Avoiding Obscure Topics and Generalising Findings Produces Higher Impact Research[1]

Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK.

Much academic research is never cited and may be rarely read, indicating wasted effort from the authors, referees and publishers. One reason that an article could be ignored is that its topic is, or appears to be, too obscure to be of wide interest, even if excellent scholarship produced it. This paper reports a word frequency analysis of 874,411 English article titles from 18 different Scopus natural, formal, life and health sciences categories 2009-2015 to assess the likelihood that research on obscure (rarely researched) topics is less cited. In all categories examined, unusual words in article titles associate with below average citation impact research. Thus, researchers considering obscure topics may wish to reconsider, generalise their study, or to choose a title that reflects the wider lessons that can be drawn. Authors should also consider including multiple concepts and purposes within their titles in order to attract a wider audience.
Keywords: Citation analysis; article titles; research impact; word frequencies; academic English

Introduction

Academic publications represent a substantial investment of expert time to create, referee, edit and publish. It is therefore worrying for the participants as well as the funders or taxpayers that financed the study if the results are rarely read. One reason why something might be ignored is that it is, or appears to be, about a rarely researched topic so that few people find it or think that it is relevant (e.g., Fox & Burns, 2015). Conversely, some believe that obscure research is essential to science and has been highly successful in the past (Gamboa, 2015; James, 2014; Mexal, 2010). This article uses a term frequency approach to assess the hypothesis that obscure topics are rarely cited. Assuming that obscure (i.e., rarely researched) topics will often be reflected by the presence of unusual title terms, this article assesses whether the presence of such terms associates with low citation rates. Whilst there will be articles on obscure topics without any obscure terms in their titles (e.g., *Fish farming technology in South Turkey during the Bronze Age*) and articles on common topics with obscure words in their titles (e.g., *Is interindexer consistency a hobgoblin?)* the hypothesis is that these are the exceptions rather than the rule.

Many paths may lead to an article being found and read (Tenopir, Wilson, Vakkari, Talja, & King, 2010). These include keyword searching, citation chaining and journal browsing. The choice of an obscure topic, or at least an obscure title for an article, may reduce the likelihood that searchers will enter a relevant keyword and find the article, unless the author keywords are more relevant (e.g., Rostami, Mohammadpoorad, & Hajizadeh, 2014). Similarly, people that cherry-pick interesting articles to read in current issues of journals may ignore topics that are apparently not relevant to their needs. Thus, titles indicating a rarely researched topic (including one that is just very specific) may tend to alienate potential readers (e.g., Sagan, 2013). Conversely, some strange titles may

[1] Thelwall, M. (in press). Avoiding obscure topics and generalising findings produces higher impact research. Scientometrics. Final version at: http://dx.doi.org/10.1007/s11192-016-2159-z

provoke enough interest amongst people to read an article for current awareness even if it does not seem to be directly relevant.

Article title lengths may affect the decision to read an article. The American Psychological Association (APA) guide recommends using a maximum of 12 words but most article titles tend to be longer, and this length has increased over time (Hallock & Dillner, 2016; Guo, Zhang, Ju, Chen, Chen, & Li, 2015). An analysis of the titles of the 25 most cited and 25 least cited articles in medical journals from 2005 found that longer titles were more cited (Jacques & Sebire, 2010). Conversely, longer titles associate with fewer citations in both biology (including biochemistry) and the social sciences and there is no relationship for chemistry (Didegah & Thelwall, 2013) or management science (Nair & Gibbert, 2016). A negative relationship between title length and citations has also been found for UK-authored articles in health and life sciences, natural sciences, geography and economics (Hudson, in press). The most cited psychology articles tend to have shorter titles than typical for psychology articles, but this may be due to higher impact journals tending to have shorter title styles (Subotic & Mukherjee, 2014). Thus, evidence about title length is mixed but suggests that shorter titles are beneficial.

