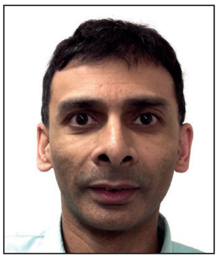
A COMPARISON OF TITLE WORDS FOR JOURNAL ARTICLES AND *WIKIPEDIA* PAGES: COVERAGE AND STYLISTIC DIFFERENCES?

Comparación de palabras de títulos de artículos de revista y páginas de *Wikipedia*: ¿Diferencias de cobertura y de estilo?

Mike Thelwall and Pardeep Sud

Mike Thelwall leads the *Statistical Cybermetrics Research Group* at the *University of Wolverhampton*, UK. He has developed and evaluated free software and methods for systematically gathering and analysing web and social web data, including for sentiment analysis, altmetrics and webometrics, and for *Mendeley*, *Twitter*, *YouTube*, *Google Books*, blogs and the general Web. He also conducts evaluation exercises for large organisations using web data, including for various divisions within the *United Nations* and *European Commission*. He has co-authored hundreds of refereed journal articles and has written three books.
http://www.scit.wlv.ac.uk/~cm1993/mycv.html
https://orcid.org/0000-0001-6065-205X

m.thelwall@wlv.ac.uk



Pardeep Sud is a member of the *Statistical Cybermetrics Research Group* and a senior lecturer in mathematics and statistics at the *University of Wolverhampton*, UK. He has published 11 refereed journal articles about alternative indicators for research evaluation. Pardeep also has several years' experience in the actuarial profession, where he specialised in asset-liability modelling for defined benefit pension schemes.
https://orcid.org/0000-0002-3304-0469

p.sud@wlv.ac.uk

University of Wolverhampton, Statistical Cybermetrics Research Group
School of Mathematics and Computer Science
Wulfruna Street, Wolverhampton WV1 1LY, UK

Abstract

This article assesses whether there are gaps in *Wikipedia*'s coverage of academic information and whether there are non-obvious stylistic differences from academic journal articles that *Wikipedia* users and editors should be aware of. For this, it analyses terms in the titles of journal articles that are absent from all English *Wikipedia* page titles for each of 27 *Scopus* subject categories. The results show that English *Wikipedia* has lower coverage of issues of interest to non-English nations and there are gaps probably caused by a lack of willing subject specialist editors in some areas. There were also stylistic disciplinary differences in the results, with some fields using synonyms of "analysing" that were ignored in *Wikipedia*, and others using the present tense in titles to emphasise research outcomes. Since *Wikipedia* is broadly effective at covering academic research topics from all disciplines, it might be relied upon by non-specialists. Specialists should therefore check for coverage gaps within their areas for useful topics and librarians should caution users that important topics may be missing.

Keywords

Wikipedia; Research communication; Encyclopedia; Science communication.

Resumen

Este artículo evalúa si hay vacíos en la cobertura de la información académica de *Wikipedia* y si existen diferencias estilísticas no obvias entre los artículos de revistas académicas que los usuarios y editores de *Wikipedia* deben conocer. Para ello se analizan los términos en los títulos de artículos de revistas que están ausentes de todos los títulos de las páginas de *Wikipedia* en inglés para cada una de las 27 categorías temáticas de *Scopus*. Los resultados muestran que la *Wikipedia* en inglés tiene menor cobertura de los temas de interés para las naciones que no son de habla inglesa, y existen lagunas probablemente

Manuscript received on 01-09-2017
Accepted on 24-10-2017

causadas por la falta de editores especialistas dispuestos en algunas áreas. También se encontraron diferencias de estilo según las disciplinas, con algunos campos que utilizan sinónimos de "análisis" que fueron ignorados en *Wikipedia*, y otros que usan el tiempo presente en títulos para enfatizar los resultados de la investigación. Dado que *Wikipedia* es muy eficaz en la cobertura de temas de investigación académica de todas las disciplinas, puede ser utilizada por personas no especializadas. Por lo tanto, los especialistas deben verificar las lagunas de cobertura dentro de sus áreas para encontrar temas útiles y los bibliotecarios deben advertir a los usuarios que pueden faltar temas importantes.

Palabras clave

Wikipedia; Comunicación de investigación; Enciclopedia; Comunicación científica.

Thelwall, Mike; **Sud, Pardeep** (2018). "A comparison of title words for journal articles and *Wikipedia* pages: Coverage and stylistic differences?". *El profesional de la información*, v. 27, n. 1, pp. 49-64.

https://doi.org/10.3145/epi.2018.ene.05

1. Introduction

Wikipedia, ranked the 5th most popular website in August 2017 by *Alexa.com*, is a source of a wide variety of mostly accurate information.
http://www.alexa.com/siteinfo/wikipedia.org

Its knowledge is valuable not only for its easy accessibility but also because many people would seek less informative free web alternatives if it did not exist (**Fallis**, 2008). *Wikipedia* is widely used in education (e.g., **Henderson** *et al.*, 2015; **Lim**, 2009), by the public for health-related issues (e.g., **Thomas** *et al.*, 2013) and probably also for professional, recreational and other needs. For example, junior doctors may consult *Wikipedia* regularly (**Hughes** *et al.*, 2009) and digital archives may link to *Wikipedia* for contextual information (**Szajewski**, 2013). Although institutions and research funders finance open access journal articles to make academic knowledge available to all (**Lange**, 2016; **Pinfield**; **Salter**; **Bath**, 2016), scholarly topics on *Wikipedia* may well be consulted by a wider section of the population than read journal articles. It is therefore important to understand how *Wikipedia* covers academic information and assess the comprehensiveness of its coverage (e.g., **Rush**; **Tracy**, 2010). For example, *Cochrane* is working with *WikiProject Medicine* to ensure that, when possible, *Wikipedia* articles on medical topics are supported by state of the art evidence from *Cochrane* reviews (**Mathew** *et al.*, 2013).
https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine

> Although *Wikipedia* covers many academic research topics, it is not known whether it has substantial gaps in content

Although *Wikipedia* covers many academic research topics (**Mesgari** *et al.*, 2015), it is not known whether it has substantial gaps in content. If such gaps were found, then researchers and research funders may consider taking extra steps to ensure that steps are taken to remedy this issue.

Wikipedia articles on specialist research topics are presumably often written or edited by field specialists or postgraduate students that are casual editors of *Wikipedia* rather than experienced *Wikipedians*. It would also therefore be useful to gain insights into any aspect of the way in which *Wikipedia*'s style differs from that of journal articles. This could help casual editors to tailor their style when contributing to *Wikipedia* or when translating academic research.

The two goals of this article are to get insights into (but not comprehensively evaluate):

a) the extent to which *Wikipedia* covers academic research and

b) stylistic differences in presentation between *Wikipedia* and journal articles.

Both are addressed by comparing the words used in the titles of academic articles with the words used in *Wikipedia* page titles. Of course, there are likely to be major obvious differences between the two due to their different functions. Nevertheless, this approach can give *insights* into gaps in *Wikipedia* and stylistic difference between the two sources of information across many different fields. It was chosen as a practical way to make large scale comparisons between *Wikipedia* and academia, although many areas of scholarship are not covered. Article keywords could also have been analysed but these are less rich than title words and for some journals are restricted to controlled vocabularies, such as *MeSH*, that may be out of date or change the focus of the study to the controlled vocabulary itself. Since article keyword styles differ between journals (different controlled vocabularies, different controlled vocabulary versions, non-use of controlled vocabularies), any interdisciplinary analysis of these would be necessarily complex and would not be able to give fully comparable results between disciplines.

2. *Wikipedia*

An encyclopedia is,

> "A literary work containing extensive information on all branches of knowledge, usually arranged in alphabetical order",

or

> "An elaborate and exhaustive repertory of information on all the branches of some particular art or department of knowledge; esp. one arranged in alphabetical order" (*OED*, 2016).

Wikipedia is an example of the former kind because its co-

verage is not restricted by topic. Whilst earlier forms were conceived as comprehensive knowledge for education (**Lærke**, 2014), later encyclopaedias serve more of a reference function – hence perhaps the shift from subject-based groupings of topics to an alphabetical list of entries (**Loveland**, 2013).
https://www.britannica.com/topic/encyclopaedia

Wikipedia's model of user-edited content seems to have largely eclipsed previous encyclopedias due to its more comprehensive free coverage (**Gralla**, 2009) combined with a similar level of accuracy (**Giles**, 2005; **Mesgari** *et al.*, 2015; **Stankus**; **Spiegel**, 2010), albeit with less credibility (**Flanagin**; **Metzger**, 2011; **Kubiszewski**; **Noordewier**; **Costanza**, 2011; but see: **Gorichanaz**, 2016). Its coverage may be less accurate for topics that attract non-expert participation, such as those that are politically sensitive (**Wilson**; **Likens**, 2015). There is some evidence that editors can be casual about citing sources (**Luyt**, 2015). *Wikipedia* does not cover news, however.
https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

A consequence of the differences between *Wikipedia* and a traditional encyclopaedia is that its contents and coverage style are likely to differ. For example, in comparison to the UK-based *Encyclopedia Britannica*, *Wikipedia* articles about large businesses seem to express more sentiment, be longer and to cover ethical issues more (**Messner**; **DiStaso**, 2013).

Wikipedia's unpaid contributors tend to act for their own personal reasons rather than from an external imperative (**Yang**; **Lai**, 2010) and enjoy contributing (**Nov**, 2007). They are rarely motivated by a need for public recognition, but tend to believe that the activity is useful, that they are competent to edit, and that contributing is fair because they also use the information (**Cho**; **Chen**; **Chung**, 2010; **Lai**; **Yang**, 2014). Another motivation is to support personal development (**Xu**; **Li**, 2015). There do not seem to be any studies of the reasons why researchers contribute to *Wikipedia*, but the need to organise to ensure high quality coverage of academic-related topics is evident in initiatives like *WikiProject Medicine*.

> *Wikipedia* coverage may be less accurate for topics that attract non-expert participation, such as those that are politically sensitive

2.1. *Wikipedia* content

Wikipedia's size allows it to subsume the functions of a specialist encyclopedia. These have historically allowed a more detailed treatment of a single broad subject area of cultural (e.g., from the 1732 *Musikalisches Lexikon* to the modern "companions" to areas of literature from Oxford and Cambridge) or scientific interest (e.g., from the 1821 *Dictionary of Chemistry* to modern handbooks for areas of research produced by scholarly publishers). For instance, an encyclopedia for a field may be expected to summarise its important features (e.g., **Dick**, 2015), sometimes keeping this up to date by using digital formats (**Remy**, 2015). The restriction to a specialist area probably allows a more technical language to be used.

