Scientometrics (2020) 122:1769–1790
https://doi.org/10.1007/s11192-020-03361-4

Exploring linguistic characteristics of highly browsed and downloaded academic articles

Bikun Chen[1] · **Dannan Deng**[1] · **Zhouyan Zhong**[1] · **Chengzhi Zhang**[1]

Received: 1 November 2019 / Published online: 30 January 2020

Abstract

Views and downloads of academic articles have become important supplementary indicators of scholarly impact. It is assumed that linguistic characteristics have an influence on article views and downloads to some extent. To understand the relationship between linguistic characteristics and article views and downloads, this study selected 63,002 full-text articles published from 2014 to 2015 in the PLoS (Public Library of Science) journals (PLoS Biology, PLoS Computational Biology, PLoS Genetics, PLoS Medicine, PLoS Neglected Tropical Diseases, PLoS One and PLoS Pathogens), and introduced seven indicators (title length, abstract length, full text length, sentence length, lexical diversity, lexical density and lexical sophistication) to measure linguistic characteristics of articles, grouped into Top 20% viewed and downloaded (proxy of highly browsed and downloaded articles), total and Bottom 20% viewed and downloaded categories. The results suggested that most linguistic characteristics played little role in article views and downloads in our data sets in general, but some linguistic characteristics (e.g. title length and average sentence length) in specific PLoS journal and platform (PLoS platform or PubMed Central platform) played certain role in article views and downloads. Also, journal differences and platform differences regarding linguistic characteristics of highly viewed and downloaded articles were existed.

Keywords Linguistic characteristics · Linguistic complexity · Usage metrics · PLoS · PubMed Central

Mathematical Subject Classification 91C99

JEL Classification C80

✉ Bikun Chen
chenbikun@njust.edu.cn

1 Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China

Introduction

Usage metrics of academic articles have become increasingly popular in scientometrics. Article views and downloads are frequently used as important supplementary indicators to measure scholarly impact, identify latest research trends of disciplines and explore user usage patterns. Meanwhile, researchers conduct correlation analysis to investigate relationship among article views, downloads, citations, co-author counts, funding data and so on, trying to probe why articles are viewed or downloaded. However, most researches above are limited to bibliographic data.

Usage metrics are required to be studied in a broader vision. The increasing availability of full text from scientific articles in machine readable electronic formats is an opportunity to greatly impact scientometrics. In-text citations and entity metrics are typical examples of full-text analysis in scientometrics (Ding et al. 2013). Similarly, it is potential and valuable to introduce full-text analysis to usage metrics.

In this study, it is assumed that linguistic characteristics have an influence on article views and downloads to some extent. To understand the relationship between linguistic characteristics and article views and downloads, linguistic characteristics (including title length, abstract length, full text length, sentence length, lexical diversity, lexical density and lexical sophistication) jointly with usage metrics are investigated.

Literature review

Usage metrics

Usage metrics in scientometrics mainly focused on the following topics. Firstly, user behavior patterns, such as scientists' working timetable (Wang et al. 2012, 2013a), user preferences (Chen 2018; Chen et al. 2018; Davis and Solla 2003; Davis 2006; Wang et al. 2016a, b) and user temporal usage patterns (Chen et al. 2017; Khan and Younas 2017). Secondly, obsolescence of articles from diachronic or synchronic perspective. For example, Moed (2005) and Moed and Halevi (2016) studied diachronic and synchronic obsolescence of usage data from perspective of journals and countries. Gorraiz et al. (2014) done the similar research from perspective of disciplines.

Thirdly, identifying latest research trends of disciplines (Bollen et al. 2002; Wang et al. 2013b). Fourthly, indicators to evaluate performance of journals, authors, groups and countries (Chi and Glänzel 2018; De Sordi et al. 2016; Wan et al. 2010) or supplementary metrics jointly with altmetrics measures (Bollen et al. 2005; Kurtz and Henneken 2016). Finally, correlation between specific usage types, including downloads and citations (Kurtz and Bollen 2010; O'Leary 2008; Schloegl et al. 2014; Subotic and Mukherjee 2014; Zhao 2017), usage data among different platforms (Chen 2018; Chen et al. 2017), usage data and co-author counts (Chi and Glänzel 2017), or funding data (Zhao et al. 2018).

Full-text analysis in scientometrics

Full text contains additional information that has not been available in bibliographic data. At a minimum this includes reference position, proximity of cited references within the text, multiple references at the same reference point, multiple mentions of references (so-called op. cit.), section information, and words indicating how an author feels about a reference (i.e., citation contexts or sentiments). Full text also contains a relatively high level of detail about motivation, methods, data, instruments, results, and conclusions that authors typically report when documenting and submitting their work for publication (Boyack et al. 2013).

Full-text analysis in scientometrics mainly focused on the following topics. Firstly, in-text citations, such as reference position (e.g. Hu et al. 2013; Boyack et al. 2018), proximity of cited references (e.g. Gipp and Beel 2009; Liu and Chen 2012; Boyack et al. 2013; Kim et al. 2016), citation contexts or sentiments (e.g. Small 2011; Liu and Chen 2013; Ding et al. 2014; Lu et al. 2017), and citation motivation or behavior (e.g. Brooks 1986; Cano 1989; Bonzi and Snyder 1991; Case and Higgins 2000; Zhang et al. 2018). Secondly, entity metrics, such as scientific concepts (e.g. Ding et al. 2013; Mckeown et al. 2016), datasets (e.g. Belter 2014), softwares (Pan et al. 2015, 2016, 2018, 2019) and algorithms (e.g. Wang and Zhang 2018). Finally, linguistic complexity of scientific writing styles and scientific impacts (e.g. Lu et al. 2019a, b) and characteristics of a highly cited article (e.g. Elgendi 2019).

Most researches of usage metrics focus on numerical analysis. Also, a few researches analyze textual contents jointly with usage metrics, but they are limited to traditional bibliometric methods (e.g. keyword frequency and ratio, bibliographic coupling, co-word analysis and correlation analysis) and bibliographic data. Full-text analysis in scientometrics mainly focus on in-text citations and entity metrics. And it is increasingly expanding to more hot topics in scientometrics, such as scientific writing styles and scientific impacts.

Research questions

The application of full-text analysis to understand the relationship between linguistic characteristics and article views and downloads has not been thoroughly investigated. To address this research gap, full-text analysis is used to explore linguistic characteristics of highly browsed and downloaded papers. In this study, we are interested in following research questions in the context of seven journals published by PLoS:

1. Are there any relationships between linguistic characteristics and highly browsed academic articles?
2. Are there any relationships between linguistic characteristics and highly downloaded academic articles?
3. Are there journal and platform differences of linguistic characteristics in highly browsed and downloaded academic articles?

Methodology

Data

The data in this study consist of 63,002 full-text articles published from 2014 to 2015 in the PLoS journal family, a set of peer-reviewed journals covering various disciplines. In PLoS, usage counts along with other metadata are collected between November 1st and November 7th, 2018. The PLoS journals are also indexed by PMC (PubMed Central) and

Table 1 Datasets

Journal	# Of publications
PLoS Biology (BIO)	288
PLoS Computational Biology (CBI)	1115
PLoS Genetics (GEN)	1514
PLoS Medicine (MED)	171
PLoS Neglected Tropical Diseases (NTD)	1372
PLoS One (ONE)	57,361
PLoS Pathogens (PAT)	1181

Web of Science (WoS). In PMC and WoS, usage counts along with other metadata are also crawled between November 1st and November 7th, 2018. PLoS usage counts, PMC usage counts and WoS usage counts of each article along with other metadata are aggregated by DOI or article title.

The published time span of the data set ranges from January 2014 to December 2015. Because the usage counts, especially citations, accumulate to a steady level in the first 2 or 3 years after publication (Lippi and Favaloro 2013). The editorials and letters are excluded, only research articles pre-labeled by PLoS are kept, the final datasets are shown in Table 1.