The content of article titles affects decisions to read them. In computer science, journals have different styles for titles (Anthony, 2001) and in one linguistics journal, about a third of titles were found to describe each of: the topic (only); methods; and results (Sahragard & Meihami, 2016). In psychology, articles with amusing titles 1985-1994 tended to receive a below average number of citations (Sagi & Yechiam, 2008), but this factor does not affect management science (Nair & Gibbert, 2016). In medicine, short titles mentioning the results are more frequently cited (Paiva, Lima, & Paiva, 2012). This supports a previous argument that informative titles are more useful (Hartley, 2005; McGowan & Tugwell, 2005). Articles with questions in their titles may be less frequently cited in most disciplines (Jamali & Nikzad, 2011; Hudson, in press). More generally, the presence of non-alphanumeric characters, such as colons and hyphens, within article titles is common throughout academia and associates with higher citation rates, perhaps because their absence marks articles as unusual (Buter & van Raan, 2011).

Some research has used a term frequency approach to analyse the individual words or phrases within article titles. For example, the distribution of nanotechnology-related terms within the titles of relevant journals follows a power law (Bartol & Stopar, 2015). A comparison of word frequencies within article titles in history, sociology, economics and education found history to use substantially rarer terms than the other fields and these were often people or place names (Nagano, 2009). An analysis of changing computing-related term frequencies over time in the titles, abstracts or keywords of library and information science articles discovered that terms that rose and declined in frequency tended to be associated with topical issues or terminologies (Thelwall & Maflahi, 2015). A study of research-related clichés in medical article titles (e.g., “paradigm shift”, “out of the box”) also found these to rise and fall in popularity over time (Goodman, 2012). Within economics 1890-2012 there have also been similar popularity changes in individual terms, such as *tax*, which was the second most popular substantive title term in the 1950s but was out of the top 10 before then and again after the 1960s (Guo, Zhang, Ju, Chen, Chen, & Li, 2015). A comparison between the most frequently used terms between scholarly and trade technical communication publications found trade publication terms to relate to people more (e.g., you, your) whilst scholarly publication terms were more often about the research process (e.g., study, design, research) (Boettger & Friess, 2014). The inclusion of a

country name within a medical article title may associate with fewer citations (Jacques & Sebire, 2010). Country names suggest that an article is primarily or exclusively of interest to a single nation and so their association with low cited articles is unsurprising. Similarly, within ecology, article titles that mention a specific organism are likely to be less cited and articles that mention broad issues are likely to be more cited (Fox & Burns, 2015). In both cases research that seems to be obscure, in the sense of being specific, is less cited. The last paper is the closest to the topic of the current study.

Function words are class of common words that have no topic meaning but serve to bind sentences together. They are in this sense the opposite of rare terms describing obscure topics. Function words include articles (e.g., the, a), pronouns (e.g., it, my, she), conjunctions (e.g., and, but), particles (e.g., if, then), prepositions (e.g., in, under), and auxiliary verbs (e.g., some uses of: has, do) (Selkirk, 1996). Function words are useful to analyse as the polar opposite to obscure terms in order to detect whether it is possible for topic-neutral terms to associate with high or low citation impact. Function words do not seem to have been analysed in the field of scientometrics before. Despite their apparently neutral and topic independent nature, function words can convey useful meta-information about texts and authors. For example, an increased frequency of the personal pronoun *I* can occur in periods of stress (Chung & Pennebaker, 2007). More relevant to the current paper, translations from Japanese to English have been shown to use fewer articles (a, an, the) (Chung & Pennebaker, 2007) and so, extrapolating, it is possible that the presence or absence of specific function words in a title can be faint evidence that an article was originally not written in English but has been translated. Function words in the full text of articles can also point to the likely author gender in some types of texts such as blogs (Koppel, Schler, & Argamon, 2009). This association can be due to an indirect connection to topics. For example, *my* in blogs is an indicator of a likely female author and associates with relationships (e.g., mom, boyfriend, love) whereas *the* in blogs is a male authorship indicator, associating with computing (e.g., software, system, game). From this it is possible, but not obvious, that function words in article titles could associate with topics, and thus, indirectly, with higher or lower citation areas of a field.

Research questions

Despite the above findings, and with one partial exception (Fox & Burns, 2015), no previous study has addressed the general issue of whether obscure research is less cited across academia. This article uses a term frequency approach to address this question, using the presence of unusual words in article titles as an indicator of likely topic obscurity. In addition, the overall relationship between term frequency and citations does not seem to have been addressed at all and so this is a secondary research question. The research questions therefore target these two gaps, with RQ3 serving to counterbalance RQ1 with an analysis of common neutral terms.