> One issue that *Wikipedia* may not be good at dealing with is the need to provide comprehensive information about topics from the perspective of all relevant types of user

One issue that *Wikipedia* may not be good at dealing with is the need to provide comprehensive information about topics from the perspective of all relevant types of user. For example, if *Wikipedia* is relied upon in education then gaps in coverage could cause problems to students that rely upon it (**Azer**, 2015). *Wikipedia* pages are also widely cited by patents, often using them to support knowledge claims (**Orduña-Malea**; **Thelwall**; **Kousha**, 2017) and so gaps in *Wikipedia* may translate to problems in patents. Whilst a casual editor may be likely to make accurate edits, they may not be well positioned to judge whether important information is missing. Medical information is particularly critical in this regard on *Wikipedia* because article inaccuracies or omissions may have serious health consequences (**Hasty** *et al.*, 2014; **Masukume** *et al.*, 2016). There are also cultural biases in the content of articles. Biographies of individuals vary between different language versions of *Wikipedia* (**Callahan**; **Herring**, 2011) and businesses are more extensively discussed in *Wikipedia* articles within their own languages (**Roessing**; **Einwiller**, 2016). The term "bias" here is used in a descriptive rather than pejorative sense.

Many previous studies have analysed the comprehensiveness or accuracy of *Wikipedia* for specific topics (**Mesgari** *et al.*, 2015) but there have been no recent empirical assessments of its relationship with academic knowledge with wide coverage. It does not have systematic and appropriate coverage of scholars (**Samoilenko**; **Yasseri**, 2014) and some broad subject areas were under-represented compared to books in 2006, with gaps in content compared to subject-specific encyclopedias (**Halavais**; **Lackaff**, 2008). Popular and more current topics also have longer articles (**Royal**; **Kapila**, 2009).

2.2. *Wikipedia* and academic journal articles

The most obvious difference between *Wikipedia* and an academic journal article is that *Wikipedia* does not allow pages to report original research.
https://en.wikipedia.org/wiki/Wikipedia:No_original_research

Wikipedia also attempts to summarise knowledge, and by extension to exclude unnecessary fine details, whereas only academic journal articles that are literature reviews have knowledge summarisation as a primary function.
https://en.wikipedia.org/wiki/Wikipedia:Scope
https://en.wikipedia.org/wiki/Wikipedia:Too_much_detail

An academic journal article may be expected to contain details that will not inform the general reader but would be important for other researchers, such as information that would allow an experiment to be reproduced. *Wikipedia* articles should be targeted at general readers and should

therefore ignore specialist terminology and assumed prior knowledge as far as possible, differing in this respect from journal articles.
https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

> Whilst a casual editor may be likely to make accurate edits, they may not be well positioned to judge whether important information is missing. Medical information is particularly critical in this regard

The *Encyclopedia Britannica* claims that an encyclopedia typically summarises published scholarship, which suggests a close relationship with academic research and would exclude most popular culture information.
https://www.britannica.com/topic/encyclopaedia

In contrast, *Wikipedia* grew based on anyone being able to add content, without the need to be an expert on the subject area and without the need for this content to be reviewed by a trusted person (**De-Laat**, 2012). This opens the door to non-academic content and popular culture (**Yam**, 2016). Nevertheless, at least one journal, *PLoS Computational Biology*, has explicit *Wikipedia*-friendly policies to help make high quality research-informed information more accessible. This includes publishing *Wikipedia*-friendly versions of articles (**Wodak** *et al.*, 2012). Other academics have made specific pleas for disseminating an area of scholarship via *Wikipedia* (**Signore**; **Serio**; **Santamaria**, 2014).

The content of *Wikipedia* has previously been compared to published academic research. Journals cited by English language *Wikipedia* articles in April 2007 were compared to those listed in the *Journal Citation Reports* (*JCR*) multidisciplinary set (**Nielsen**, 2007; see also: **Kousha**; **Thelwall**, 2017). From the 30,368 matches found, higher impact journals were more likely to be cited (e.g., *Nature*: 787; *Science*: 669) and medical and astronomy journals seemed to be overrepresented. There was also at least one active area within *Wikipedia* with many citations to relatively obscure journals, such as Australian Banksia species articles citing Australian botany journals, but this study did not report a systematic comparison of academic research topics and *Wikipedia* content. A study of 4721 journals in 26 fields confirmed that articles were more likely to be referenced if they were in higher impact journals but also found that open access journals were more likely to be cited (**Teplitskiy**; **Lu**; **Duede**, 2017).

A topic-based investigation of citations in *Wikipedia* articles focused on one of its high-profile areas, astronomy. Older research was less cited in *Wikipedia* than newer research (**Thelwall**, 2016), suggesting that shifts in the focus of academic research could lead to changes in the attention given to different topics. High profile issues, such as the planet status of Pluto, seemed to result in an increase in *Wikipedia* editor activity. A study of wind power pages in *Wikipedia* found that a quarter of their references cited academic publications but, less than 1% of academic wind power articles had been cited in wind power *Wikipedia* pages (**Serrano-López**; **Ingwersen**; **Sanz-Casado**, 2017). If typical, this suggests that the vast majority of individual academic papers are ignored by *Wikipedia*.

2.3. *Wikipedia* and academic article title styles

Wikipedia page titles can be either a name or a topic description, according to the official guidelines. They should be recognisable, natural, precise, and concise, as well as being consistent with the titles of similar articles. Disambiguation pages can be used for cases where different topics could have the same name. There are also some specific conventions for naming types of articles, such as books, people, organisations and events.
https://en.wikipedia.org/wiki/Wikipedia:Article_titles

Academic journal article titles have more flexibility than *Wikipedia* pages, although some journals and field norms may be prescriptive. They should ideally summarise the content of an article to help someone decide whether to read the abstract for more information (**Swales**, 1990). The most common formats are declarations of results or descriptions of the paper (**Jamali**; **Nikzad**, 2011). Titles may express the purpose or results of a study, its methods (**Méndez**; **Alcaraz**; **Salager-Meyer**, 2014; **Paiva**; **Lima**; **Paiva**, 2012) or the overall study design (**Ubriani**; **Smith**; **Katz**, 2007). A descriptive chemistry title, for example, may centre around variants of the phrase "an analysis of" (**Sano**; **Fujiwara**, 1993). Nevertheless, authors may be inefficient or ignore the function of article titles (**Hartley**, 2005) and can also find creative solutions by adopting or inventing alternative strategies to attract attention (**Hartley**, 2007). Titles can be questions, but these are rare in all disciplines (**Cook**; **Plourde**, 2016; **Méndez**; **Alcaraz**; **Salager-Meyer**, 2014) although they are increasing in frequency (**Ball**, 2009).

> *Wikipedia* grew based on anyone being able to add content, without the need to be an expert on the subject area and without the need for this content to be reviewed by a trusted person (De-Laat, 2012)

Acronyms can also be present in article titles (**Rostami**; **Mohammadpoorasl**; **Hajizadeh**, 2014) but these can also occur in *Wikipedia* redirection and disambiguation page titles even if they are rare in standard *Wikipedia* page titles. Present participles (verbs ending in -ing) can also be used to emphasise the importance of results (**Wang**; **Bai**, 2007) and these seem likely to be absent from *Wikipedia* titles since they do not need to perform this function. The same is true for past participles (**Wang**; **Bai**, 2007).

There are disciplinary differences in the constructions of article titles. For example, complete sentences are more common in some disciplines than others and the use of compound titles, such as with a colon in the middle, is particularly common in the social sciences (**Soler**, 2011). Compound titles may start with a general theme and then finish with a particularising aspect (**Hartley**, 2007) and terms playing the latter role may be too specific for a *Wikipedia* page. Older fields may also tend to have longer article titles (e.g., **White**; **Hernandez**, 1991) as articles become more specialised to fill gaps left by previous research.

3. Research hypotheses

The following hypotheses drive the research. They are partly informed from the literature review above and partly by initial explorations of the data. The primary goal of the paper is to seek evidence of coverage gaps, which relate primarily to hypotheses 4 and 5, and the secondary goal is to seek evidence of stylistic differences, which relate to hypotheses 1, 2, and 3. The secondary goal also aids the primary goal by separating term differences that are not due to coverage gaps.

Hypothesis 1 (complexity)

Terms that are common in *Scopus* article titles for a subject but absent from *Wikipedia* titles are often *complex* in format or in the ideas represented. This seems likely because encyclopedias summarise knowledge for the non-expert (**Béjoint**, 2000) and therefore seem likely to use simpler language than that of journal articles, which target subject experts or specialists. Moreover, one form of complexity is specificity and so articles may be about topics that are too specific to merit their own *Wikipedia* page.

Hypothesis 2 (research process descriptions)

Terms that are common in *Scopus* article titles for a subject but absent from *Wikipedia* titles include terminology that describes the research process (e.g., **Méndez**; **Alcaraz**; **Salager-Meyer**, 2014; **Paiva**; **Lima**; **Paiva**, 2012; **Ubriani**; **Smith**; **Katz**, 2007). This seems likely because the research process is important in academia but research outcomes are more relevant to the summarising functions of an encyclopedia. Nevertheless, the research process itself can be an object of study and hence may be described in separate *Wikipedia* pages.

Hypothesis 3 (stylistic and structural differences)

Some terms that are common in *Scopus* article titles for a subject but absent from *Wikipedia* titles are the result of stylistic differences in the way in which academic titles are formed (e.g., **Jamali**; **Nikzad**, 2011) in comparison to *Wikipedia* page titles, or structural differences in the types of entities that are given their own pages.

Hypothesis 4 (culture bias)

Terms that are common in *Scopus* article titles for a subject but absent from *Wikipedia* titles include terminology that is *geographically or culturally* tied to languages other than English (see: **Roessing**; **Einwiller**, 2016). This will be most evident in the arts and humanities (even though scholars tend to write in their own language, humanities deal with many location-specific phenomena). It will also occur in sciences that deal with location-specific phenomena, such as geography, social sciences and applied sciences. It will not be evident in pure sciences.

