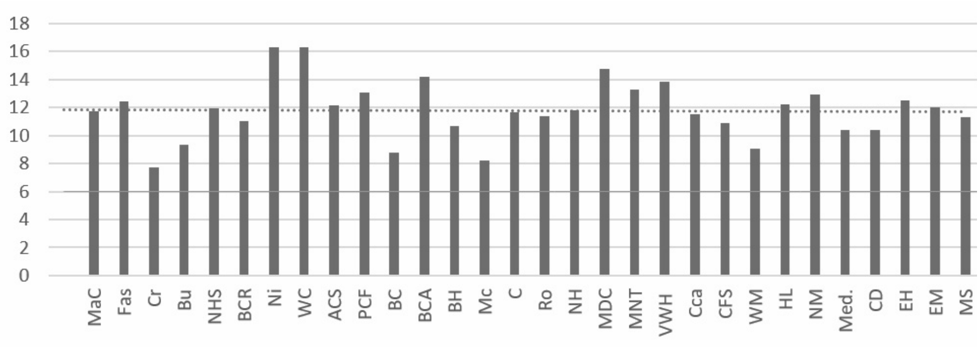

Fig. 2. Average Reading Grade

The Fry graph shows that none of the 30 websites meets the recommended health materials literacy level equal to or lower than the 6th grade in the US (or the age of 11 in the UK). In fact, most websites are written at the level of the 12th grade or higher. The Fry graph for all articles (200 documents) shows that only one text is at a health literacy level equal to the recommended 6th grade (from cancerresearchuk.org). The cancerresearchuk.org website (Cr) cares about the language level of content as it is certified by the Plain English Campaign and thus the overall readability value for the website is 7.6, which is the lowest of all websites.

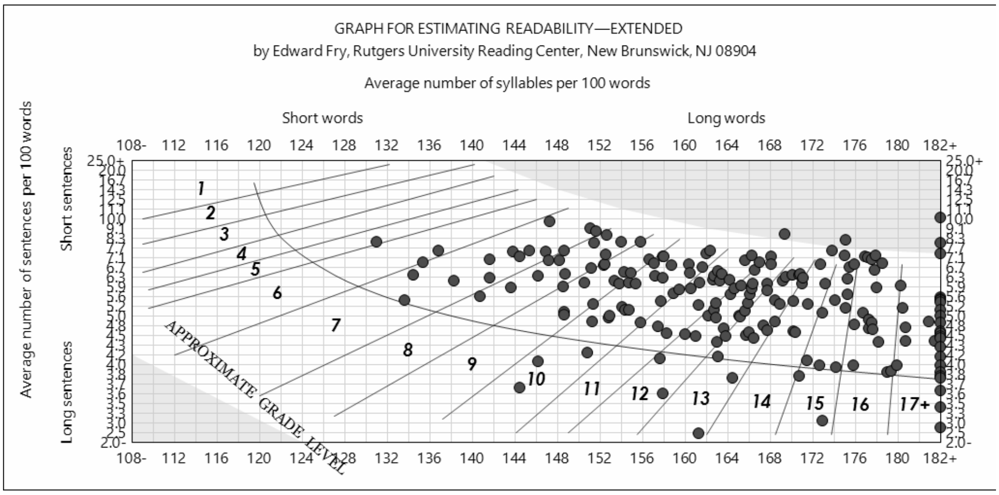

Fig. 3. Fry readability graph

None of the websites meets the demand recommended by the USDHHS of patient-oriented text being at the level of the 6th grade. There are thus no texts classified by the USDHHS as being of "easy" readability. Only four websites (13.33%) correspond to what USDHHS labels as of "average" readability level, which leaves 26 websites (86.66%) illustrating texts of "difficult" readability.

The biggest differences are between: (1) UK certified not-for-profit websites and (a) UK non-certified commercial websites, as well as between (b) American certified and non-certified not-for-profit websites; (2) UK non-certified commercial websites and American certified commercial websites. The mean and standard deviation for all types of websites across variables (country, certification and domain-based affiliation) are shown in Figure 4 (no calculations were performed for groups with only one sample).