- RQ1: Do articles containing unusual words in their titles (i.e., words that are rarely used in other titles from articles in the same field and time period) tend to be less cited?
- RQ2: Does the relationship between the relative frequency of title words and the citation impact of the articles differ between subjects?
- RQ3: Do function words within article titles tend to have a neutral association with citation counts? (i.e., do articles with titles containing a given function word tend to attract the same number of citations as other articles in the same subject?)

Although these questions address all areas of academia, the arts and humanities and social sciences (except within the health sciences) are not included in the results for the reasons outlined in the methods. Hence, the scope of the research questions is the natural, life and formal sciences, as well as engineering and the health sciences.

Methods

The data used was recycled from a previous article (Fairclough & Thelwall, 2015) that did not analyse term frequencies. The data consists of every third subject in each broad Scopus category, giving a wide sample of subject areas encompassing most of academia. The choice of every third subject was arbitrary, driven only by the need for a systematic selection procedure. The non-English versions of article titles were excluded in cases where two languages were provided (e.g., there were Spanish and English versions of the same title for some journals). Categories including any journals publishing articles with exclusively non-English titles were rejected because the presence of other languages would affect the results. Although it would have been possible to remove these non-English journals, this would have changed the nature of the categories and so this was not attempted. This left 18 Scopus categories out of the initial set of 25, each containing journal articles (excluding reviews and non-article documents) from all years from 2009 to 2015 (with partial coverage), as gathered in April and May 2015.

The seven-year time period 2009-2015 seems to be long enough to get enough data on article titles to reliably identify unusual terms. Nevertheless, terms at the start of the period may have been common in a previous period and terms at the end may be common in the future and so this choice has limitations.

Citation counts are widely used in scientometrics as an indicator of scholarly impact for articles. For statistical analyses, citation counts for sets of articles have the disadvantage that they are highly skewed and so a logarithmic transformation is needed to eliminate this (Lundberg, 2007). Another disadvantage of citation counts is that the average number of citations per article varies greatly between years, so the raw citation counts need to be normalised in order to be comparable between years. A simple way to do this is to divide each article's citation count by the average citation count in its field and year (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). After this, normalised citation counts from different years can fairly be grouped together. The following paragraph explains how these ideas were applied to the raw data.

For each category, the subject and year normalised log-transformed citation count $\check{c}$ was obtained for each article so that a value of 1 always indicates an average number of citations, irrespective of subject and year. For this, the Scopus citation count c for each article was first log transformed using $ln(1+c)$ to reduce skewing (Lundberg, 2007; Thelwall, 2016). The arithmetic mean of the log transformed citation counts $\overline{ln(1+c)}$ was then calculated separately for each subject and year. The subject and year normalised log-transformed citation count for each article was computed using $\check{c} = ln(1+c)/\overline{ln(1+c)}$ where the average in the denominator is taken over all articles in the subject and year containing the article. For each subject and year, the resulting set of subject and year normalised log-transformed citation counts should be approximately normally distributed (Thelwall, 2016) with an arithmetic mean of 1. This property is retained if different subjects and/or years are merged.

For each subject separately (but combining all years) a vocabulary was created recording all words in all article titles in all seven years, together with the number of article

titles containing each word. For example, a term with frequency 2 would be in exactly two different article titles within the subject (2009 to 2015), but they might be from different years and the term could occur multiple times in one or both titles.

Within each subject, the average citation impact $\check{c}_t$ of each term t was calculated by taking the arithmetic mean of the subject and year normalised log-transformed citation counts $\check{c}$ for all article titles in the subject containing the term. For example, if there were ten article titles containing the term *study*, then $\check{c}_{study}$ would be the arithmetic mean of the ten $\check{c}$ values for these articles.

Within each subject, the average citation impact $\check{c}_f$ of each term *frequency* f was calculated by taking the arithmetic mean of average citation impact $\check{c}_t$ of each term t with frequency f. For example, if 1000 terms each only occurred in one article title in a subject then the average citation impact $\check{c}_1$ of term frequency 1 for that subject would be the average citation impact of these 1000 terms. Similarly, if there were 100 terms with frequency 2 then $\check{c}_2$ would be the arithmetic mean of the $\check{c}_t$ values for all of these 100 terms.