Hypothesis 5 (editor gaps)

There are gaps in the coverage of *Wikipedia* that are reflected in some important subject terms being absent from *Wikipedia* page titles. These seem likely to occur because the voluntary nature of *Wikipedia* and the need for specialists to edit academic topics means that areas lacking specialist volunteers may lack content or may omit content that is relevant to some users (**Azer**, 2015).

4. Methods

The research hypotheses were addressed by identifying and manually examining a large set of words that frequently occur in academic journal article titles but rarely in *Wikipedia* page titles. To start, different *Scopus* subject categories were chosen to represent a wide range of academic fields. *Scopus* categories were chosen as a transparent method of collecting together large and reasonably coherent collections of subject-based journal articles. Although some journals are interdisciplinary or occasionally publish articles from out of their core scope, this is unlikely to affect the primary research method, as described below. *Scopus* was selected in preference to the *Web of Science* for its finer grained categories. The seventh field was selected from each broad *Scopus* category, replacing this with the next alternative (counting in cycles) for categories with less than 7 subjects. This produced a set of 27 subject categories from diverse fields from the arts and humanities, social sciences, engineering, formal sciences, physical sciences, life sciences, medicine and health science (Table 1).

The titles of all journal articles published between 1996 and 2015 (20 years) were downloaded from *Scopus*. The year 1996 was the starting date because the coverage of *Scopus* expanded in this year. English language articles dominate *Scopus* and non-English articles were *not* excluded because some topics of international interest may be primarily discussed in other languages (e.g., Spanish, German, French, Chinese). This might have resulted in non-English function words occurring in the stylistic analyses but this did not occur due to the dominance of English in *Scopus* titles and abstracts. The final year was 2015 to ensure that there had been sufficient time (almost 1 year) for all articles to have been added to *Wikipedia*. There was a system limit of 10,000 articles per year and in large subject areas and more recent years, this resulted in incomplete coverage. In such cases, only the first and last 5000 articles from the subject and year were obtained. This should not have a major impact on the results or bias them, given the large total numbers involved. For each of the 27 subjects, a complete list was built of all words in all titles, together with their frequency. The free software *Webometric Analyst* was used for this [*Tab-sep* menu; *Count frequency of words in text or column n* (e.g., *Wikipedia* titles) menu item].

The titles of all *Wikipedia* articles were extracted from a data dump of *Wikipedia* from December 12, 2016 of a “List of all page titles in the main namespace”.
https://dumps.wikimedia.org/enwiki

This includes standard page titles and disambiguation pages, which are the main pages that typical visitors will see.

It excludes meta/background pages, such as those for editor discussions and image files.
https://meta.wikimedia.org/wiki/Help:Namespace

All words were extracted from all these *Wikipedia* page titles and used to form a word frequency list in the same way as for the *Scopus* article titles. This resulted in 2,417,043 different words with frequencies up to 1,064,797 (*of*).

No word stemming was applied to either term list because the focus of the study is on academic language, which by its nature involves rare words for which the normal rules of grammar may not apply. Moreover, given that the comparisons span very large word lists, even rare systematic errors can impact on the word frequency comparison results. Hyphens in the middle of words or at their end were retained within the words. Hyphens are an important part of names in some fields (**Burke**, 2008; **Frey-Klett** *et al.*, 2011; **Gill** *et al.*, 2009).

For each of the 27 subjects, each term in the term frequency list was checked against the *Wikipedia* list and the presence or absence of a match was recorded. The 25 most frequent words in the subject that were absent from the *Wikipedia* data were then saved for manual analysis, resulting in 27 sets of 25 terms (675 in total). These are therefore all words that are frequent in article titles for a specific *Scopus* subject category but completely absent from *Wikipedia* page titles. For example, “mucoadhesive” occurred 474 times in *Scopus* journal articles from the category Pharmaceutical Science but not in any *Wikipedia* page titles.

The primary analysis was an investigation into the linguistic properties of the 675 terms extracted. Although the goal of this paper is to assess academic knowledge gaps, initial testing indicated strong linguistic commonalities that were relevant to the research hypotheses. For the main manual analysis, the 27 sets of 25 terms were manually examined and their

Table 1. The 27 *Scopus* subject categories examined and the number of articles extracted from them for the years 1996-2015 (as of December 2016). The Table includes equivalent information for *Wikipedia* (bottom row). The top term is the most common term in article titles that is not in a *Wikipedia* page title. See Table 2 for the codes used.

Subject area	Articles	Title words	Mean title words	SD title words	Top term
App Microbiol & Biotech.	171,555	2,266,862	13.2	4.63	Enhances (PT)
Atomic & Mol Phys & Optics	199,990	2,089,197	10.4	4.18	All-optical (Hy)
Cell Biology	199,308	2,600,433	13.0	4.76	Regulates (PT)
Comp. Vis. & Pattern Recog.	70,807	641,427	9.1	3.58	Wavelet-based (Hy)
Control & Systems Eng.	177,184	1,661,292	9.4	3.52	Observer-based (Hy)
Dental Assisting	89	700	7.9	3.45	Perimplantitis (Su)
Dermatology	140,309	1,500,975	10.7	5.39	HIV-infected (Hy)
Discrete Math & Combin.	51,666	394,323	7.6	3.51	Subgraphs (Pl)
Emergency Nursing	16,082	147,366	9.2	5.46	Out-of-hospital (Hy)
Endocrine & Autonomic	14,195	188,404	13.3	4.69	Modulates (PT)
Finance	90,791	820,166	9.0	3.95	Spillovers (Co)
Fluid Flow & Transfer Proc	85,996	961,890	11.2	4.41	Impinging (PP)
Forestry	96,934	1,257,994	13.0	4.51	Provenances (Pl)
Fuel Technology	148,800	1,646,493	11.1	4.31	Non-premixed (Hy)
Geology	151,359	1,979,169	13.1	4.94	Ore-forming (Hy)
Health, Toxicology & Mut	117,852	1,583,038	13.4	4.75	Subchronic (Pr)
History & Philosophy of Sci	29,996	291,535	9.7	4.96	Preservice (Pr)
Human Factors & Ergonom	21,378	233,041	10.9	4.36	Quantifying (PP)
Medical Laboratory Tech	17,600	213,768	12.1	5.40	Clinicopathologic (Co)
Org Behav & Hum Res Man	56,732	567,651	10.0	4.24	Moderating (PP)
Pharmaceutical Science	171,798	1,972,024	11.5	4.82	Dispersions (Pl)
Polymers & Plastics	199,852	2,165,758	10.8	4.21	Hyperbranched (Pr)
Small Animals	9,204	105,265	11.4	5.51	Frozen-thawed (Hy)
Social Psychology	94,661	1,055,939	11.2	4.47	Moderating (PP)
Spectroscopy	172,971	2,205,312	12.8	4.93	Preconcentration (Pr)
Stats, Prob & Uncertainty	65,796	576,474	8.8	3.60	Change-point (Hy)
Transplantation	77,971	1,006,897	12.9	4.89	Undergoing (PP)
***Scopus* 27**	2,650,876	30,133,393	11.0	4.50	
Wikipedia	12,922,668	35,370,976	2.7	1.86	

key linguistic features were extracted and compared. This was achieved by the first author reading the lists, identifying factors in common that were relevant to the research hypotheses and then classifying all the terms for these factors. These categories are described in the results section. This process was repeated until no further common factors could be identified. A native English-speaking coder independent of the project and with a degree in English independently re-classified 100 of the texts using the guidelines in the table caption and examples from the first coder, achieving a Cohen's kappa value of 0.836 for inter-coder consistency (**Cohen**, 1960) for 87% agreement. Cohen's kappa assesses the agreement between two coders on the same categorisation task, factoring out chance rates of agreement. Values above 0.8 could be described as "almost perfect agreement" (e.g., **Landis**; **Koch**, 1977). Even accounting for chance rates of agreement, this indicates that the two coders almost always agreed. This confirms that the classification scheme is transparent and straightforward, with only a small element of subjectivity.

Table 2. Characteristics of the 25 terms that are most frequent in article titles but absent from *Wikipedia* titles. Cu=related to non-English-speaking cultures; PT=present tense verbs; PP=present participle verbs (-ing, but not gerund or a verbal noun); Hy=hyphenated terms; Co=nominal compound or acronym, Po=possessives; Pl=plurals; Pr=term with prefix; Su=term with suffix; Ot=others. Categories toward the left have priority over categories toward the right.

Subject area	Cu	PT	PP	Hy	Co	Po	Pl	Pr	Su	Ot
App. Microb. & Biotech.	0	8	0	5	4	0	3	2	3	0
At. & Mol. Physics, Optics	0	0	0	14	3	0	3	2	3	0
Cell Biology	0	18	0	4	3	0	0	0	0	0
Comp. Vis. & Pattern Rec.	0	0	2	17	2	0	0	2	2	0
Control & Systems Eng.	0	0	0	20	1	0	2	1	1	0
Dental Assisting	3	0	1	7	2	3	2	3	3	1
Dermatology	0	7	3	7	4	0	0	3	1	0
Discr. Math. & Comb.	0	0	2	10	0	0	11	1	0	1
Emergency Nursing	1	1	1	12	2	0	1	3	3	1
Endocrine & Auto. Sys.	0	14	1	8	1	0	0	0	1	0
Finance	1	1	2	9	0	3	6	2	1	0
Fluid Flow & Trans. Proc.	0	0	0	15	3	0	2	3	2	0
Forestry	0	0	2	13	4	0	3	2	1	0
Fuel Technology	0	0	0	12	6	0	2	4	1	0
Geology	3	0	1	9	4	0	3	3	2	0
Health, Tox. & Mut.	0	6	1	9	4	0	2	3	0	0
History & Phil. of Sci.	0	0	7	4	1	3	2	4	3	1
Human Factors & Erg.	0	0	3	9	6	0	2	2	3	0
Medical Lab. Tech.	0	1	2	6	8	0	0	8	0	0
Org. Behav, & HRM	1	1	2	14	1	6	0	0	0	0
Pharmaceutical Science	0	5	0	7	5	0	4	0	3	1
Polymers & Plastics	0	0	0	8	2	0	4	7	4	0
Small Animals	0	1	1	9	5	0	1	4	4	0
Social Psychology	0	1	2	9	1	6	1	1	4	0
Spectroscopy	0	0	0	10	9	0	2	3	1	0
Stats, Prob. & Uncert.	0	0	3	14	1	0	3	2	1	1
Transplantation	0	6	1	9	4	0	0	5	0	0
Overall average	0.3	2.6	1.4	10.0	3.2	0.8	2.2	2.6	1.7	0.2
Overall percentage	1%	10%	5%	40%	13%	3%	9%	10%	7%	1%

For a secondary cross-check, when possible, key properties identified from the main manual analysis were followed up by comparing relevant properties for the set of terms from the subject that were not found in *Wikipedia* with the corresponding set of terms from the subject that were also in *Wikipedia*. This comparison allows a check of whether any differences identified also occur outside of the top 25 terms, and that they are genuinely the result of differences with *Wikipedia*.