Usage counts of PLoS

PLoS offers Article-Level Metrics (ALMs) to each journal article. PLoS ALMs draw from the sources below. Viewed: PLoS Journals (HTML, PDF, XML), PubMed Central (HTML, PDF); Saved: CiteULike, Mendeley; Cited: CrossRef, Datacite, Europe PMC, PubMed Central, Scopus, Web of Science; Recommended: F1000 Prime; Discussed: PLoS Comments, Facebook, Reddit, Twitter, Wikipedia.[1]

PLoS articles are provided in three different formats—page views, PDF downloads, and XML downloads and we record the online activity of users across these three formats. This “usage”, comprised of the three types, is provided as an aggregate metric or broken down, month-by-month in graphical format. Online usage via the PLoS platform is presented according to industry standard definitions of usage and is COUNTER-compliant.

We also display COUNTER 3-compliant PMC usage data for each article. PMC individually counts the number of page views and PDF downloads of the article on their site. The results are only made available to PLoS once a month, not in real-time. As a result, articles may experience a lag with the display of PMC data of up to one month. This will also impact the data shown on recently published articles, which may not show PMC usage data for their first month of publication.[2]

Usage counts of WoS

The usage count is a measure of the level of interest in a specific item on the WoS platform. The count reflects the number of times the article has met a user’s information needs

[1] https://www.plos.org/article-level-metrics.

[2] http://www.lagotto.io/plos/.

Table 2 Descriptive statistics of views per journal

Journal	Type	Median	Mean	Journal	Type	Median	Mean
BIO	PLoS HTM	8727	10,411	NTD	PLoS View	2544	3053
	PMC View	530	602		PMC View	466	742
	WoS 2013	14	21		WoS 2013	8	10
CBI	PLoS View	3640	4645	ONE	PLoS View	1708	2513
	PMC View	314	416		PMC View	411	599
	WoS 2013	10	12		WoS 2013	10	15
GEN	PLoS View	3950	4995	PAT	PLoS View	3870	4540
	PMC View	529	644		PMC View	554	644
	WoS 2013	10	15		WoS 2013	9	12
MED	PLoS View	10,547	13,701	–	–	–	–
	PMC View	1078	1475	–	–	–	–
	WoS 2013	13	20	–	–	–	–

"–" means null

as demonstrated by clicking links to the full-length article at the publisher's website (via direct link or Open-Url) or by saving the article for use in a bibliographic management tool (via direct export or in a format to be imported later). The usage count is a record of all activity performed by all WoS users, not just activity performed by users at your institution. Usage counts for different versions of the same item on the WoS platform are unified. Usage counts are updated daily. There are two kinds of usage counts in WoS platform.

Last 180 days This is the count of the number of times the full text of a record has been accessed or a record has been saved in the last 180 days. This count can move up or down as the end date of the fixed period advances.
Since 2013 This is the count of the number of times the full text of a record has been accessed or a record has been saved since February 1, 2013. This count can increase or remain static over time.[3]

In general, ''usage'' often refers to HTML views and PDF downloads. HTML views and PDF downloads of PLoS and PMC belong to traditional usage data. XML downloads of PLoS are new usage data. Different from traditional definitions, WoS defines "usage" as "clicking" and "saving" (Wang et al. 2016a). More accurately, WoS usage should be "HTML view" and "saving" (Chen 2018). From Tables 2 and 3, they show that usage counts in WoS and XML downloads of PLoS are considerably less than HTML views and PDF downloads of PLoS and PMC. Considering traditional definitions of usage data, only HTML views and PDF downloads of PLoS and PMC are investigated in this study. Besides, usage counts in PMC are also significantly less than that of PLoS, which means that PLoS official websites are the primary channel for users to view and download articles.

[3] http://images.webofknowledge.com/WOKRS519B3/help/WOK/hp_usage_score.html.

Table 3 Descriptive statistics of downloads per journal

Journal	Type	Median	Mean	Journal	Type	Median	Mean
BIO	PLoS PDF	1277	1482	NTD	PLoS PDF	414	504
	PLoS XML	51	61		PLoS XML	29	35
	PMC PDF	193	233		PMC PDF	174	216
CBI	PLoS PDF	584	782	ONE	PLoS PDF	350	475
	PLoS XML	28	37		PLoS XML	33	35
	PMC PDF	104	129		PMC PDF	153	200
GEN	PLoS PDF	699	882	PAT	PLoS PDF	701	882
	PLoS XML	29	37		PLoS XML	28	33
	PMC PDF	184	231		PMC PDF	214	254
MED	PLoS PDF	1118	1607	–	–	–	–
	PLoS XML	58	74	–	–	–	–
	PMC PDF	438	515	–	–	–	–

"–" means null

Table 4 Category assignment based on different usage data

Category	BIO	CBI	GEN	MED	NTD	ONE	PAT
Top20% ranked by PLoS View	57	223	302	34	274	11,472	236
Bottom20% ranked by PLoS View	57	223	302	34	274	11,496	237
Top20% ranked by PLoS PDF	57	224	302	34	275	11,488	236
Bottom20% ranked by PLoS PDF	57	223	303	34	275	11,552	236
Top20% ranked by PMC View	57	224	302	34	274	11,476	236
Bottom20% ranked by PMC View	57	224	304	34	274	11,489	237
Top20% ranked by PMC PDF	57	224	302	34	274	11,492	237
Bottom20% ranked by PMC PDF	57	228	304	34	283	11,474	240

Methods

Article classification strategy

In terms of Pareto principle (or 80/20 rule), highly browsed and downloaded academic articles in this study are defined by Top 20% papers ranked by HTML views and PDF downloads in PLoS and PMC platforms respectively. In order to comparatively uncover linguistic characteristics of Top 20% papers, total papers and Bottom 20% papers are also incorporated (detailed number of publications in each category are shown in Table 4).

Indicators measuring linguistic characteristics

Linguistic complexity comprises two aspects: syntactic and lexical complexity. Syntactic complexity consists of quantitative variables on sentence length, sentence complexity, and others (Ferris 1994; Kormos 2011; Ojima 2006). Lexical complexity is made up of lexical diversity, lexical density, and lexical sophistication (Vajjala and

Table 5 Indicators measuring linguistic characteristics

Indicator	Description	Formula
Title Length	Calculating total number of words in each article title	$TTL = \sum_{i=1}^{N} \text{Title}$
Abstract Length	Calculating total number of words in each article abstract	$TAL = \sum_{i=1}^{N} \text{Abstract}$
Full-text Length	Calculating total number of words in each article (main body)	$TFL = \sum_{i=1}^{N} \text{Full text}$
Sentence Length	Calculating average number of words in sentences of each article	$MSL = \frac{\sum_{i=1}^{N} SL_i}{N}$
Lexical Diversity	Type-Token Ratio in each article	$TTR = \frac{\#\text{ of Distinct words}}{\#\text{ of Tokens}}$
Lexical Density	Counting the ratio of lexical items in tokens in each article based on their part of speech (lexical class)	$\text{Type Ratio} = \frac{\#\text{ of Type items}}{\#\text{ of Tokens}}$
Lexical Sophistication	Counting the length of nouns, verbs, adjectives, and adverbs	$MWL = \frac{\sum_{1}^{N} WL_i}{N}$

Meurers 2012). Lu et al. (2019a, b) selected several indicators (sentence length, sentence complexity, lexical diversity, lexical density and lexical sophistication) to measure linguistic complexity.

In this study, more comprehensive indicators measuring linguistic characteristics were adopted compared with former research. The indicators selected follow the structures of academic article, which are “Title–Abstract–Keyword–Full text–Sentence–Word”. Specifically, title length, abstract length, full-text length, sentence length, lexical diversity, lexical density and lexical sophistication are incorporated in this study (shown in Table 5). “Keyword number” is not applied in this study, because there are no keywords in original articles in PLoS platform. “Co-author number” is not selected because it measures co-authorship (Chi and Glänzel 2017). Strictly speaking, it can’t be classified to the category of linguistic characteristics. In addition, punctuation marks are removed from the calculations.