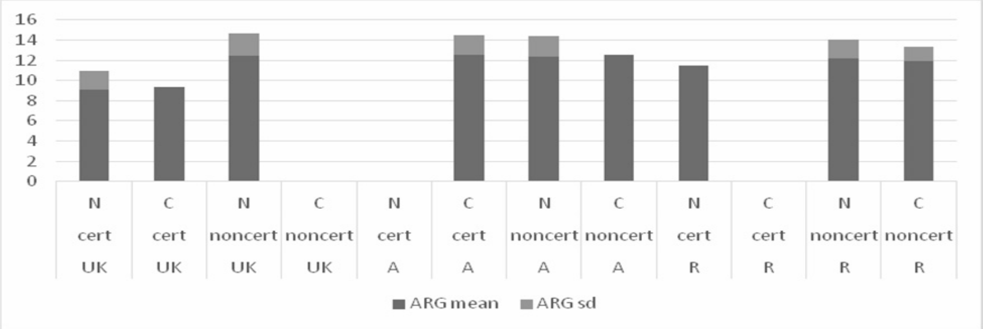

Fig. 4. Means and standard deviations of all types of websites across three variables (country, certification and domain-based affiliation) calculated for average reading grade (ARG)

Significant differences were observed between the certified versus non-certified groups (t= –3.6546, p=.0013, d=.382), and between websites published in different countries (H=11.265, p=.00358, ε^2 =0.4897), yet not between commercial versus not-for-profit groups (U=22.5, p=.34212, z= –.94519, r_{rb}=.021).

The certified websites labelled with Hon and IS have a statistically significant difference in the mean (Hon: 11.85; IS: 8.63; t=2.499, p=.0157, g=1.66). The IS labelled websites are all published in the UK; two of them are also certified with Plain English Campaign (PE). The low level of the mean for IS websites may thus be attributed not only to the fact that they are IS-certified but also to the factor of the domain-based country of origin and the PE certificate. The lowest readability mean of all has the British PE-labelled website.

4.2. Lexical sophistication

In this study, lexical sophistication of content words (LS_c) only was examined as it was assumed that this parameter would generate more precise information about LS than if LS of all words were taken into account (i.e. including function words, such as *a*, *the*, etc.) (Linnarud 1986, Hyltenstam 1988). LS_c is calculated by dividing the number of rare content words (also known as lexical words) by the number of all content words in a text, so the formula is N_{slex}/N_{lex} (Lu 2012). Lexical sophistication of verbs (LS_v) was also examined (Laufer 1994, Lu 2012). Analogically, lexical sophistication of content verbs can be computed (T_{sv}/T_v).

The values for LS, calculated with the use of Lexical Complexity Analyser (LCA, Lu 2012), demonstrate that a higher readability score entails a higher lexical sophistication score (LS_c r=.4677, p=.0091), and that content verbs (VS_c r=0.5447, p=.0018) used in higher readability score texts contain more sophisticated (more rare and advanced) verbs. Contrary to expectations, it is not only medical terms (i.e. nouns) that are rare words, but also verbs seem to be too advanced for a patient-oriented website text.

In keeping with the above results, texts of greater readability (lower readability scores) should contain words of higher frequency, i.e. ones of greater familiarity. The analysis of high-frequency content words was conducted using four databases: BNC (data for written texts), COCA (data for academic texts, for magazines and the news), the Brown Corpus (B) and London-Lund Corpus (LL). A free TAALES software (Kyle and Crossley 2015) was used to conduct the calculations. The results based on the four databases generated similar results with no statistical significance of differences between the corpora (Fig. 5), thus supporting the significance of the LS values for the colon cancer websites regardless of the reference corpus used and ultimately confirming the LS values.

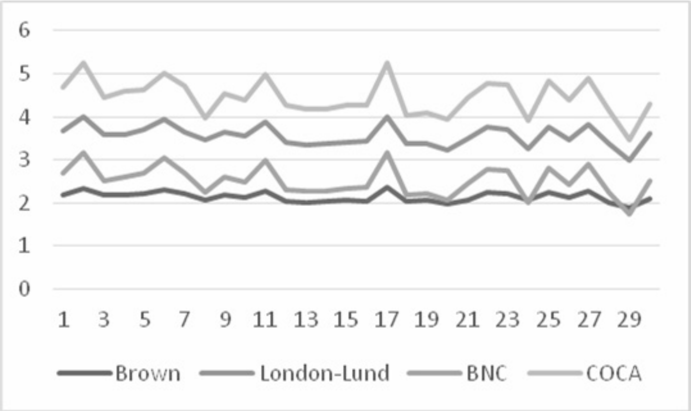

Fig. 5. Lexical sophistication: logarithm of lexical frequency of content words in four corpora