5. Results

A total of 2,650,876 articles were extracted from the 27 subject categories, but the Dental Assisting category is too small to give useful data (Table 1). The full data for this article is available online:
https://figshare.com/s/1e8774053297ddd4257e

Journal article titles were typically longer (11.0 words) than *Wikipedia* page titles (2.7 words). The length difference is probably due to the many *Wikipedia* pages that are nouns or noun phrases, such as names of people (e.g., Tom Tureen), animals or plants (e.g., Yucatan Jay), objects (e.g., SS King James), places (e.g., Senaki District), concepts (e.g., Religious Satanism) and cultural products (e.g., Dangerous Girls).

Each of the top 25 terms in a *Scopus* journal article title but absent from all *Wikipedia* page titles were placed in one of ten categories according to the likely reason for their absence: Cu=of primary interest in non-English-speaking cultures; PT=present tense verbs; PP=present participle verbs (-ing, but not gerund or a verbal noun); Hy=hyphenated terms; Co=nominal compound term (i.e., with two separate major word parts, such as vibrotactile but not microring because micro- is a suffix) or acronym; Po=possessives; Pl=plurals; Pr=term with prefix (e.g., micro-, di-, nano-); Su=term with suffix (e.g., -ive, -ed, -ological); Ot=others. Categories toward the left have priority over categories toward the right, with each word being allocated to the leftmost category that it fits. This order was chosen to reveal the most important patterns in the data.

The key attributes of the categories are described below, matched with the hypotheses that they are most relevant for, although some relate to multiple hypotheses.

5.1. Hypothesis 1: Complexity

Compound and hyphenated terms are indicators of complexity at the word level because they bind together different concepts. The same is true for possessives since they connect to another concept in the title.

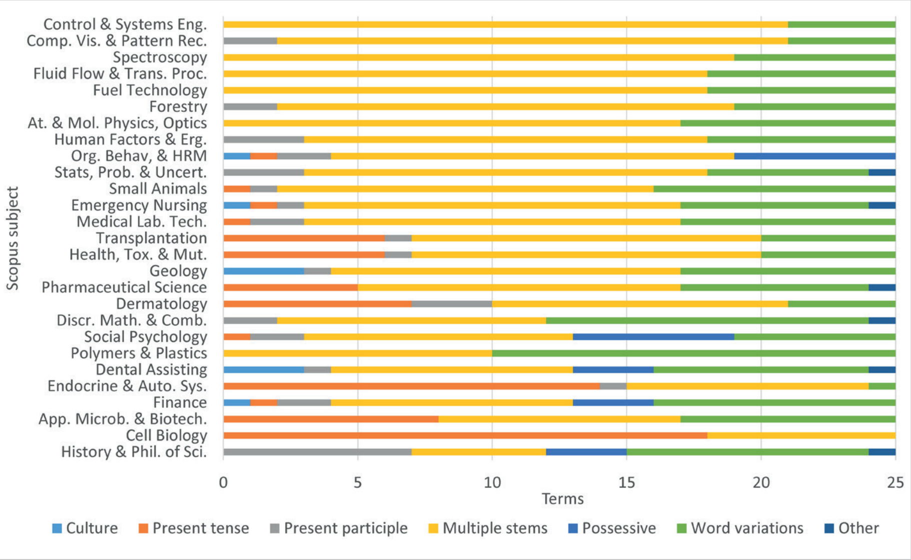


Figure 1. Major categories of the 25 terms that are most frequent in article titles but absent from *Wikipedia* page titles. The categories and data are the same as in Table 2 except: hyphenated terms and nominal compounds or acronyms are merged into "Multiple stems"; plurals and terms with prefixes or suffixes are merged into "Word variations". Subjects are arranged in decreasing order of the size of the Multiple stems category.

Compound and hyphenated terms form a majority, on average, of the 25 terms for each subject that are most frequent in article titles but absent from *Wikipedia* page titles (Figure 1; Table 2). These take the form of hyphenated words (40% overall) or non-hyphenated compound words (13% overall) both of which are complex in the sense of merging together two distinct substantial terms. This type of term was particularly common in Control and Systems Engineering (84%), including five '-based' terms and two '-feedback' and '-dependent' terms. In this area, hyphenated terms seem to be primarily useful as adjectives to describe the distinctive features of a control system. Compound and hyphenated terms were least common in History and Philosophy of Science (20%).

In three subjects, several of the most frequent hyphenated terms had a common term. Seven Computer Vision and Pattern Recognition terms had the same second part (wavelet-based, feature-based, gradient-based, HMM-based, appearance-based, block-based, SVM-based) for descriptions of the key features of algorithms. In Endocrine and Autonomic Systems three terms had the same second part (anxiety-like, depressive-like, depression-like), for describing symptoms. In Social Psychology there were three terms with a common first part (self-perceptions, self-other, self-reports) and two other terms included a specific role (parent-adolescent, mother-child). This aligns with the social psychology focus on individuals in a social context, but it is not clear why these terms would not be in a *Wikipedia* page title.

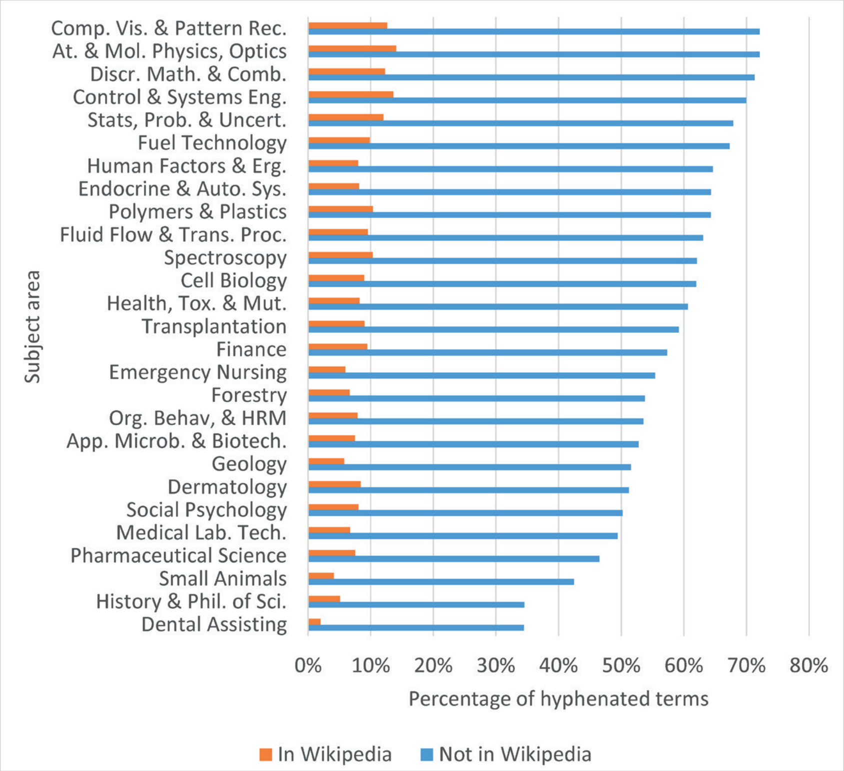


Figure 2. The percentage of terms in the titles of *Scopus* articles that are hyphenated, broken down by whether they are also in *Wikipedia*.

Possessives were rare overall (3%: Table 2) but were the most common in the two disciplines with a focus on human behaviour in groups. Perhaps unsurprisingly, the Organizational Behavior and Human Resource Management possessives related to organisational groups (firms', auditors', followers', subordinates', organizations', managers') whereas the Social Psychology possessives were more general or more to non-work contexts (couples', preschoolers', therapists', adolescents', individuals', jurors') but again this does not explain their absence from *Wikipedia*. Possessives function to bind words together and so are an additional complexity flag. Possessives played a different role in two other subjects. For History and Philosophy of Science they were exclusively used to refer to academic theories (Carnap's, Vygotsky's, Bourdieu's). For the tiny Dental Assisting subject, the terms were mixed (therapy's, hygienist's, hands').

Overall, there is strong evidence of complexity in the top 25 terms, with a majority being explicitly the combination of two or more words. For the largest category, hyphenated terms, those that are also in *Wikipedia* are rare compared to those that are not for the full data sets (Figure 2). They also form a majority of terms absent from *Wikipedia* page titles in most subject categories. Thus, hyphenation is a universal major difference between journal article titles and *Wikipedia* page titles.

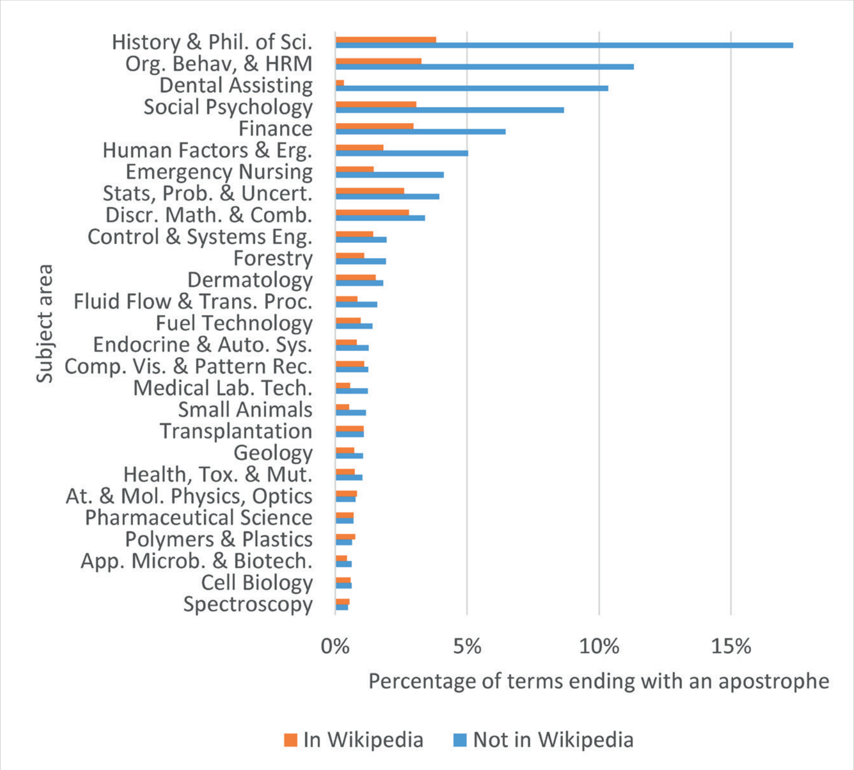


Figure 3. The percentage of terms in the titles of *Scopus* articles that end in an apostrophe or 's, normally indicating a possessive, broken down by whether they are also in *Wikipedia*.