Applicability of indicators to PLoS and PubMed views and downloads

In this study, it is assumed that linguistic characteristics have an influence on article views and downloads to some extent. It means that linguistic characteristics may be the reason why the paper is viewed or downloaded. After conducting experiments of browsing and downloading papers on PLoS and PubMed platforms, it is found that before browsing full text of the paper, only paper title and author name can be seen, and after clicking paper title, abstract and full text can be read. Then, readers can choose to download the paper by clicking the “Download” button on the full-text page. So, before the paper is viewed, readers only know title and author name, and before the paper is downloaded, the readers know title, author name, abstract, full text and so on. Therefore, only indicator “title length” is applicable to PLoS and PubMed views, and indicators “title length, abstract length, full-text Length, sentence length, lexical diversity, lexical density and lexical sophistication” are applicable to PLoS and PubMed downloads.

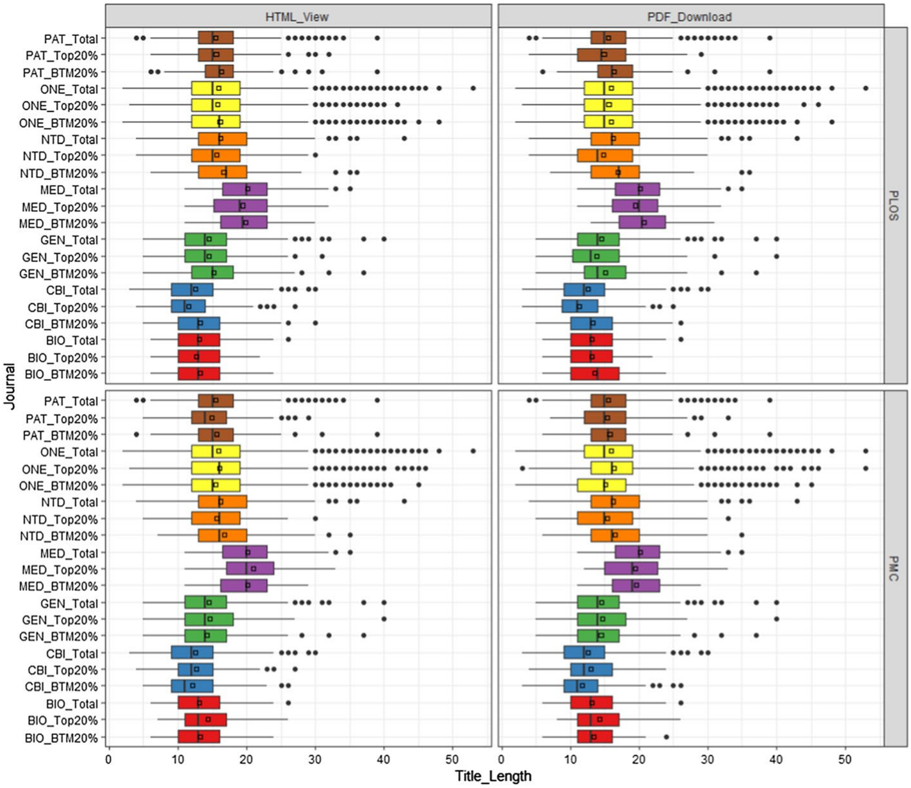

Fig. 1 Title word length distribution in PLoS and PMC platforms (color figure online)

Results

Title length distribution

In Fig. 1, each journal is plotted by a unique color, articles of three categories in each journal are plotted as points inside or outside the boxes, vertical line and hollow square inside or outside each box denote median and mean respectively. Mean of title length by three categories in each journal are listed in Table 6.

From Fig. 1 and Table 6, it reveals that Top 20% viewed and downloaded papers have less title length in average than total and Bottom 20% papers in each journal in PLoS platform (only Top 20% viewed papers of journal *PLoS Pathogens* and Top 20% downloaded papers of journal *PLoS Biology* are excluded). But it shows no regular characteristics in each journal in PMC platform. Then, generally, Top 20% viewed and downloaded papers of each journal in PLoS platform have less average title length than that of PMC platform. Finally, in Top 20% viewed and downloaded papers, *PLoS Medicine* have the most title length in average, then *PLoS One*, *PLoS Pathogens* and *PLoS Neglected Tropical Diseases*, *PLoS Genetics*, *PLoS Computational Biology* and *PLoS Biology* are the least in general.

After tracing submission guidelines of each PLoS journal, it is found that there are character limits of title, no more than 250 characters in *PLoS Biology*, *PLoS One* and *PLoS*

Table 6 Mean of title and abstract length by three categories in each journal

Category	PLoS title view	PMC title view	PLoS title download	PMC title download	PLoS abstract download	PMC abstract download
BIO_BTM20%	13.298	13.193	13.544	13.404	214.544	226.947
BIO_Top20%	12.702	14.316	13.158	14.228	237.719	218.719
BIO_Total	13.056	13.056	13.056	13.056	221.587	221.587
CBI_BTM20%	13.260	12.143	13.287	11.754	229.668	221.974
CBI_Top20%	11.610	12.732	11.286	12.960	228.996	238.045
CBI_Total	12.534	12.534	12.534	12.534	231.053	231.053
GEN_BTM20%	15.189	14.237	15.089	14.451	232.568	228.507
GEN_Top20%	14.460	14.629	13.868	14.725	237.788	235.384
GEN_Total	14.463	14.463	14.463	14.463	231.014	231.014
MED_BTM20%	19.912	20.088	20.706	19.618	369.500	380.206
MED_Top20%	19.441	21.029	19.441	19.412	366.559	371.912
MED_Total	20.152	20.152	20.152	20.152	379.123	379.123
NTD_BTM20%	16.682	16.774	16.953	16.551	254.942	260.286
NTD_Top20%	15.624	15.588	14.829	15.310	265.000	267.854
NTD_Total	16.220	16.220	16.220	16.220	261.966	261.966
ONE_BTM20%	16.119	15.426	15.973	15.135	232.989	225.529
ONE_Top20%	15.747	16.094	15.572	16.298	237.643	246.135
ONE_Total	15.901	15.901	15.901	15.901	235.824	235.824
PAT_BTM20%	16.367	15.637	16.415	15.721	237.907	228.100
PAT_Top20%	15.559	14.881	14.873	15.418	236.419	240.873
PAT_Total	15.549	15.549	15.549	15.549	234.800	234.800

Neglected Tropical Diseases, no more than 200 characters in *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Pathogens* and *PLoS Medicine*. Words consist of characters, so title character length (spaces are removed) distribution in PLoS and PMC platforms are shown in Fig. 2, which reveals that character length of most articles in each journal are within the limits of submission guidelines. But within 200 characters, different journal has the unique title character length.

Abstract length and full-text length distribution

From Fig. 3a and Table 6, it reveals that Top 20% downloaded papers have more abstract length in average than total or Bottom 20% papers in most journals, but the differences are marginal. Then, in Top 20% downloaded papers, *PLoS Medicine* have the most average length in average, then *PLoS Neglected Tropical Diseases*, other journals are the least in general. After checking submission guidelines of each journal, it is found that there are no words limits of abstract in *PLoS Biology*, no more than 300 words in *PLoS Computational Biology*, *PLoS Genetics*, *PLoS One* and *PLoS Pathogens*, less than 250–300 words in *PLoS Neglected Tropical Diseases* and less than 500 words in *PLoS Medicine*. Probably, the average abstract length of each journal is affected by submission guidelines.