Approximate confidence intervals were calculated for each word frequency average citation impact $\check{c}_f$ from the standard normal distribution formula from the complete set of subject and year normalised log-transformed citation counts used to calculate it. If n_f is the number of terms with frequency f, then the sample size would be fn_f because each term occurs in f different articles and there are n_f terms. Here the same article can be counted multiple times if its title contains different terms with the same frequency f. This is a hybrid calculation in most cases. For a frequency count of 1, it is a precise confidence interval for the average impact of *all unique terms*. For frequency counts with only one associated term (e.g., if only one term occurred in exactly 500 articles) then the confidence interval is for the average impact of the individual *term*. Between these two extremes, the confidence interval is a purely illustrative hybrid between the two.

Results

Unique words (i.e., terms that occur in only one article title in a subject) were analysed to address the first research question, since unique words are the most unusual in the corpus in terms of frequency in article titles. In all subjects, unique words in article titles associate with lower citation counts (Table 1). Except for Assessment and Diagnosis, the 95% confidence intervals for the citation counts exclude 1, giving statistical evidence of having a below average citation count for the subject. In other words, in all subject areas except Assessment and Diagnosis, if an article from 2009-2015 includes within its title at least one term that is in no other title in the subject area during 2009-2015 then that article can be expected to receive a below average number of citations for its subject and year.

Some of the unique terms are specialist and rare words, such as amentacea (ciboria amentacea is a fungus species that grows on willow and elder tree catkins), Boswelic (Boswelic acid is a tree resin traditionally used in Ayurvedic medicine and being investigated for its anti-inflammatory properties), FACSCanto (in title: *Comparison of two single-platform ISHAGE-based CD34 enumeration protocols on BD FACSCalibur and FACSCanto flow cytometers*), sunnhemp (referring to a hemp plant) and BMP5 (Bone Morphogenetic Protein 5, a protein coding gene). The apparent obscurity of the topics associated with these terms shows that the hypothesis that rare title terms associate with unusual topics has some support in the data. Not all unique terms associate with unusual topics, however. Some

appear to be typographically unusual, such as 10's (article title: *The HI Chronicles of LITTLE THINGS BCDs II: The Origin of IC 10's HI Structure*). Some are lists, such as b2-b8 (article title: *Effect of peptide fragment size on the propensity of cyclization in collision-induced dissociation: Oligoglycine b_2-b_8*). Others are more common words that may be rarely used in titles within a field, such as issuing, tigress, and algorithmically. Overall, then, the results (Table 1) are consistent with the hypothesis that articles on obscure (i.e., rarely researched) topics are more rarely cited because a substantial proportion of the unique title word terms associate with apparently obscure topics. The results are not definitive, however, because the judgement of topics being obscure is qualitative and it is possible that the unique words not referring to obscure topics have more influence, for example if they reflect awkward language constructions by junior researchers or researchers with low fluency in English.

Table 1. Field and year normalised average citation impacts of articles containing unique words in their titles (i.e., words occurring in no other article title from the subject 2009-2015) together with 95% confidence intervals. Articles are counted once for each unique word. The overall average for all articles in each subject is 1 (n=874,411 article titles).

Subject	Average citation impact	Articles in subject	Example random unique term
Assessment and Diagnosis	0.975 (0.922, 1.028)	2830	vaccinations
Ceramics and Composites	0.963 (0.949, 0.977)	69950	sunnhemp
Computational Theory and Mathematics	0.952 (0.931, 0.974)	54455	NQS
Biochemistry	0.911 (0.900, 0.922)	69824	polymethyl
Immunology	0.907 (0.897, 0.918)	67814	FACSCanto
Pharmaceutical Science	0.902 (0.888, 0.917)	69531	Boswelic
Food Animals	0.892 (0.866, 0.918)	16760	BMP5
Complementary and Manual Therapy	0.890 (0.831, 0.949)	2643	PNF
Catalysis	0.867 (0.858, 0.876)	69875	b2-b8
Electrochemistry	0.857 (0.844, 0.869)	65868	ketals
Biological Psychiatry	0.854 (0.836, 0.872)	24378	PMDD
Fuel Technology	0.843 (0.826, 0.860)	65695	7H2
Animal Science and Zoology	0.841 (0.828, 0.854)	67020	amentacea
Astronomy and Astrophysics	0.826 (0.812, 0.840)	68529	10's
Computers in Earth Sciences	0.815 (0.791, 0.838)	10661	algorithmically
Ecology	0.805 (0.793, 0.816)	65390	tigress
Analysis	0.794 (0.766, 0.821)	59430	dicritical
Automotive Engineering	0.755 (0.725, 0.785)	23758	issuing