Possessives are relatively rare and in the full data sets are usually, but not always, not found in *Wikipedia* page titles (Figure 3). There are huge disciplinary differences in the prevalence of possessives. In the subject areas for which they are most common, possessives are mostly absent from *Wikipedia* page titles (as far down as Transplantation in Figure 3). There is a general tendency for people focused disciplines to include more possessives, although this is not true for the two topics that focus on anatomical parts of people: Dermatology and Transplantation. Finance fits this trend through its focus on actors in the financial sector (e.g., analysts', firms', auditors', management's, SEC's). As in the research process discussion below, for some categories possessives could be the names of academics, which effectively form a compound term with the entity that they are associated with. There were also mathematics and statistics terms that are grammatically incorrect in standard English usage since they should not be used to pluralise acronyms (e.g., PDE's, as in "Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's") and so not all the terms classed as possessives in this category are correctly classified (en.oxforddictionaries.com/punctuation/apostrophe). Overall, however, the possessive analysis supports the hypothesis of increased complexity for all titles outside of *Wikipedia*.

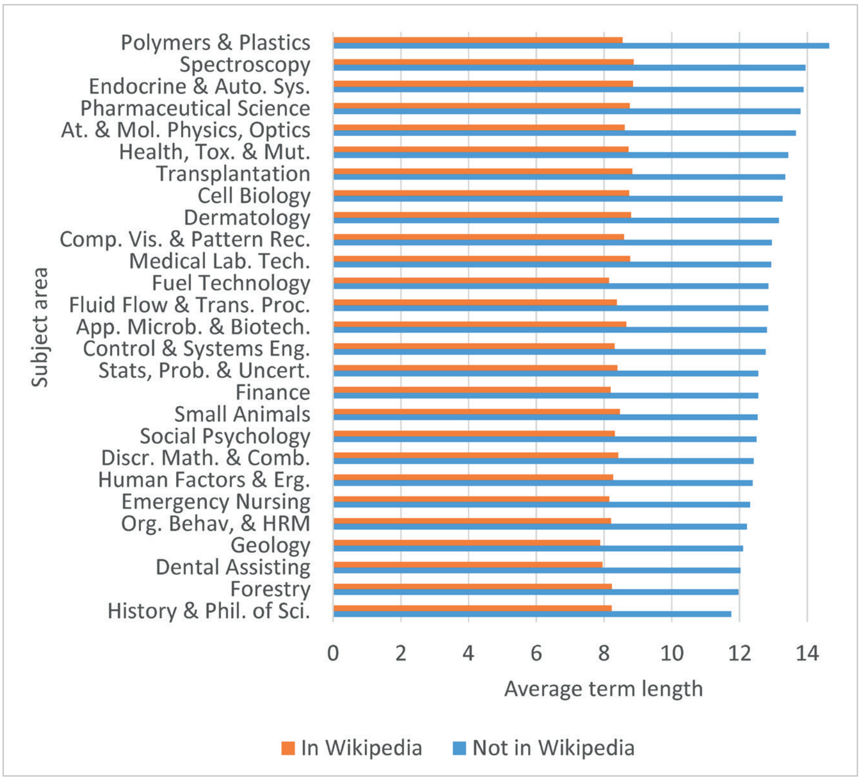


Figure 4. The average length of terms in the titles of *Scopus* articles broken down by whether they are also in *Wikipedia*.

The simplest complexity comparison is to compare the length of words inside and outside of *Wikipedia* page titles in the full data set. In all cases, longer terms in *Scopus* titles tend not to be in *Wikipedia* page titles (Figure 4). Even allowing for the biasing effect of short common terms (e.g., it, the), the large numbers involved and the substantial differences support the complexity hypothesis. However, this may be largely a consequence of the general absence of hyphenated terms in *Wikipedia* page titles reported earlier, as hyphenation will necessarily result in increased average term length.

5.2. Hypothesis 2: Research process descriptions

Present participles had the strongest association with the research process and there is also an association for some possessives.

The **present participle** is relatively rare (5%: Table 2) but common in only one large subject. In History and Philosophy of Science (reconsidering, moderating, disentangling, historicizing, re-thinking, cointegrating, untangling), these present participles are used to describe how the author is approaching the topic investigated. These terms therefore indicate the research process, albeit in a very general way.

Some **possessives** (see above for the History and Philosophy of Science examples) also refer to aspects of the research process in the form of the name of the researcher. In the full data set this was evident in three areas. For Statistics, Probability and Uncertainty, possessives refer to researchers' methods (e.g., Kak's [Three-Stage Quantum Protocol], Zadeh's [fuzzy logic], Simes' [method]) or application areas (e.g., insurer's, firms'). For Discrete Mathematics and Combinatorics, possessives tend to refer to mathematicians' theories or unsolved problems (e.g., Heilbronn's [triangle problem], Gallai's [theorems], Thomassen's [conjecture], Hajos' [theorem]), analogously to the History and Philosophy of Science use of possessives for theories, as discussed above. Although these possessives can also indicate complexity, they may also serve to name something in the absence of a more logical invented term (abstract maths may not have many real-world referents to derive a name from) or to confer authority by naming an important researcher. The latter case relates indirectly to the research process.

Some **suffix** terms also denoted the research process in Polymers and Plastics, to some extent, by describing the process that had been applied to a molecule: compatibilizer, compatibilized, and plasticized.

> Journal article title terms may be absent from *Wikipedia* page titles due to stylistic differences although the underlying concepts are present in a different form

Overall, there is strong evidence to support the research process description hypothesis as important in one area, History and Philosophy of Science, and weaker evidence for Statistics, Probability and Uncertainty, for Discrete Mathematics and Combinatorics, and for Polymers and Plastics.

5.3. Hypothesis 3: Stylistic and structural differences

Journal article title terms may be absent from *Wikipedia* page titles due to stylistic differences although the underlying concepts are present in a different form. The most obvious way in which this can occur is if one uses a singular form of a noun whereas the other uses the plural form. It can also occur if a term is not present in a *Wikipedia* page title but the stem term without any prefix or suffix is (see "Word variations" in Figure 1).

Plurals were a substantial minority overall (11%: Table 2), except in Discrete Mathematics and Combinatorics (subgraphs, matchings, labelings, colourings, transversals, nonlinearities, subsequences, -graphs, edge-colorings, hyperovals, labellings, -factors, k-trees), where they were in most cases probably used as a generic term (e.g., "labellings for subgraphs" implies that a general method for labelling all subgraphs will be described). It may be primarily a stylistic issue that *Wikipedia* might use the singular form to denote generalisation in mathematics instead. A similar issue seems to occur in the vastly different field of Organizational Behavior and Human Resource Management (spillovers, adapt-abilities).

Terms with a **prefix** or **suffix** were also a substantial minority overall (17% combined, Table 2). They were particularly common in Polymers and Plastics, including the three related prefix adjectives (diblock, triblock, and multiblock) and three related suffix adjectives, (compatibilizer, compatibilized, plasticized), as already mentioned.

The **present tense** is a large minority (10%: Table 2) that is common in two subjects and absent from most of the rest (14). In both Cell Biology (regulates, inhibits, enhances, modulates, interacts, stimulates, activates, suppresses, contributes, prevents, attenuates, impairs, encodes, facilitates, determines, confers, disrupts, potentiates) and Endocrine and Autonomic Systems (modulates, enhances, attenuates, regulates, inhibits, impairs, prevents, stimulates, suppresses, facilitates, contributes, modifies, potentiates, disrupts) present tense verbs are used to describe the actions of the entity investigated. Presumably if *Wikipedia* covers the topic of these articles then it would form part of an article on the entity rather than having a separate article on one of its actions. Thus, these terms probably reflect a focus on objects rather than actions in *Wikipedia* (nouns rather than verbs), a structural difference.

There are also some individual style differences in the use of **hyphens and compound words**. For example, although palmprint (Computer Vision and Pattern Recognition title frequency: 57) occurs in no *Wikipedia* page titles, there is a *Wikipedia* Palm print page.
https://en.wikipedia.org/wiki/Palm_print

Similarly, for flowshop (Statistics, Probability and Uncertainty frequency: 46) with a *Wikipedia* Flow shop scheduling page.
https://en.wikipedia.org/wiki/Flow_shop_scheduling

Acronyms and short forms of words may also be acceptable in journal article titles for an audience that would be familiar with them but not be present in *Wikipedia* pages. For example, Re-Os (Geology frequency 235) has an entire *Wikipedia* page called Rhenium-osmium dating.
https://en.wikipedia.org/wiki/Rhenium-osmium_dating

Another example is surface electromyography (SEMG), which is described in a section in the *Wikipedia* electromyography (EMG) page.
https://en.wikipedia.org/wiki/Electromyography

5.4. Hypothesis 4: Culture bias

Possessives could be markers of cultural bias if they referred to people that were of less interest in the English-speaking world. This only occurred in the History and Philosophy of Science and only for three people (Carnap's, Vygotsky's, Bourdieu's). None are native English speakers but all are widely known in the USA (Carnap moved there, and the others are internationally famous) and have large *Wikipedia* pages, so this is not strong evidence of cultural bias. In forestry, two terms refer to issues that are more important in warmer climates, *silvopastoral* (combining cattle and trees) and *postfire* (dealing with the aftereffects of devastating fires) but both are issues in the USA and discussed in this context in the literature.