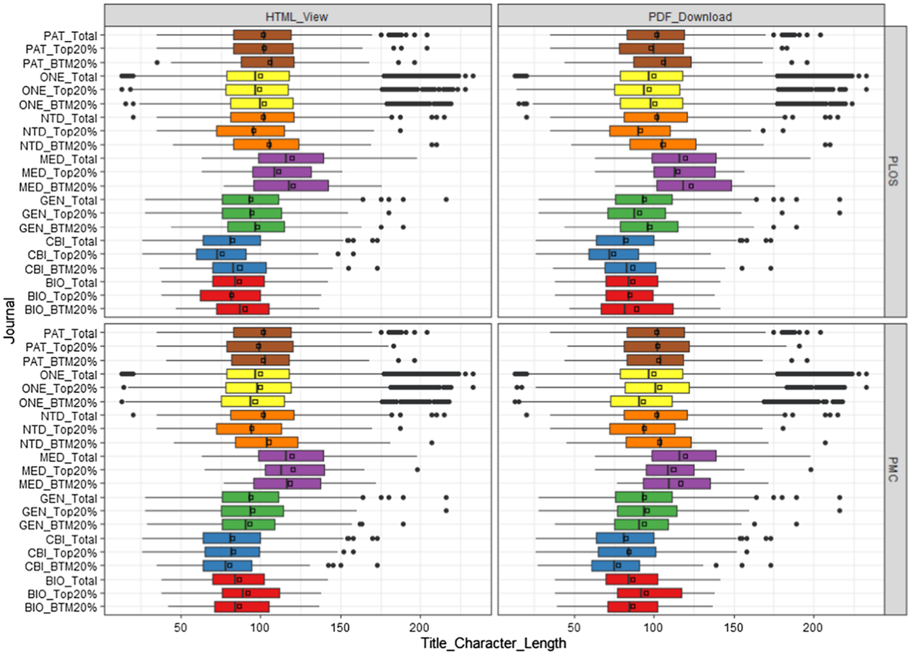

Fig. 2 Title character length distribution in PLoS and PMC platforms (color figure online)

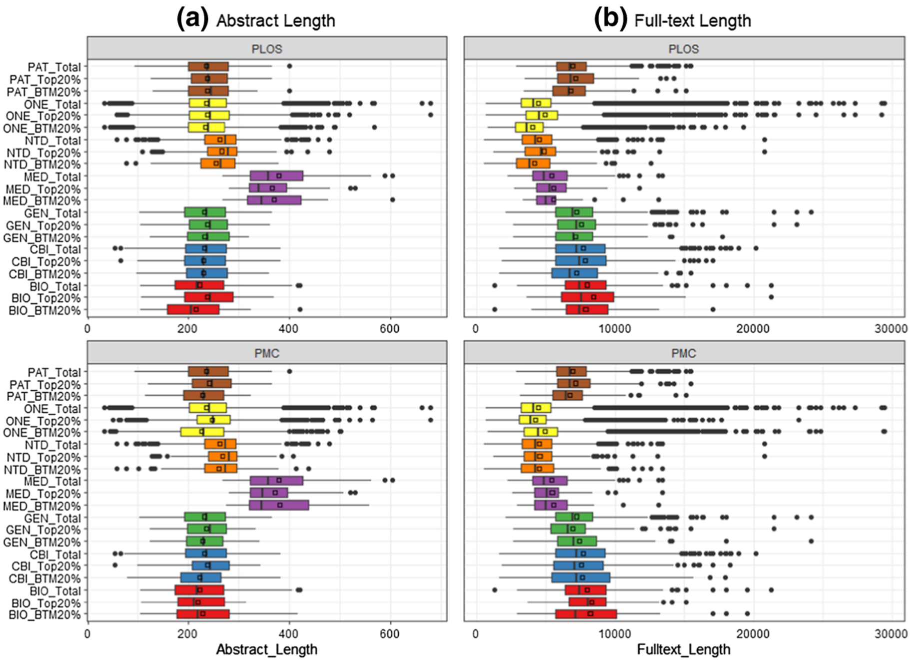

Fig. 3 Abstract length and full-text length ($\leq$30,000 words) distribution in PLoS and PMC platforms (color figure online)

Table 7 Download mean of full-text, sentence length and lexical diversity by three categories in each journal

Category	PLoS full text	PMC full text	PLoS sentence	PMC sentence	PLoS lexical diversity	PMC lexical diversity
BIO_BTM20%	7863.421	8208.316	24.062	25.044	0.225	0.218
BIO_Top20%	8450.772	8317.842	25.017	23.299	0.220	0.230
BIO_Total	7974.319	7974.319	23.911	23.911	0.227	0.227
CBI_BTM20%	7216.628	7648.026	23.691	24.206	0.201	0.199
CBI_Top20%	7809.321	7535.031	23.801	23.403	0.200	0.205
CBI_Total	7661.292	7661.292	23.630	23.630	0.199	0.199
GEN_BTM20%	7134.413	7405.829	23.243	23.915	0.230	0.222
GEN_Top20%	7525.510	6906.444	23.050	22.496	0.226	0.240
GEN_Total	7208.145	7208.145	22.896	22.896	0.232	0.232
MED_BTM20%	5531.206	5539.882	26.902	27.368	0.234	0.235
MED_Top20%	5572.059	5451.441	26.992	26.744	0.235	0.245
MED_Total	5424.035	5424.035	26.919	26.919	0.240	0.240
NTD_BTM20%	4188.084	4539.777	23.416	23.938	0.286	0.278
NTD_Top20%	4894.236	4559.478	23.324	23.076	0.270	0.275
NTD_Total	4497.738	4497.738	23.283	23.283	0.280	0.280
ONE_BTM20%	4017.922	4919.450	23.323	23.734	0.275	0.253
ONE_Top20%	4946.275	4229.520	23.076	22.652	0.262	0.277
ONE_Total	4477.960	4477.960	23.081	23.081	0.269	0.269
PAT_BTM20%	6781.716	6707.604	23.145	23.654	0.240	0.243
PAT_Top20%	7125.322	7154.751	22.708	22.612	0.241	0.238
PAT_Total	6965.152	6965.152	22.877	22.877	0.240	0.240

In order to reveal more details, only papers with "full-text length ≤ 30,000 words" are captured from the global graph. From Fig. 3b and Table 7, it reveals that only Top 20% downloaded papers in PLoS platform (top right) have more full-text length in average than total and Bottom 20% papers. Then, generally, in Top 20% downloaded papers, *PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics* and *PLoS Pathogens* have the most full-text length in average, then *PLoS Medicine*, *PLoS One* and *PLoS Neglected Tropical Diseases*. Finally, Top 20% downloaded papers of each journal in PLoS platform have more average full-text length than that of PMC platform (only journal *PLoS Pathogens* is excluded).

Sentence length and lexical diversity distribution

In order to reveal more details, only papers with "average sentence length ≤ 50 words" are captured from the global graph. From Fig. 4a and Table 7, it reveals that Top 20% downloaded papers have less average sentence length than total and Bottom 20% papers

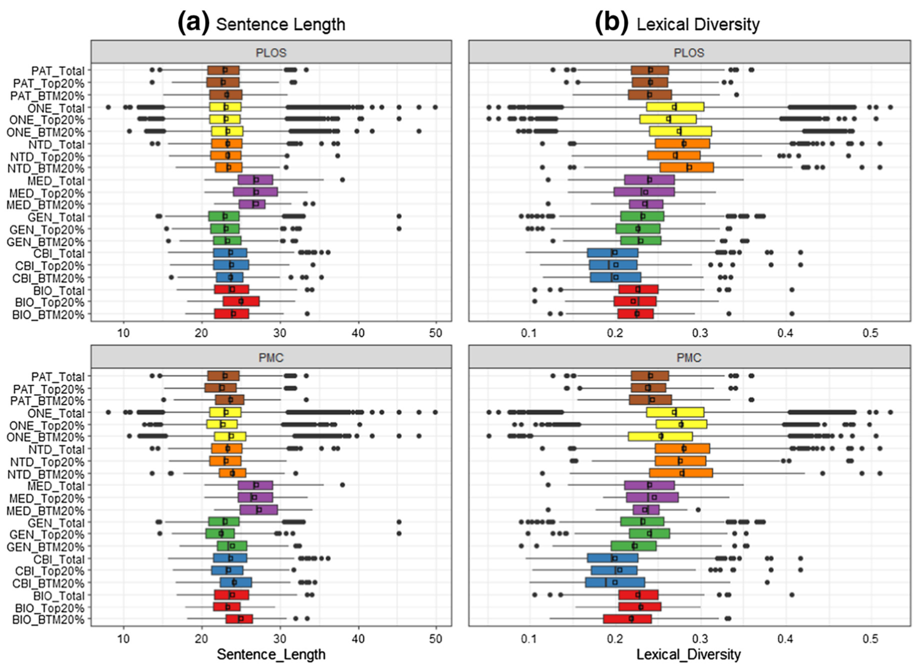

Fig. 4 Sentence length (≤50 words) and lexical diversity distribution in PLoS and PMC platforms (color figure online)

in general, especially in PMC platform, although the differences are marginal. Also, generally, in Top 20% downloaded papers, *PLoS Medicine* have the most average sentence length, then *PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS Pathogens*.