A strong correlation between readability scores and content words frequency was noticed in all four corpora (B: r= –.641, p=.000135; LL: r= –.8093, p=.00001; BNC: r= –.7006, p=.000017; COCA$_{ac}$: r= –.0451, p=0.012372, COCA$_{mag}$: r= –.7625, p= .00001, COCA$_{news}$: r= –.6851, p=.00003). Two conclusions arise from the above calculations. Websites with a high readability score have a higher number of frequent content words. In other words, websites with easier language have more words that are frequently used. Secondly, the language of colon cancer websites has features of written and formal discourse and is similar to the language of magazines and the news.

To determine the frequency bands of the words appearing on the websites a free online software VocabProfile (www.lextutor.ca) was used, which is a standard software dedicated to calculations of lexical frequency profiles of texts. Two databases (the BNC-COCA and the BNC) were employed to identify the percentage of words outside the 2,000 most frequent words occurring in the two corpora. In Lextutor, words within 2k typically constitute on average over 80% of all words used in texts (Fig. 6); hence the words above 2k were treated as “rare” words in the present study.

It should be mentioned in passing that this hands-on observation of 2k being the borderline between the frequent and the infrequent vocabulary is in fact bolstered by earlier research conducted by Brezina and Gablasova (2013). They noticed that there existed “a stable vocabulary core”, which added up to 2,122 items (which made 70.7%), and that emerged out of four language corpora they used (BNC, LOB, BE06 and EnTenTen12). The core vocabulary list was enriched by these scholars by a shorter list of most frequently used words occurring in the BE06 and EnTenTen12 corpora that represent "the current language use" to finally achieve a new-GWL of ca. 2,500 items, which constitutes about 80% of the source corpora.

Returning to the study at hand, the result shows that the percentage of words used on the 30 websites that represent the frequency band beyond 2k (infrequent words) differs significantly depending on which database (BNC–COCA or BNC) is used as a reference corpus (U=287, p=.0164, *z*-score=2.4028). In consequence, correlations between the readability score and the percentage of infrequent words were conducted for both databases separately. The results demonstrate that there is a statistically significant correlation between readability level grade and the frequency bands of words calculated for both BNC–COCA (r_s=.4736, p=.0082) and BNC database (r_s=.3904, p=.0329). From this it transpires that the correlation is observable; however, a greater correlation was achieved by resorting to the BNC–COCA data. Therefore, the readability score is correlated with the percentage of words representing less frequently used words: the higher the score in readability tests (i.e. the less understandable the texts), the more infrequent words the texts contain (i.e. words above the 2,000 most frequent words band). The VocabProfile test supports the results obtained by the TAALES software and reveals more precise information.

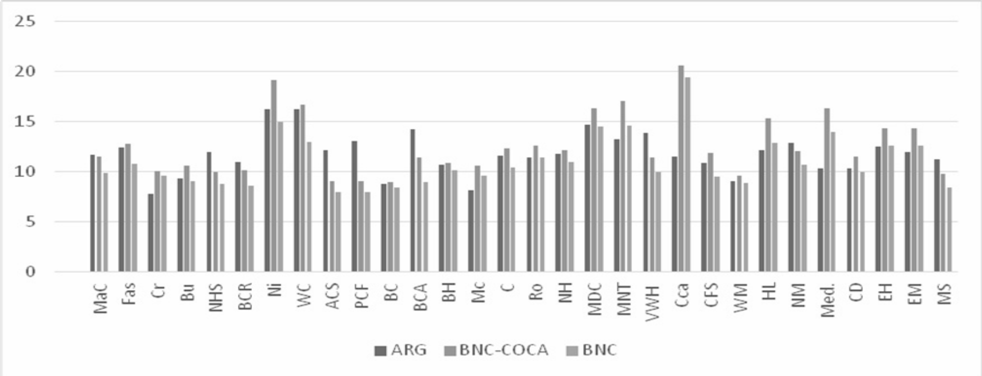

Fig. 6. Lexical frequency profile of colon cancer websites: percentage of words above 2k band calculated for two databases (BNC–COCA, BNC) against average readability grade (ARG)