In answer to the second research question, a visual inspection of the overall term frequency pattern of each subject (see a complete set of graphs in the online supplement: https://dx.doi.org/10.6084/m9.figshare.3806265.v1) suggests that they are all broadly similar, with one partial exception, Assessment and Diagnosis (Figure 1). This subject is unusual because most term frequencies have an above average citation impact. This counterintuitive attribute is due to the presence of many articles from *Nursing* magazine with short titles (e.g., *Break through your fears*) and no citations. Thus, whilst the overall

per-*article* average normalised citation impact is 1, the overall per-*term* average normalised citation impact is 1.4.

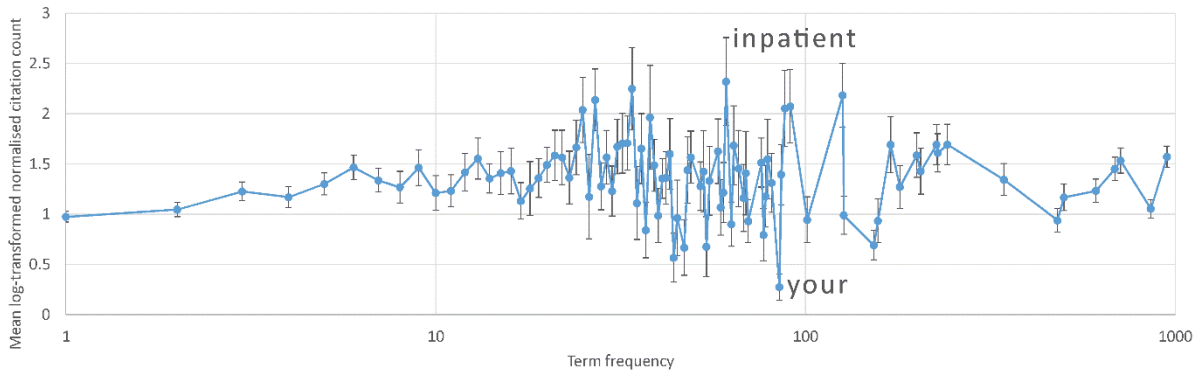

Figure 1. The year-normalised, log-transformed average citation count of title words by frequency for Assessment and Diagnosis 2009-2015. Error bars show estimated 95% confidence intervals. The highest and lowest impact points are annotated with the generating term.

Catalysis, one of the two middle subjects in Table 1, has a typical shape for subjects other than Assessment and Diagnosis, in the sense of an increasing slope on the left, a fuzzy shape with an average value of 1 in the middle, and a jagged line of high frequency values on the right.

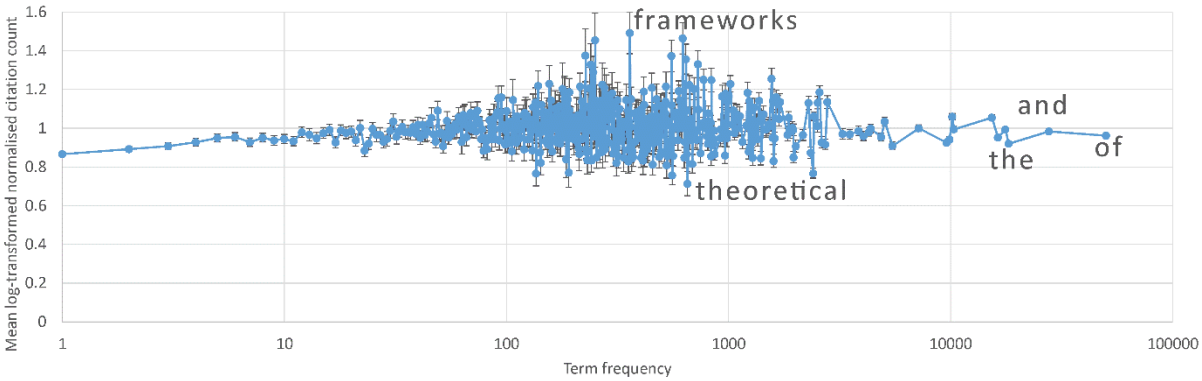

Figure 2. The year-normalised, log-transformed average citation count of title words by frequency for Catalysis 2009-2015. Error bars show estimated 95% confidence intervals. The highest and lowest impact points are annotated, as are the three highest frequency terms.