From Fig. 4b and Table 7, it reveals that average Type-Token Ratios of Top 20% downloaded papers are greater than 20%. Then, it reveals that Top 20% downloaded papers have less average Type-Token Ratios than total and Bottom 20% papers in most journals in PLoS platform, but the results are opposite in PMC platform. Also, in Top 20% downloaded papers, *PLoS Neglected Tropical Diseases* and *PLoS One* have the most lexical diversity in average, then *PLoS Pathogens*, *PLoS Medicine*, *PLoS Genetics* and *PLoS Biology*, *PLoS Computational Biology* is the least.

Lexical density distribution

Lexical density is only measured by lexical items, including nouns, verbs, adjectives and adverbs, whereas other types of words, for example, preposition, are not considered in this study. In order to reveal more details, papers with "noun ratio ≤ 0.6" are captured from the global graph. From Figs. 5 and 6, it is found that among the lexical items, nouns are used most frequently, then verbs and adjectives, adverbs are the least.

In Fig. 5a, it reveals that Top 20% downloaded papers have more average noun ratio than total and Bottom 20% papers in most journals in PMC platform, but show no regular differences among three categories in PLoS platform. Also, in Top 20% downloaded

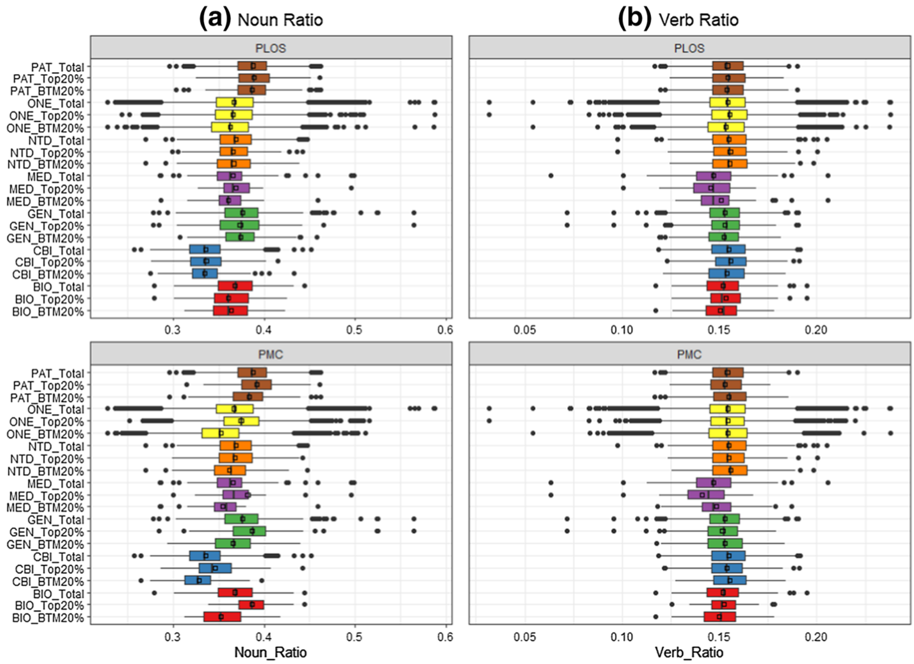

Fig. 5 Noun ratio (≤ 0.6) and verb ratio distribution in PLoS and PMC platforms (color figure online)

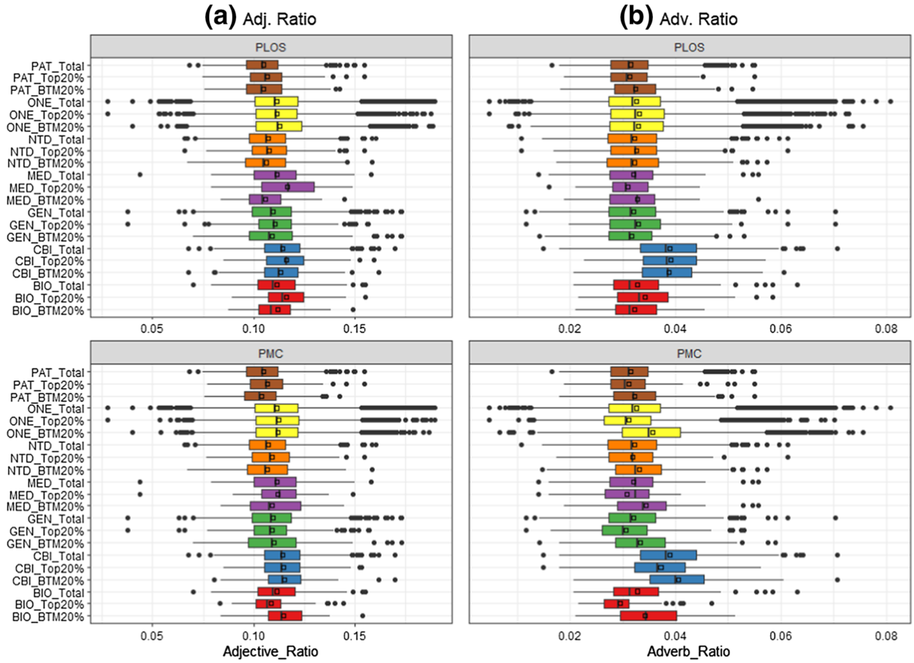

Fig. 6 Adjective and adverb ratio distribution in PLoS and PMC platforms (color figure online)

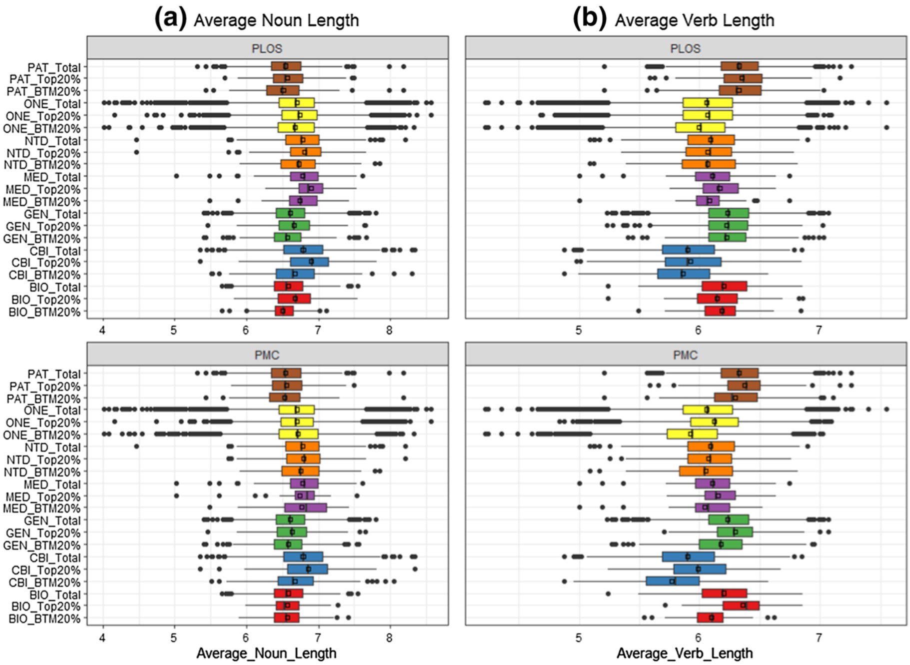

Fig. 7 Average noun and verb length distribution in PLoS and PMC platforms (color figure online)

papers, *PLoS Pathogens* and *PLoS Genetics* have the most noun ratio in average (about 38–39%), then *PLoS Neglected Tropical Diseases*, *PLoS One*, *PLoS Medicine* and *PLoS Biology* (about 36–37%), *PLoS Computational Biology* is the least (around 34%).