In Geology, three terms were specific to China. Shahejie (in 115 journal article titles but no *Wikipedia* page titles) is a town and geological feature in China (e.g., "Diagenetic history and diagenetic stages prediction of Shahejie formation in the Qikou Sag"). For context, the word Shahejie occurs inside some English *Wikipedia* articles about Chinese railways and the only *Wikipedia* page with a title containing the term is Swedish, although Chinese *Wikipedia* contains the original term 沙河街, including a page dedicated to it.
https://sv.wikipedia.org/wiki/Shahejie_(h%C3%A4radshuvudort_i_Kina,_Jiangxi_Sheng,_lat_29,61,_long_115,89
https://zh.wikipedia.org/wiki/沙河街镇

Tazhong (110) is an oil bearing area of China (e.g., "Oil and gas accumulations in the Ordovician carbonates in the Tazhong Uplift of Tarim Basin, west China"). Xujiahe (109) is a geological feature in China (e.g., "Analysis on provenance-supply system of Upper Triassic Xujiahe Formation, Sichuan basin"). This is the clearest evidence of cultural bias. In the small Dental Assisting category, the three Italian words in the top 25 are not relevant (frequency 1 each). In Emergency Nursing, Xuebijing (15) is a traditional Chinese medicine and all articles about it are written by Chinese authors. The term is not mentioned in a *Wikipedia* page from any language. The original term血必净 does not seem to be present in Chinese *Wikipedia* (this was double-checked by a native Chinese speaker) although it has a dedicated page in another Chinese online encyclopaedia.
https://wapbaike.baidu.com/item/血必净

No evidence of cultural bias was found in the other subjects.

> There is evidence of cultural bias, albeit only against China and for Geology and Emergency Nursing

In conclusion, there is evidence of cultural bias, albeit only against China and for Geology and Emergency Nursing. The appearance of Chinese terms is probably due to a combination of the large size of the country and the indexing of some Chinese journals in *Scopus*, although the three terms found also occurred in international journals.

5.5. Hypothesis 5: Editor gaps

The three Chinese terms found above also serve as evidence of editor gaps. Presumably there were no experts on Shahejie, Tazhong or Xujiahe that were willing to write articles on them in the English *Wikipedia*. Given the existence of large national petrochemical companies, such as the China Petrochemical Corporation, it would make sense for the experts on these to be mainly resident Chinese.

> The results confirm that there are gaps in the coverage of academic topics by *Wikipedia*

To search for specific editor gaps, the 675 terms were manually examined for terms, and particularly singular nouns, that might refer to topic areas absent from *Wikipedia*.

- Pharmaceutical Science: The brand name Eudragit (subject category frequency: 183) does not occur in any English *Wikipedia* page titles and only seems to occur in four English *Wikipedia* pages as a very minor mention in each case.
  https://en.wikipedia.org/wiki/Ethyl_acrylate
- Human Factors and Ergonomics: The concept of macroergonomics (14; macroergonomic: 7) is largely missing from *Wikipedia*, although it gets four passing mentions in the main *Wikipedia* article, two in method names as part of a list, and two in sentences starting with, "As applied to macroergonomics…". This is clear evidence of a research concept having no meaningful content in *Wikipedia*, although relatively low term frequencies are involved.
  https://en.wikipedia.org/wiki/Human_factors_and_ergonomics
- Small Animals: preantral (29) occurs only as two minor mentions in the page on folliculogenesis ("in contrast to so a called preantral follicle that still lacks an antrum") and is not mentioned in the page on antral follicles. This term is therefore implicitly defined but not discussed.
  https://en.wikipedia.org/wiki/Folliculogenesis
  https://en.wikipedia.org/wiki/Antral_follicle
- Emergency Nursing: Fireground (33) is defined in *Wikipedia* but not extensively discussed. In academic articles, it seems to be a background term rather than a topic of discussion, however (e.g., "Establishing adequate fall protection on the fireground"). The Penehyclidine (11) [Hydrochloride] drug name is not in any *Wikipedia* page from any language version.

The above results confirm that there are gaps in the coverage of academic topics by *Wikipedia*.

It is possible that some areas have wider gaps in knowledge than others and a simple way to seek evidence of this is to assess whether any disciplines have a particularly high percentage of terms that are not in *Wikipedia* page titles. This is a very crude test, however, since term mismatches can be the result of many factors, as the above discussions show. The percentage of terms likely to be absent from *Wi-*

kipedia is also likely to be statistically related to the size of a field and so points that are the highest above the trend line in Figure 5 are the most likely to have gaps. From this evidence, Discrete Mathematics and Combinatorics either has the most missing knowledge or other systematic factors, as described above. Pure mathematics might be the general area that is least accessible to a non-expert because of its hierarchical nature and extreme abstraction and so it would make sense if this subject was the least well represented in *Wikipedia*.

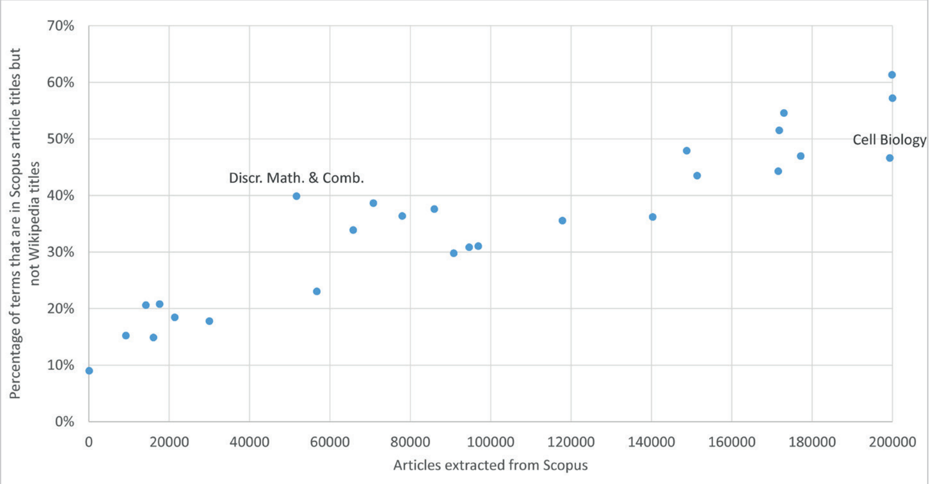


Figure 5. The percentage of terms that are in *Scopus* article titles but not in *Wikipedia* article titles against subject area size. The highest positive and negative outliers are named (identified from the residuals after linear regression).

6. Limitations and discussion

The results are limited in several respects in addition to those discussed in the methods section. Not all subject areas were analysed and the results may be different for those that were not chosen. The social sciences and humanities are not well represented in the fields covered and so important patterns in these areas have may been overlooked. The results are likely to be substantially different in a non-English *Wikipedia* and further research is needed to investigate this. Most importantly, the word frequency approach used here is indirect for the first four hypotheses and does not provide definitive evidence for them. Any failure to find evidence in support of a hypothesis is not evidence that the hypothesis is false for any subject. Moreover, since only the top 25 terms were examined for the primary analysis, the importance of one category of word for a subject would push out words from other categories, hiding the evidence that they provide. For the manual analysis, the choice of 25 terms rather than any other number is arbitrary and may have affected the results, as may the decision not to use word stemming.

> Article titles tend to be more complex than *Wikipedia* page titles across all subjects analysed

The results tend to confirm the hypotheses, at least partially, in all cases. In terms of the words contained in them, article titles tend to be more complex than *Wikipedia* page titles across all subjects analysed (H1), with longer terms and more hyphenated terms and, at least in the top 25, more compound words. This is consistent with *Wikipedia* carrying out the role of synthesising knowledge and simplifying and summarising it for a lay audience.

The strongest evidence of research process descriptions being absent from *Wikipedia* pages was in History and Philosophy of Science (H2), where the missing terms described general analytical approaches, but there was also some evidence for three other areas. Presumably, the most common academic methods have their own *Wikipedia* pages and these are exceptions.

There was clear evidence of systematic stylistic or structural differences in four subject areas (H3), but not all. These included the use of plurals to signify abstraction in Discrete Mathematics and Combinatorics and Organizational Behavior and Human Resource Management, and the present tense to describe actions in both Cell Biology and Endocrine and Autonomic Systems that would be described within *Wikipedia* pages but would not have their own pages. There seem likely to be additional stylistic differences in all areas despite the failure to find more evidence.

> There was clear evidence of systematic stylistic or structural differences in four subject areas, but not all

Although cultural biases may be present in other areas, they were only obvious in two, Geology and Emergency Medicine, and in both cases topics of interest mainly within China were absent from *Wikipedia* (H4). There may well have been subtler cultural biases that were not identified, however, such as interest in chemicals because of their applications in local industries within a specific country. In this context and due to the difficulty in checking the background to the use of all terms, only areas that mention academics' names, geographic concepts, or localised professional practices would have a reasonable chance of producing obvious evidence of cultural biases.

Non-trivial coverage gaps were found in three subjects (H5) in terms of concepts that were not well covered in *Wikipedia* despite appearing frequently in academic journal titles (ignoring the minor Emergency Medicine example). These are in addition to the cultural bias gaps that are described above for two further areas, and the research description gaps found for a sixth. There were probably gaps in all areas but these results provide evidence that gaps do occur in *Wikipedia*.

7. Conclusions

Wikipedia appears to be performing the role of a specialist encyclopaedia for all areas of academia, although it is imperfect by having gaps in some areas. Any other specialist encyclopaedia would presumably also have gaps, perhaps for the same reason (the lack of a willing specialist contributor). Thus, the main finding is that *Wikipedia* has wide coverage of academia, but with some omissions. Whilst extensive coverage of research must help it to reach a wider audience, it may cause users to rely on *Wikipedia* and not expect any important topics to be missing.

> *Wikipedia* has wide coverage of academia, but with some omissions

Gaps in *Wikipedia* should not be a problem for field specialist academics who can find and read the original research if a simpler explanation is not available in *Wikipedia*. Other information seekers may turn to lower quality alternatives (**Fallis**, 2008), especially if they cannot understand or afford relevant scholarly publications or do not have the time to read them. There may also be negative consequences if decisions are made upon the apparently reasonable assumption that the absent information *does not exist*. For example, macroergonomic research might be ignored by managers that rely upon *Wikipedia* but have not heard of macroergonomics.

Given these potential negative consequences of topics missing from *Wikipedia*, *academics should ensure that their specialism is adequately covered in Wikipedia.* This especially applies to topics that are useful for non-specialists who have not heard of them or could not find out about them from alternative sources.