The individual high and low impact Catalysis terms (Table 2) associate with high or low impact research topics (e.g., batteries: 1.455; propane: 0.771), research types (frameworks: 1.492; study: 0.767; investigation: 0.757; theoretical: 0.713), or claims (e.g. first: 0.821). The low impact association of the general terms in this list contrasts with a previous study for ecology (Fox & Burns, 2015). The low impact association of the term *first* is surprising given that it sometimes signals an explicit novelty claim (e.g., *The first example of asymmetric hydrogenation of imines with Co 2(CO)8/(R)-BINAP as catalytic precursor*). The reason is that it was often used within the phrase "first principals", to denote a research approach that was perhaps less cited than others (e.g., *Selectivity in propene dehydrogenation on Pt and Pt3Sn surfaces from first principles*).

Table 2. The 10 highest and 10 lowest average citation impact terms for Catalysis 2009-2015, together with 95% confidence intervals. Terms must occur in at least 100 different article titles. The average citation impact is the average of the field normalised, log transformed citation counts for articles containing the term in their title.

Term	Average citation impact	Articles
frameworks	1.492 (1.385, 1.598)	359
solar	1.465 (1.384, 1.547)	622
batteries	1.455 (1.315, 1.594)	251
arylation	1.376 (1.240, 1.512)	227
co2	1.374 (1.294, 1.455)	554
visible	1.357 (1.280, 1.435)	643
tio2	1.331 (1.261, 1.401)	729
dots	1.329 (1.223, 1.434)	240
photocatalytic	1.256 (1.204, 1.309)	1563
fluorescent	1.252 (1.197, 1.306)	772
novel	0.831 (0.797, 0.865)	1598
inhibitors	0.829 (0.768, 0.890)	354
first	0.821 (0.752, 0.889)	313
model	0.816 (0.768, 0.864)	687
presence	0.813 (0.757, 0.868)	456
kinetics	0.809 (0.742, 0.876)	527
propane	0.771 (0.694, 0.848)	191
study	0.767 (0.741, 0.792)	2405
investigation	0.757 (0.709, 0.805)	558
theoretical	0.713 (0.650, 0.775)	653

The jagged line on the right hand side of Figure 2 indicates, because of the non-overlapping confidence intervals, small but significant differences in the average impact of individual high frequency terms. These differences pervade all subjects, but are not always the same (Figure 3).