In Fig. 5b, it reveals that average verb ratios among Top 20%, total and Bottom 20% downloaded papers of each journal are marginal, precisely 15% or so. Similarly, from Fig. 6a, it reveals that average adjective ratios among three categories of each journal are marginal, around 11%, but journal *PLoS Pathogens* shows less average adverb ratio than others (10.5% vs. 11%). From Fig. 6b, it reveals that average adverb ratios among three categories of each journal are also marginal, around 3%, but journal *PLoS Computational Biology* shows more average adverb ratio than others (4% vs. 3%).

Lexical sophistication distribution

Figures 7 and 8 show the distributions of average lexical word (noun, verb, adjective and adverb) length by category respectively. Generally, in “total papers” category, average length of nouns (6.68) is longer than that of verbs (6.13) and adverbs (6.48), but shorter than that of adjectives (7.92). In “Top 20% downloaded papers” category, average length of adjectives (7.96) is the longest of all, then nouns (6.72), adverbs (6.53) and verbs (6.17).

From Fig. 7a, *PLoS Computational Biology* and *PLoS Medicine* have the most average noun length, then *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS*

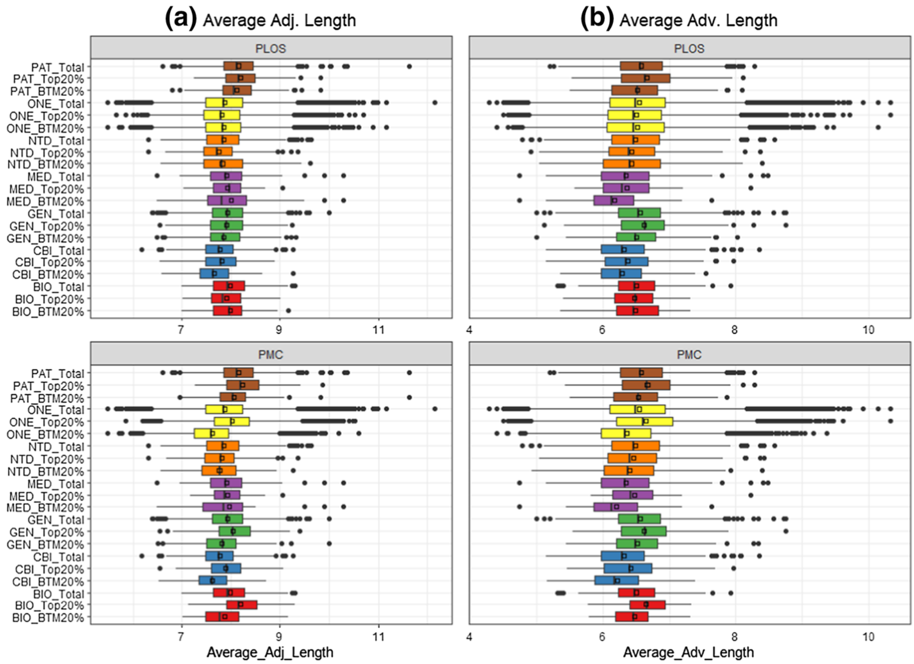

Fig. 8 Average adjective and adverb length distribution in PLoS and PMC platforms (color figure online)

Table 8 *p* values of Kolmogorov–Smirnov test of title length between "Top 20% viewed" and "BTM 20% viewed" in different journals and platforms

	BIO	CBI	GEN	MED	NTD	ONE	PAT
PLoS	0.9807	0.0007339	0.05432	0.9727	0.1398	4.349e−05	0.04042
PMC	0.3442	0.396	0.4851	0.9727	0.03758	6.994e−15	0.05228

Genetics, *PLoS Biology* and *PLoS Pathogens* are the least. From Fig. 7b, *PLoS Pathogens* has the most average verb length, then *PLoS Biology*, *PLoS Genetics*, *PLoS One*, *PLoS Medicine* and *PLoS Neglected Tropical Diseases*, *PLoS Computational Biology* is the least. From Fig. 8a, *PLoS Pathogens* has the most average adjective length, then *PLoS Biology*, *PLoS Genetics*, *PLoS One*, *PLoS Medicine*, *PLoS Computational Biology* and *PLoS Neglected Tropical Diseases*. From Fig. 8b, *PLoS Pathogens* and *PLoS Genetics* have the most average adverb length, then *PLoS One*, *PLoS Biology*, *PLoS Neglected Tropical Diseases* and *PLoS Medicine*, *PLoS Computational Biology* is the least.

Table 9 Spearman's correlation coefficient between views and title length of Top 20% viewed articles in different journals and platforms

	BIO	CBI	GEN	MED	NTD	ONE	PAT
PLoS	0.208	−0.022	−0.139*	0.050	−0.141*	−0.053***	−0.201**
PMC	0.229	−0.013	0.046	0.109	−0.014	−0.036***	−0.073

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$

Statistical significance test

Two-sample Kolmogorov–Smirnov (K–S) test for linguistic characteristics between "Top 20% papers" and "BTM 20% papers" categories are provided and p-values of K-S test are shown in Tables 8 and 10, indicating that the differences of the characteristics among categories are statistically significant or not. Also, Spearman's correlation coefficient between usage data of Top 20% articles and linguistic characteristics are investigated and shown in Tables 9 and 11.

Form Tables 8 and 10, about 40% K–S test results of linguistic characteristics between "Top 20% viewed and downloaded" and "BTM 20% viewed and downloaded" categories in different journals and platforms suggest statistical significance. For "Top 20% viewed" and "BTM 20% viewed" categories, title length of *PLoS Computational Biology*, *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS Pathogens* suggest statistical significance, especially in PLoS platform. For "Top 20% downloaded" and "BTM 20% downloaded" categories, title length of *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS Pathogens* suggest statistical significance, especially in PLoS platform. For "Top 20% downloaded" and "BTM 20% downloaded" categories, average sentence length of *PLoS Biology*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS Pathogens* suggest statistical significance, especially in PMC platform. For other linguistic characteristics between "Top 20% downloaded" and "BTM 20% downloaded" categories, they also show statistical significance, but they depend on different journals and platforms.

Form Tables 9 and 11, they show very weak positive or slightly negative correlation between usage data and linguistic characteristics in general. In Top 20% viewed articles, they show slightly negative correlation between views and title length in *PLoS Genetics*, *PLoS Neglected Tropical Diseases*, *PLoS One* and *PLoS Pathogens*, especially in PLoS platform. In Top 20% downloaded articles, they show slightly negative correlation between downloads and title length in *PLoS Neglected Tropical Diseases* and *PLoS One* in PLoS platform. For Top 20% downloaded articles in *PLoS One*, they show slightly negative correlation to lexical diversity, noun ratio, verb length, adjective length and adverb length, especially in PLoS platform. For Top 20% downloaded articles in *PLoS Biology*, they show weak negative correlation to adjective ratio and adverb length in PMC platform. For Top 20% downloaded articles in *PLoS Biology*, they show moderate negative correlation to adverb length in PLoS platform.