5. Discussion

The study on readability is consistent with prior research on medical online text readability regarding *inter alia* colorectal cancer (Grewal and Alagaratnam 2013), Graves' disease and thyroid-associated ophthalmopathy (Edmunds, Denniston, Boelaert, Frankly and Durrani 2014), skin cancer (Dobbs, Neal, Hutchings and Whitaker 2017), and pancreatic cancer (Storino et al. 2016). All these studies revealed that the readability level far exceeded the 6th grade in schooling recommended by the USDDH. Only about 13% of texts in our study are at the "average" level of readability, and as many as nearly 87% of texts are classified by the USDHHS as "difficult". This result should be disturbing considering the fact that almost 72% of adult Internet users seek health-related information on the Internet (Fox and Duggan, 2013), and that health literacy may negatively affect clinical treatment and all-cause mortality (Baker et al. 2007), at least in the elderly.

The lowest score in readability tests obtained websites published in the UK with IS or PE certificates, relatively low score was also noticed in the case of British not-for-profit websites relative to American websites and those published in other countries. No significant differences were noticed between commercial and not-for-profit websites, which supports results from some earlier studies. Contrary to some earlier study (Edmunds et al. 2014), there is a noticeable difference in readability scores between UK and other countries, with a tendency to have higher readability level for UK domains.

As for lexical sophistication, the study supports the intuitive judgment that the words typical for higher frequency bands (i.e. above 2,000 and higher) are more likely to occur in texts of low comprehensibility. For a text to be understandable, therefore, preferably words from the first 2,000 words should be used, which constitute around 80% of most frequent words across several language corpora (see Fig. 6). Alternatively, words from the band 2,500 words could be used to achieve a similar percentage, as suggested by Brezina and Gablasova (2013), who employed the following corpora: BNC, COCA, EnTenTen12 and BE06, which illustrate more current language.

There are several limitations to the study. Firstly, the analysis of readability shows how comprehensible a text is based on the complexity of words (number of syllables, signs, words in a sentence) and their frequency. However, even a text written with an easy language does not guarantee that it will be understood by a reader, as the success in text comprehensibility depends on a number of other factors related to the reader, e.g. the reader's mood, attitude to a text and subject matter, the general knowledge about the world, the intellectual capacity, etc. This study did not focus on the readers; instead, it was a text-oriented investigation. Filling the gap between examining text features and text comprehensibility by readers would require further research. Secondly, all of the readability tests involved in this study have some limitations. Thirdly, only 2 subsequent pages with website addresses retrieved by search engines (and only three of them) were taken into account.

This limitation, however, should not influence the results, considering the outcomes of other studies claiming that Internet users rarely go beyond the first two pages of results returned by search engines, or even that they focus only on the first five URL addresses (Eysenchach and Kohler 2002). Finally, the graphic content (diagrams, photographs, etc.) presented on the websites was not considered in assessment of the readability level of the internet health-related content, even though it certainly supported the understanding of the discussed health issues.

6. Conclusions

The websites devoted to colon cancer far overestimate the reading abilities of an average internet user. The vocabulary used in the texts published online for patient information should be simplified. One way of making them more readable for an average patient of low health literacy could be using words of higher frequency, preferably from the frequency lists containing about 2,000 words, or up to 2,500 words that constitute the core vocabulary of the current English language.

References

Baker, D. W., Wolf, M. S., Feinglass, J., Thompson, J. A., Gazmararian, J. A. & Huang, J. (2007). Health literacy and mortality among elderly persons. *Archives of Internal Medicine*, 167(14), 1503-1509.

Bland, J. M. & Altman, D. G. (1997). Statistics notes: Cronbach's Alpha, *Br Med J*, 314, 572. DOI: 10.1136/bmj.314.7080.572

Brezina, V. & Gablasova, D. (2013). Is there a Core Vocabulary? Introducing the *New General Service List*. *Applied Linguistics*, 36(1), 1–22.

Chall, J.S. & Dale, E. (1995). *Manual for the New Dale-Chall Readability Formula*. Cambridge.

Cobb, T. (2012). *Compleat lexical tutor*, available at http://www.lextutor.ca/. Accessed 10 March 2019.

Coleman, M. & Liau, T.L. (1975). A computer readability formula designed for machine scoring, *Journal of Applied Psychology,* 60.2, 283-284.