English *Wikipedia* seems to include an element of cultural bias (not a pejorative term here), although this seems to be relatively minor. Perhaps academics that are not native English speakers but work in disciplines that publish primarily in English contribute to English *Wikipedia* and help to reduce its cultural bias. Thus, scholars should be careful to check areas that may be of less interest to academics in English speaking countries if the topics may be useful for non-academics that use English *Wikipedia*.

> Academics should ensure that their specialism is adequately covered in *Wikipedia*

Research process descriptions seem to be well covered in *Wikipedia*, except for the generic terms of the History and Philosophy of Science (reconsidering, moderating, disentangling, historicizing, re-thinking, cointegrating, untangling). These seem to be describing analytical approaches that are too general to merit explicit description, however.

The stylistic and structural differences found in some subject areas confirm that academic research in *Wikipedia* is framed differently from in journal articles, with strong linguistic differences in some subject areas. *When editing content, care should be taken to conform with the differing style within Wikipedia in comparison to academic English*. Conversely, the evidence of stylistic differences presented here may also be useful to highlight to junior researchers that they need to adjust their language to cope with the differing formats required by specific subject areas. This may also be important for people conducting multidisciplinary research who may not notice that styles differ between fields.

Finally, librarians training information seekers should make them aware that gaps exist within *Wikipedia*, especially related to non-English speaking countries' culture and geography, but also, to a lesser extent, on any academic topic.

> When editing content, care should be taken to conform with the differing style within *Wikipedia* in comparison to academic English

8. References

Azer, Samy A. (2015). "Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics". *Advances in physiology education*, v. 39, n. 1, pp. 5-14.
https://doi.org/10.1152/advan.00110.2014

Ball, Rafael (2009). "Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966-2005". *Scientometrics*, v. 79, n. 3, pp. 667-679.
https://doi.org/10.1007/s11192-007-1984-5

Béjoint, Henri (2000). *Modern lexicography: An introduction*. Oxford, UK: Oxford University Press. ISBN: 0 19 829951 6

Burke, Ernest A. (2008). "Tidying up mineral names: an IMA-Cnmnc scheme for suffixes, hyphens and diacritical marks". *Mineralogical record*, v. 39, n. 2, pp. 131-135.
http://ima-cnmnc.nrm.se/tidyingupnames.pdf

Callahan, Ewa; **Herring, Susan C.** (2011). "Cultural bias in Wikipedia content on famous persons". *Journal of the American Society for Information Science and Technology*, v. 62, n. 10, pp. 1899-1915.
https://doi.org/10.1002/asi.21577

Cho, Hichang; **Chen, Meihui**; **Chung, Siyoung** (2010). "Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia". *Journal of the Association for Information Science and Technology*, v. 61, n. 6, pp. 1198-1212.
https://goo.gl/sDxL2c
https://doi.org/10.1002/asi.21316

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and psychological measurement*, v. 20, n. 1, pp. 37-46.
https://doi.org/10.1177/001316446002000104

Cook, James; **Plourde, Dawn** (2016). "Do scholars follow Betteridge's Law? The use of questions in journal article titles".

Scientometrics, v. 108, n. 3, pp. 1119-1128.
https://goo.gl/VNu99t
https://doi.org/10.1007/s11192-016-2030-2

De-Laat, Paul (2012). "Open source production of encyclopedias: Editorial policies at the intersection of organizational and epistemological trust". *Social epistemology*, v. 26, n. 1, pp. 71-103.
https://philpapers.org/archive/DELOSP.pdf
https://doi.org/10.1080/02691728.2011.605478

Dick, Bob (2015). "Reflections on the SAGE Encyclopedia of Action Research and what it says about action research and its methodologies". *Action research*, v. 13, n. 4, pp. 431-444.
https://goo.gl/ziRxPh
https://doi.org/10.1177/1476750315573593

Fallis, Don (2008). "Toward an epistemology of Wikipedia". *Journal of the Association for Information Science and Technology*, v. 59, n. 10, pp. 1662-1674.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1263781
https://doi.org/10.1002/asi.20870

Flanagin, Andrew J.; **Metzger, Miriam J.** (2011). "From Encyclopaedia Britannica to Wikipedia: Generational differences in the perceived credibility of online encyclopedia information". *Information, communication & society*, v. 14, n. 3, pp. 355-374.
https://goo.gl/irg1DH
https://doi.org/10.1080/1369118X.2010.542823

Frey-Klett, Pascale; **Burlinson, Peter**; **Deveau, Aurélie**; **Barret, Matthieu**; **Tarkka, M.; Sarniguet, Alain** (2011). "Bacterial-fungal interactions: hyphens between agricultural, clinical, environmental, and food microbiologists". *Microbiology and molecular biology reviews*, v. 75, n. 4, pp. 583-609.
https://doi.org/10.1128/MMBR.00020-11

Giles, Jim (2005). "Internet encyclopaedias go head to head". *Nature*, v. 438, n. 7070, pp. 900-901.
https://doi.org/10.1038/438900a

Gill, Frank B.; **Wright III, Minturn T.**; **Conyne, Sally B.**; **Kirk, Robert** (2009). "On hyphens and phylogeny". *The Wilson journal of ornithology*, v. 121, n. 3, pp. 652-655.
https://goo.gl/z6n9hi

Gorichanaz, Tim (2016). "How the document got its authority". *Journal of documentation*, v. 72, n. 2, pp. 299-305.
https://doi.org/10.1108/JD-09-2015-0117

Gralla, Preston (2009). "What was Encarta? Look it up on Wikipedia". *PCWorld*, 31 Mar.
http://www.pcworld.com/article/162320/what_was_encarta_look_it_up_on_wikipedia.html

Halavais, Alexander; **Lackaff, Derek** (2008). "An analysis of topical coverage of Wikipedia". *Journal of computer-mediated communication*, v. 13, n. 2, pp. 429-440.
https://doi.org/10.1111/j.1083-6101.2008.00403.x

Hartley, James (2005). "To attract or to inform: what are titles for?". *Journal of technical writing and communication*, v. 35, n. 2, pp. 203-213.
https://doi.org/10.2190/NV6E-FN3N-7NGN-TWQT

Hartley, James (2007). "There's more to the title than meets the eye: Exploring the possibilities". *Journal of technical writing and communication*, v. 37, n. 1, pp. 95-101.
https://goo.gl/wWsM4r
https://doi.org/10.2190/BJ16-8385-7Q73-1162

Hasty, Robert T.; **Garbalosa, Ryan C.**; **Barbato, Vicenzo A.**; **Valdes Jr, Pedro J.**; **Powers, David W.**; **Hernandez, Emmanuel**; **John, Jones S.**; **Suciu, Gabriel**; **Qureshi, Farheen**; **Poparadu, Matei**; **San-José, Sergio**; **Drexler, Nathaniel**; **Patankar, Rohan**; **Paz, José R.**; **King, Christopher**; **Gerber, Hilary N.**; **Valladares, Michael G.**; **Somji, Alyaz A.** (2014). "Wikipedia vs peer-reviewed medical literature for information about the 10 most costly medical conditions". *Journal of the American Osteopathic Association*, v. 114, n. 5, pp. 368-373.
https://doi.org/10.7556/jaoa.2014.035

Henderson, Michael; **Selwyn, Neil**; **Finger, Glenn**; **Aston, Rachel** (2015). "Students' everyday engagement with digital technology in university: exploring patterns of use and 'usefulness'". *Journal of higher education policy and management*, v. 37, n. 3, pp. 308-319.
https://goo.gl/cWJWu4
https://doi.org/10.1080/1360080X.2015.1034424

Hughes, Benjamin; **Joshi, Indra**; **Lemonde, Hugh**; **Wareham, Jonathan** (2009). "Junior physician's use of Web 2.0 for information seeking and medical education: a qualitative study". *International journal of medical informatics*, v. 78, n. 10, pp. 645-655.
https://doi.org/10.1016/j.ijmedinf.2009.04.008

Jamali, Hamid R.; **Nikzad, Mahsa** (2011). "Article title type and its relation with the number of downloads and citations". *Scientometrics*, v. 88, n. 2, pp. 653-661.
https://doi.org/10.1007/s11192-011-0412-z

Kousha, Kayvan; **Thelwall, Mike** (2017). "Are Wikipedia citations important evidence of the impact of scholarly articles and books?". *Journal of the Association for Information Science and Technology*, v. 68, n. 3, pp. 762-779.
https://goo.gl/YHnCgF
https://doi.org/10.1002/asi.23694

Kubiszewski, Ida; **Noordewier, Thomas**; **Costanza, Robert** (2011). "Perceived credibility of Internet encyclopedias". *Computers & education*, v. 56, n. 3, pp. 659-667.
https://www.uvm.edu/giee/pubpdfs/Kubiszewski_2011_Computers_and_Education.pdf
https://doi.org/10.1016/j.compedu.2010.10.008

Lærke, Mogens (2014). "Leibniz, the encyclopedia, and the natural order of thinking". *Journal of the history of ideas*, v. 75, n. 2, pp. 237-259.