Table 10 p values of Kolmogorov–Smirnov test of linguistic characteristics between "Top 20% downloaded" and "BTM 20% downloaded" in different journals and platforms

	Title length	Abstract length	Full-text length	Sentence length	Lexical diversity	Noun ratio	Verb ratio	Adj. ratio	Adv. ratio	Noun length	Verb length	Adj. length	Adv. length
BIO_PLoS	0.9103	0.02241	0.4761	0.3466	0.788	0.9141	0.4796	0.2404	0.4796	0.001644	0.3466	0.4796	0.9824
BIO_PMC	0.9103	0.9103	0.3442	0.03818	0.1604	4.391e−08	0.4796	0.003331	2.626e−05	0.9824	4.447e−07	0.001644	0.06393
CBI_PLoS	0.000675	0.974	0.06854	0.4814	0.6225	0.5584	0.1739	0.1331	0.7872	2.969e−08	0.2403	0.009742	0.2519
CBI_PMC	0.04741	0.005116	0.6018	0.01908	0.03142	1.903e−10	0.3691	0.7813	0.0002985	5.02e−06	2.505e−07	1.004e−07	0.0001167
GEN_PLoS	0.005451	0.3668	0.431	0.444	0.5056	0.07871	0.3787	0.003585	0.07679	0.00121	0.4888	0.5471	0.006845
GEN_PMC	0.3158	0.07971	0.01763	1.497e−07	1.056e−08	5.095e−13	0.5518	0.1371	2.16e−05	0.09991	7.038e−08	4.589e−07	0.01755
MED_PLoS	0.8558	0.8558	0.8558	0.6727	0.8632	0.3068	0.8632	0.01319	0.4728	0.1859	0.1859	0.8632	0.6727
MED_PMC	0.9994	0.6649	0.8558	0.9762	0.3068	0.05637	0.1859	0.3068	0.6727	0.4728	0.4728	0.4728	0.1057
NTD_PLoS	2.231e−05	0.02988	0.000323	0.4611	0.004429	0.2058	0.9804	0.1411	0.4611	0.01049	0.3991	0.06072	0.3991
NTD_PMC	0.02413	0.1419	0.4305	0.0001405	0.1145	0.05081	0.3753	0.1206	0.07627	0.2695	0.1278	0.26	0.8371
ONE_PLoS	1.842e−06	2.2e−16	2.2e−16	2.238e−08	2.2e−16	2.2e−16	2.2e−16	2.885e−11	0.1172	2.2e−16	2.2e−16	2.471e−06	0.01023
ONE_PMC	2.2e−16	2.2e−16	2.2e−16	2.2e−16	2.2e−16	2.2e−16	2.428e−06	0.4415	2.2e−16	3.489e−13	2.2e−16	2.2e−16	2.2e−16
PAT_PLoS	0.0002553	0.4985	0.114	0.03409	0.5729	0.65	0.9598	0.365	0.114	0.07216	0.1741	0.3075	0.04414
PAT_PMC	0.1168	0.06395	0.07946	1.848e−05	0.2103	0.000104	0.3044	0.01633	0.02749	0.3775	6.193e−05	0.000376	0.01576

Table 11 Spearman's correlation coefficient between downloads and linguistic characteristics of Top 20% downloaded articles in different journals and platforms

	Title length	Abstract length	Full-text length	Sentence length	Lexical diversity	Noun ratio	Verb ratio	Adj. ratio	Adv. ratio	Noun length	Verb length	Adj. length	Adv. length
BIO_PLoS	0.118	−0.182	−0.017	0.044	−0.054	−0.147	−0.106	0.080	0.255	0.186	−0.098	−0.090	−0.047
BIO_PMC	0.084	0.185	−0.081	0.028	−0.072	0.161	−0.157	**−0.294***	−0.056	−0.170	−0.016	0.069	**−0.330***
CBI_PLoS	−0.101	−0.058	0.074	−0.064	−0.041	0.042	−0.017	0.052	−0.017	−0.047	−0.046	0.030	0.005
CBI_PMC	−0.004	0.086	−0.031	−0.061	0.036	0.075	0.016	−0.005	−0.050	0.031	0.045	0.154*	0.062
GEN_PLoS	0.014	−0.030	−0.072	0.081	−0.020	−0.078	−0.006	0.032	0.033	0.141*	−0.010	−0.015	−0.073
GEN_PMC	0.025	0.089	0.015	0.076	−0.002	0.044	−0.079	−0.062	−0.024	0.032	0.003	0.020	0.096
MED_PLoS	0.024	−0.075	0.047	0.019	0.032	0.068	0.216	−0.087	−0.060	−0.033	−0.075	−0.125	**−0.444****
MED_PMC	0.196	0.061	0.110	−0.274	0.137	0.282	−0.095	0.003	−0.197	0.026	−0.007	−0.133	−0.184
NTD_PLoS	**−0.222*****	−0.012	0.053	−0.050	0.007	−0.008	0.046	0.083	0.113	0.044	−0.018	−0.040	0.086
NTD_PMC	−0.037	0.033	0.094	−0.026	−0.076	0.035	−0.101	0.071	0.052	0.031	−0.001	−0.131	−0.043
ONE_PLoS	**−0.064*****	0.011	0.057***	0.064***	**−0.058*****	**−0.116*****	0.060***	0.041***	0.095***	0.059***	**−0.079*****	**−0.103*****	**−0.080*****
ONE_PMC	0.008	0.058***	0.026**	−0.014	**−0.039*****	**−0.037*****	0.032***	0.021*	0.039***	0.045***	0.004	**−0.019***	−0.010
PAT_PLoS	0.027	−0.035	−0.044	0.027	−0.003	−0.053	−0.018	0.059	−0.014	0.026	0.112	0.011	0.071
PAT_PMC	0.026	0.080	0.028	−0.041	−0.034	−0.080	0.079	0.006	0.014	0.029	−0.088	−0.076	0.018

Negatively correlated and statistically significant results ($p \leq 0.001$) are in bold font style

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$

Discussions and conclusions

This paper applied computational linguistics to understand the relationship between linguistic characteristics and article views and downloads. The mean and median results show marginal differences for most linguistic characteristics among different categories; statistical significance test results indicate no statistical significance generally; however, for certain linguistic characteristics (e.g. title length and average sentence length) in different PLoS journals and platforms, they are still statistically significant.

Despite most linguistic characteristics play little role in article views and downloads in our data sets in general, some linguistic characteristics (e.g. title length and average sentence length) in specific PLoS journal and platform play certain role in article views and downloads in our data sets. Also, academic papers in this study follow some patterns of linguistic characteristics. For example, the average length of sentences in sample papers is usually greater than 22 words; average Type-Token Ratios of sample papers are greater than 20%; average ratios of noun, verb, adjective and adverb are about 35–39%, 15%, 11% and 3% respectively. Besides, each journal has its own linguistic characteristics. Differences of linguistic characteristics between two platforms are also existed.

Jamali and Nikzad (2011) found that articles with longer titles were downloaded slightly less than the articles with shorter titles, but Duan and Xiong (2017) found that there were only weak correlations between total downloads and title length and held that the correlation between downloads and title length could be different due to data differences. In our mind, social factors should be considered, for example, each journal has unique submission guidelines to limit characters or words of article length.

Apart from submission guidelines of journals (eg. word limits of title, abstract and full-text length), other social factors also should be incorporated into to understand linguistic characteristics of academic articles. First of all, each discipline follows its own research paradigm and covers unique terminology. Then, diverse users with various ages, positions and academic backgrounds prefer different academic platforms to acquire academic papers. In empirical research, to keep balance between disciplines and journals in sampling and grouping, and to compare or combine usage data from different academic platforms should also be valued.

There are also some limitations in this study. Only papers published between 2014 and 2015 in PLoS journals are investigated, therefore, the conclusions might be different in more samples and other journals, which need further experiments. Only several basic indicators measuring linguistic characteristics are adopted, more diversified and semantic indicators can be incorporated. Although these limitations exist, we hope that this first introduction of multi-granularity linguistic characteristics to usage metrics would provide a new perspective. In further study, in-depth interviews and experiments of user behaviors will be combined with linguistic characteristics to investigate user motivation and behavior pattern of usage metrics.