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly,* 34/2, 213–38.

Coxhead, A. (1998). *An academic word list*. Wellington.

Davis, M. (2010). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available from english-corpora.org/coca.

Dobbs, T., Neal, G., Hutchings, H.A., & Whitaker I.S. (2017). The Readability of Online Patient Resources for Skin Cancer Treatment. *Oncol Ther*, 149–160.

Edmunds, M.E., Denniston A.K., Boelaert K, Franklyn J.A. & Durrani O.M. (2014). Patient Information in Graves' Disease and Thyroid-Associated Ophthalmopathy: Readability Assessment of Online Resources. *Thyroid*, 24(1), 67-72. DOI: 10.1089/thy.2013.0252

Eysenchach, G. & Kohler C. (2002). How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews, *BMJ*, 324(7337), 573–577. DOI: 10.1136/bmj.324.7337.573

Fox, S, & Duggan, M. (2013). Pew Research Center. Available from: https://www.pewinternet.org/2013/01/15/health-online-2013/. DOA: January 15 2013.

Grewal, P. & Alagaratnam, S. (2013). The quality and readability of colorectal cancer information on the internet. *International Journal of Surgery,* 11(5), 410–413. DOI: 10.1016/j.ijsu.2013.03.006

Gunning, R. (1968). *The Technique of Clear Writing*. New York.

Higginbotham, G. & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes,* 37, 127-140.

Hyltenstam, H. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development*, 9, 67–84.

Kincaid, J. P., Lt. Fishburne, R.P. Jr., Rogers, R.J. & Chissom, B.S. (1975). United States. Naval Air Station Memphis, Research Branch. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Millington: Naval Technical Training Command.

Kyle, K. & Crossley S.A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly,* 49(4), 757–786.

Laufer, B. (1994). The Lexical Profile of Second Language Writing: Does it change over time. *RELC Journal*, 25(21), 21–33.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. "Applied Linguistics, 16(3), 307–322.

Leroy, G., Kauchak, D. & Mouradi, O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics,* 82(8), 717–730.

Linnarud, M. (1986). *Lexis in composition: a performance analysis of Swedish learners' written English*. Lund.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96, 190–208.

Morris, L. & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, 32(1), 75–87.

Ley, P. & Florio, T. (1996). The use of readability formulas in health care. *Psychology, Health and Medicine*, 1, 7–28.

McLaughlin, H. (1969). SMOG grading - a new readability formula. *Journal of Reading*, 22, 639-646.

Nielsen-Bohlman, L., Panzer, A.M. & Kindig, D.A. (2004). *Health Literacy: A Prescription to End Confusion*. Washington, DC.

Raygor, A.L. (1977). *The Raygor Readability Estimate: A Quick and Easy Way to Determine Difficulty*. Clemson.

Shultz, H.A. & Young, K.M. (2017). *Health Care USA: Understanding Its Organization and Delivery*. 7th ed. Sudbury, MA.

Smith, E. A. & Senter, R. J. (1967). United States. Air Force. Automated Readability Index, Dayton.

Silverman, J. S., Draper, J. & Kurtz, S. (2013). *Skills for Communicating with Patients*, London.

Storino, A., Castillo-Angeles, M., Watkins, A. A., Vargas, C., Mancias, J. D., Bullock, A. & Moser, J. K. (2016). Assessing the accuracy and readability of online health information for patients with pancreatic cancer. *JAMA Surgery*, 151(9), 831-837.

USDHHS (2010) - U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. National Action Plan to Improve Health Literacy, Washington, DC.

Weiss, B. D. (2007). *Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians.* AMA Foundation.

West, M. (1953). *A general service list of English words*. London.

Wittlink, H. & Oosterhaven, J. (2018). Patient education and health literacy. *Musculosceletan Science and Practice*, 38, 120-127.

Internet sources

[1] https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero

[2] https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html.

[3] https://www.who.int/cancer/PRGlobocanFinal.pdf

[4] Reliablesoft.net. Digital marketing agency. [Online]. [cited 2019. Available from: https://www.reliablesoft.net/top-10-search-engines-in-the-world/

[5] EbizMBA. Top 15 most popular search engines. [Online]. [cited 2019 Apr 12]. Available from: http://www.ebizmba.com/articles/search-engines