Lai, Cheng-Yu; **Yang, Heng-Li** (2014). "The reasons why people continue editing Wikipedia content–task value confirmation perspective". *Behaviour & information technology*, v. 33, n. 12, pp. 1371-1382.
https://doi.org/10.1080/0144929X.2014.929744

Landis, J. Richard; **Koch, Gary G.** (1977). "The measurement of observer agreement for categorical data". *Biometrics*, v. 33, n. 1, pp. 159-174.
https://doi.org/10.2307/2529310

Lange, Jessica (2016). "Scholarly management publication and open access funding mandates: a review of publisher policies". *Ticker: The academic business librarianship review*, v. 1, n. 3, pp. 15-27.
https://goo.gl/6aaGgW

Lim, Sook (2009). "How and why do college students use Wikipedia?". *Journal of the Association for Information Science and Technology*, v. 60, n. 11, pp. 2189-2202.
https://goo.gl/muAGJf
https://doi.org/10.1002/asi.21142

Loveland, Jeff (2013). "Encyclopaedias and genre, 1670-1750". *Journal for eighteenth-century studies*, v. 36, n. 2, pp. 159-175.
https://goo.gl/jMhJrx
https://doi.org/10.1111/j.1754-0208.2012.00493.x

Luyt, Brendan (2015). "Debating reliable sources: writing the history of the Vietnam War on Wikipedia". *Journal of documentation*, v. 71, n. 3, pp. 440-455.
https://doi.org/10.1108/JD-11-2013-0147

Masukume, Gwinyai; **Kipersztok, Lisa**; **Das, Diptanshu**; **Shafee, Thomas**; **Laurent, Michaël R.**; **Heilman, James M.** (2016). "Medical journals and Wikipedia: A global health matter". *Lancet global health*, v. 4, e791.
http://dx.doi.org/10.1016/S2214-109X(16)30254-6

Mathew, Manu E.; **Joseph, Anna**; **Heilman, James M.**; **Tharyan, Prathap** (2013). "Cochrane and Wikipedia: the collaborative potential for a quantum leap in the dissemination and uptake of trusted evidence". *The Cochrane database of systematic reviews*, v. 10, ED000069.
https://doi.org/10.1002/14651858.ED000069

Méndez, David I.; **Alcaraz, María-Ángeles**; **Salager-Meyer, Françoise** (2014). "Titles in English-medium astrophysics research articles". *Scientometrics*, v. 98, n. 3, pp. 2331-2351.
http://rua.ua.es/dspace/handle/10045/46404
https://doi.org/10.1007/s11192-013-1174-6

Mesgari, Mostafa; **Okoli, Chitu**; **Mehdi, Mohamad**; **Nielsen, Finn-Årup**; **Lanamäki, Arto** (2015). "'The sum of all human knowledge': A systematic review of scholarly research on the content of Wikipedia". *Journal of the Association for Information Science and Technology*, v. 66, n. 2, pp. 219-245.
https://spectrum.library.concordia.ca/978652/
https://doi.org/10.1002/asi.23172

Messner, Marcus; **DiStaso, Marcia W.** (2013). "Wikipedia versus Encyclopedia Britannica: A longitudinal analysis to identify the impact of social media on the standards of knowledge". *Mass communication and society*, v. 16, n. 4, pp. 465-486.
https://goo.gl/Wh8G4q
https://doi.org/10.1080/15205436.2012.732649

Nielsen, Finn-Årup (2007). Scientific citations in Wikipedia. *First Monday*, v. 12, n. 8.
http://firstmonday.org/article/view/1997/1872
https://arxiv.org/pdf/0705.2106.pdf

Nov, Oded (2007). "What motivates Wikipedians?". *Communications of the ACM*, v. 50, n. 11, pp. 60-64.
https://goo.gl/Fj6fhE

OED (2016). Oxford English dictionary (online).
http://www.oed.com

Orduña-Malea, Enrique; **Thelwall, Mike**; **Kousha, Kayvan** (2017). "Web citations in patents: Evidence of technological impact?". *Journal of the Association for Information Science and Technology*, v. 68, n. 8, pp. 1967-1974.
https://doi.org/10.1002/asi.23821

Paiva, Carlos-Eduardo; **Lima, João-Paulo-Da-Silveira-Nogueira**; **Paiva, Bianca-Sakamoto-Ribeiro** (2012). "Articles with short titles describing the results are cited more often". *Clinics*, v. 67, n. 5, pp. 509-513.
https://doi.org/10.6061/clinics/2012(05)17

Pinfield, Stephen; **Salter, Jennifer**; **Bath, Peter A.** (2016). "The 'total cost of publication' in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions". *Journal of the Association for Information Science and Technology*, v. 67, n. 7, pp. 1751-1766.
https://doi.org/10.1002/asi.23446

Remy, Charlie (2015). "Oxford's research encyclopedias: A new model for reference content?". *Journal of electronic resources librarianship*, v. 27, n. 3, pp. 204-210.
https://doi.org/10.1080/1941126X.2015.1059663

Roessing, Thomas; **Einwiller, Sabine** (2016). "Portrayals of large corporations in the English and German version of Wikipedia –Exploring similarities and differences". *Corporate reputation review*, v. 19, n. 2, pp. 108-126.
https://goo.gl/nMmsHc
https://doi.org/10.1057/crr.2016.3

Rostami, Fatemeh; **Mohammadpoorasl, Asghar**; **Hajizadeh, Mohammad** (2014). "The effect of characteristics of title on citation rates of articles". *Scientometrics*, v. 98, n. 3, pp. 2007-2010.
https://doi.org/10.1007/s11192-013-1118-1

Royal, Cindy; **Kapila, Deepina** (2009). "What's on Wikipedia, and what's not…? Assessing completeness of information". *Social science computer review*, v. 27, n. 1, pp. 138-148.
https://doi.org/10.1177/0894439308321890

Rush, Elizabeth K.; **Tracy, Sara J.** (2010). "Wikipedia as public scholarship: Communicating our impact online". *Journal of applied communication research*, v. 38, n. 3, pp. 309-315.
https://goo.gl/WoX2Lx
https://doi.org/10.1080/00909882.2010.490846

Samoilenko, Anna; **Yasseri, Taha** (2014). "The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics". *EPJ data science*, v. 3, n. 1, pp. 1.
https://doi.org/10.1140/epjds20

Sano, Hikomaro; **Fujiwara, Yuzuru** (1993). "Syntactic and semantic structure analysis of article titles in analytical chemistry". *Journal of information science*, v. 19, n. 2, pp. 119-124.
https://doi.org/10.1177/016555159301900203

Serrano-López, Antonio-Eleazar; **Ingwersen, Peter**; **Sanz-Casado, Elías** (2017). "Wind power research in Wikipedia: Does Wikipedia demonstrate direct influence of research publications and can it be used as adequate source

in research evaluation?". *Scientometrics*, v. 112, n. 3, pp. pp. 1471-1488.
https://doi.org/10.1007/s11192-017-2447-2

Signore, Angelo; **Serio, Francesco**; **Santamaria, Pietro** (2014). "Wikipedia as a tool for disseminating knowledge of (agro) biodiversity". *HortTechnology*, v. 24, n. 1, pp. pp. 118-126.
http://horttech.ashspublications.org/content/24/1/118.abstract

Soler, Viviana (2011). "Comparative and contrastive observations on scientific titles written in English and Spanish". *English for specific purposes*, v. 30, n. 2, pp. 124-137.
https://doi.org/10.1016/j.esp.2010.09.002

Stankus, Tony; **Spiegel, Sarah E.** (2010). "Wikipedia, Scholarpedia, and references to books in the brain and behavioral sciences: A comparison of cited sources and recommended readings in matching free online encyclopedia entries". *Science & technology libraries*, v. 29, n. 1-2, pp. 144-164.
https://doi.org/10.1080/01942620903579435

Swales, John (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press. ISBN: 978 0521338134

Szajewski, Michael (2013). "Using Wikipedia to enhance the visibility of digitized archival assets". *DLib magazine*, v. 19, n. 3/4, pp. 1-8.
https://doi.org/10.1045/march2013-szajewski

Teplitskiy, Misha; **Lu, Grace**; **Duede, Eamon** (2017). "Amplifying the impact of open access: Wikipedia and the diffusion of science". *Journal of the Association for Information Science and Technology*, v. 68, n. 9, pp. 2116-2127.
https://arxiv.org/abs/1506.07608
https://doi.org/10.1002/asi.23687

Thelwall, Mike (2016). "Does astronomy research become too dated for the public? Wikipedia citations to astronomy and astrophysics journal articles 1996-2014". *El profesional de la información*, v. 25, n. 6, pp. 893-900.
https://doi.org/10.3145/epi.2016.nov.06

Thomas, Garry R.; **Eng, Lawson**; **De-Wolff, Jacob F.**; **Grover, Samir C.** (2013). "An evaluation of Wikipedia as a resource for patient education in nephrology". *Seminars in dialysis*, v. 26, n. 2, pp. 159-163.
https://doi.org/10.1111/sdi.12059

Ubriani, Ravi; **Smith, N.**; **Katz, Kenneth A.** (2007). "Reporting of study design in titles and abstracts of articles published in clinically oriented dermatology journals". *British journal of dermatology*, v. 156, n. 3, pp. 557-559.
https://doi.org/10.1111/j.1365-2133.2006.07705.x

Wang, Yan; **Bai, Yongquan** (2007). "A corpus-based syntactic study of medical research article titles". *System*, v. 35, n. 3, pp. 388-399.
https://goo.gl/xZ8U7G
https://doi.org/10.1016/j.system.2007.01.005

White, Arden; **Hernandez, Nelda-Rae** (1991). "Increasing field complexity revealed through article title analyses". *Journal of the American Society for Information Science*, v. 42, n. 10, pp. 731-734.
https://doi.org/10.1002/(SICI)1097-4571(199112)42:10<731::AID-ASI6>3.0.CO;2-W

Wilson, Adam; **Likens, Gene E.** (2015). "Content volatility of scientific topics in Wikipedia: a cautionary tale". *PloS one*, v. 10, n. 8, e0134454.
https://doi.org/10.1371/journal.pone.0134454

Wodak, Shoshana J.; **Mietchen, Daniel**; **Collings, Andrew M.**; **Russell, Robert B.**; **Bourne, Philip E.** (2012). "Topic pages: PLoS computational biology meets Wikipedia". *PLoS computational biology*, v. 8, n. 3, e1002446.
https://doi.org/10.1371/journal.pcbi.1002446

Xu, Bo; **Li, Dahui** (2015). "An empirical study of the motivations for content contribution and community participation in Wikipedia". *Information & management*, v. 52, n. 3, pp. 275-286.
https://goo.gl/NkMFX3
https://doi.org/10.1016/j.im.2014.12.003

Yam, Shing-Chung J. (2016). "Negotiating boundaries of knowledge: Discourse analysis of Wikipedia's Articles for Deletion (AfD) discussion". *Communication and critical/cultural studies*, v. 13, n. 3, pp. 305-323.
https://doi.org/10.1080/14791420.2015.1137334

Yang, Heng-Li; **Lai, Cheng-Yu** (2010). "Motivations of Wikipedia content contributors". *Computers in human behavior*, v. 26, n. 6, pp. 1377-1383.
https://doi.org/10.1016/j.chb.2010.04.011

EPI

El profesional de la información
http://www.elprofesionaldelainformacion.com/autores.html

PRÓXIMOS TEMAS

Número	Mes año	Tema	Envío textos
27, 2	Mar 2018	Indicadores	--
27, 3	May 2018	Libro electrónico y lectura digital	--
27, 4	Jul 2018	Infomediación y automatización	10 mar 2018
27, 5	Sep 2018	Información política y redes sociales	10 may 2018
27, 6	Nov 2018	Información y comunicación biomédica	10 jul 2018
28, 1	Ene 2019	Posverdad y desinformación	10 sept 2018