Acknowledgements This paper is supported by Youth Program of National Social Science Fund in China (15CTQ035), Social Public Safety S&T Collaborative Innovation Center of Universities in Jiangsu Province and China Scholarship Council (ID: 201906845042).

Appendix

See Tables 6, 7, 8, 9, 10 and 11.

References

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE, 9*(3), e92590.

Bollen, J., Luce, R., Vemulapalli, S. S., & Xu, W. (2002). Usage analysis for the identification of research trends in digital libraries. *D-Lib Magazine*. https://doi.org/10.1045/may2003-bollen.

Bollen, J., Sompel, H. V. D., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing and Management, 41*(6), 1419–1440.

Bonzi, S., & Snyder, H. W. (1991). Motivations for citation: A comparison of self citation and citation to others. *Scientometrics, 21*(2), 245–254.

Boyack, K. W., Eck, N. J. V., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59–73.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9), 1759–1767.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the Association for Information Science and Technology, 37*(1), 34–36.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the Association for Information Science and Technology, 40*(4), 284–290.

Case, D. O., & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635–645.

Chen, B. (2018). Usage pattern comparison of the same scholarly articles between Web of Science (WoS) and Springer. *Scientometrics, 115*(1), 519–537.

Chen, B., Zhong, Z., & Zhan, C. (2017). Usage pattern analysis of academic articles from two Chinese journals. In K. Holmberg & J. Vainio (Eds.), *Proceedings of ISSI 2017* (pp. 366–375). Wuhan: Wuhan University.

Chen, B., Zhou, H., Zhong, Z., & Wang, Y. (2018). Exploring the user platform preference and user interest preference of chinese scholarly articles: A comparison based on usage metrics. *Journal of Library Science in China, 44*(6), 90–104. **(in Chinese)**.

Chi, P. S., & Glänzel, W. (2018). Comparison of citation and usage indicators in research assessment in scientific disciplines and journals. *Scientometrics, 116*(1), 537–554.

Chi, P. S., & Glänzel, W. (2017). An empirical investigation of the associations among usage, scientific collaboration and citation impact. *Scientometrics, 112*(1), 403–412.

Davis, P. M. (2006). Ejournal interface can influence usage statistics: Implications for libraries, publishers, and project counter. *Journal of the Association for Information Science and Technology, 57*(9), 1243–1248.

Davis, P. M., & Solla, L. R. (2003). An ip-level analysis of usage statistics for electronic journals in chemistry: Making inferences about user behavior. *Journal of the American Society for Information Science and Technology, 54*(11), 1062–1068.

De Sordi, O. J., Conejero, M. A., & Meireles, M. (2016). Bibliometric indicators in the context of regional repositories: Proposing the d-index. *Scientometrics, 107*(1), 235–258.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., et al. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE, 8*(8), e71416.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*(9), 1820–1833.

Duan, Y., & Xiong, Z. (2017). Download patterns of journal papers and their influencing factors. *Scientometrics, 112*(3), 1761–1775.

Elgendi, M. (2019). Characteristics of a highly cited article: A machine learning perspective. *IEEE Access, 7,* 87977–87986.

Ferris, D. R. (1994). Rhetorical strategies in student persuasive writing: Differences between native and non-native English speakers. *Research in the Teaching of English, 28*(1), 45–65.

Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA)—a new approach for identifying related work based on co-citation analysis. In B. Larsen & J. Leta (Eds.), *Proceedings of ISSI 2009* (pp. 571–575). Wuhan: Wuhan University.

Gorraiz, J., Gumpenberger, C., & Schloegl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics, 101*(2), 1077–1095.

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*(4), 887–896.

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88*(2), 653–661.

Khan, M. S., & Younas, M. (2017). Analyzing readers behavior in downloading articles from IEEE digital library: A study of two selected journals in the field of education. *Scientometrics, 110*(3), 1523–1537.

Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics, 10*(4), 954–966.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*(2), 148–161.

Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology, 44*(1), 1–64.

Kurtz, M. J., & Henneken, E. A. (2016). Measuring metrics-a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the Association for Information Science and Technology, 68*(3), 695–708.

Lippi, G., & Favaloro, E. J. (2013). Article downloads and citations: Is there any relationship? *Clinica Chimica Acta, 415,* 195.

Liu, S., & Chen, C. (2012). The proximity of co-citation. *Scientometrics, 91*(2), 495–511.

Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology, 64*(3), 627–639.

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., et al. (2019a). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., et al. (2019b). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*(5), 462–475.

Lu, C., Ding, Y., & Zhang, C. (2017). Understanding the impact change of a highly cited article: A content-based citation analysis. *Scientometrics, 112*(2), 927–945.

Mckeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology, 67*(11), 2684–2696.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology, 56*(10), 1088–1097.

Moed, H. F., & Halevi, G. (2016). On full text download and citation distributions in scientific-scholarly journals. *Journal of the Association for Information Science and Technology, 67*(2), 412–431.

Ojima, M. (2006). Concept mapping as pre-task planning: A case study of three Japanese ESL writers. *System, 34*(4), 566–585.

O'Leary, D. E. (2008). The relationship between citations and number of downloads in decision support systems. *Decision Support Systems*, *45*(4), 972–980.

Pan, X., Yan, E., Cui, M., & Hua, W. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *Journal of Informetrics, 12*(2), 481–493.

Pan, X., Yan, E., Cui, M., & Hua, W. (2019). How important is software to library and information science research? A content analysis of full-text publications. *Journal of Informetrics, 13*(1), 397–406.

Pan, X., Yan, E., & Hua, W. (2016). Disciplinary differences of software use and impact in scientific literature. *Scientometrics, 109*(3), 1–18.

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics, 9*(4), 860–871.

Schloegl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics, 101*(2), 1113–1128.

Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics, 87*(2), 373–388.

Subotic, S., & Mukherjee, (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science, 40*(1), 115–124.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the seventh workshop on building educational applications using NLP (pp. 163–173), July 8–14, 2012, Jelu Island, South Korea.

Wan, J. K., Hua, P. H., Rousseau, R., & Sun, X. K. (2010). The journal download immediacy index (DII): Experiences using a chinese full-text database. *Scientometrics, 82*(3), 555–566.

Wang, X., Fang, Z., & Sun, X. (2016a). Usage patterns of scholarly articles on Web of Science: A study on Web of Science usage count. *Scientometrics, 109*(2), 917–926.

Wang, X., Peng, L., Zhang, C., Xu, S., Wang, Z., Wang, C., et al. (2013a). Exploring scientists' working timetable: A global survey. *Journal of Informetrics, 7*(3), 665–675.

Wang, X., Wang, Z., & Xu, S. (2013b). Tracing scientist's research trends realtimely. *Scientometrics, 95*(2), 717–729.

Wang, X., Xu, S., & Fang, Z. (2016). *Tracing digital footprints to academic articles: An investigation of PeerJ* publication referral data. Retrieved October 28, 2018, from http://cn.arxiv.org/abs/1601.05271.

Wang, X., Xu, S., Peng, L., Wang, Z., Wang, C., Zhang, C., et al. (2012). Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics, 6*(4), 655–660.

Wang, Y., & Zhang, C. (2018). Using full-text of research articles to analyze academic impact of algorithms. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Proceedings of iConference* (pp. 395–401). Sheffield: University of Sheffield.

Zhang, C., Ding, R., & Wang, Y. (2018). Using behavior and influence assessment of algorithms based on full-text academic articles. *Journal of the China Society for Scientific and Technical Information, 37*(12), 1175–1187. **(in Chinese)**.

Zhao, X. (2017). Exploring the features of usage data for academic literatures. *Journal of Library Science in China, 43*(3), 44–57. **(in Chinese)**.

Zhao, S. X., Lou, W., Tan, A. M., & Yu, S. (2018). Do funded papers attract more usage? *Scientometrics, 115*(1), 153–168.