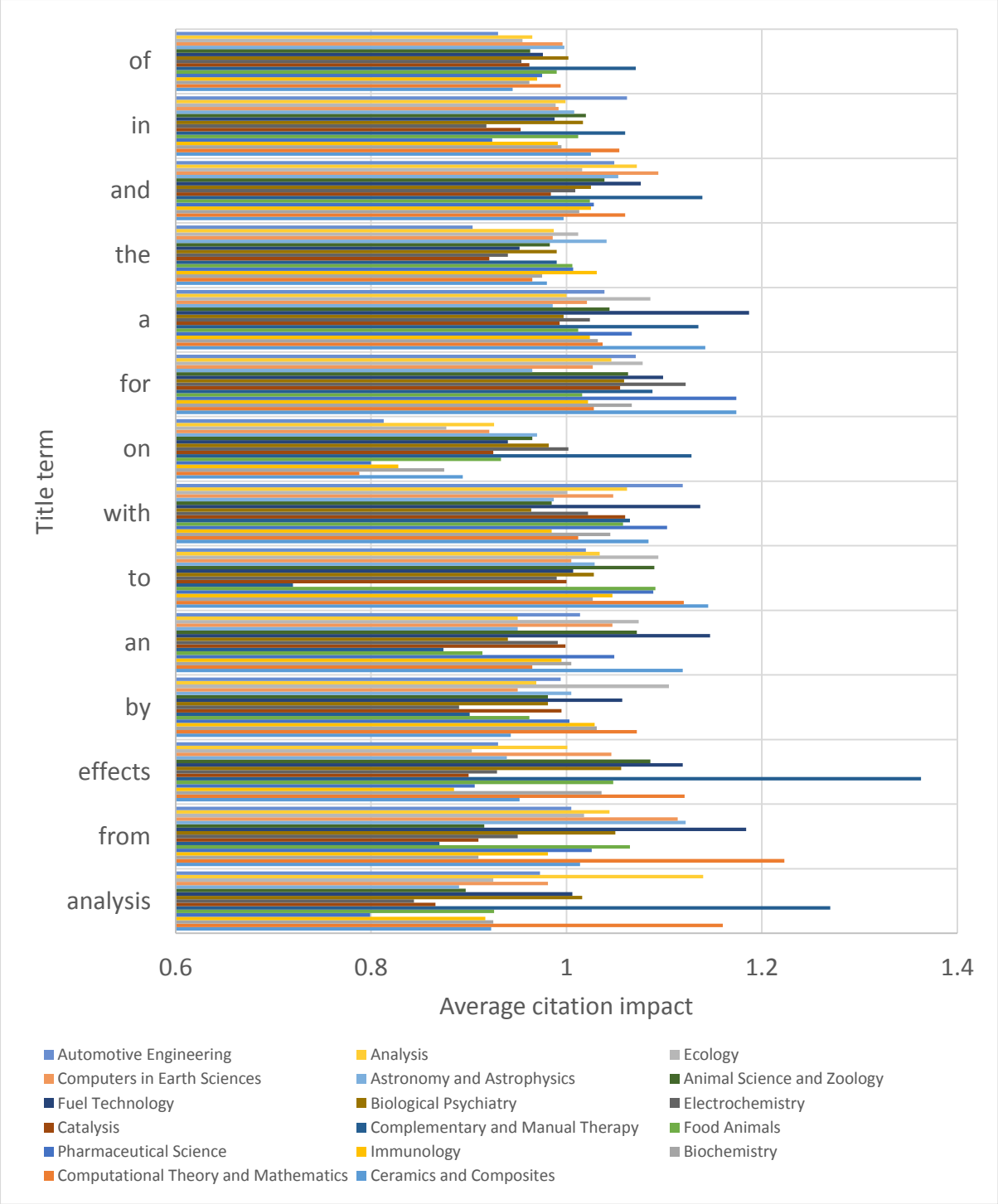

Figure 3. The average citation impact of terms occurring in the titles of all subject areas except Assessment and Diagnosis, which generates much larger outliers. This graph shows the same data as Table 3 and 4 but visualises the variability between disciplines for individual terms.

Although most function words associate with slightly higher impact research overall (Table 3), this is not true for *on* and *the*, both of which may be indicators of specificity. The highest impact common function words are *for,* perhaps indicating an application or purpose, and *and*, suggesting multiple results or applications.

Table 3. The average citation impact of all function words occurring in the titles of all subject areas.

Subject\term	from	by	an	to	with	on	for	a	the	and	in	of
Assessment and Diagnosis	1.18	1.28	1.69	0.94	1.35	1.43	1.17	1.23	1.05	1.53	1.45	1.57
Ceramics and Composites	1.01	0.94	1.12	1.15	1.08	0.89	1.17	1.14	0.98	1.00	1.03	0.95
Computational Theory and Mathematics	1.22	1.07	0.97	1.12	1.01	0.79	1.03	1.04	0.97	1.06	1.05	0.99
Biochemistry	0.91	1.03	1.01	1.03	1.05	0.88	1.07	1.03	0.98	1.01	1.00	0.96
Immunology	0.98	1.03	1.00	1.05	0.99	0.83	1.02	1.02	1.03	1.03	0.99	0.97
Pharmaceutical Science	1.03	1.00	1.05	1.09	1.10	0.80	1.17	1.07	1.01	1.03	0.92	0.98
Food Animals	1.07	0.96	0.91	1.09	1.06	0.93	1.02	1.01	1.01	1.02	1.01	0.99
Complementary and Manual Therapy	0.87	0.90	0.87	0.72	1.07	1.13	1.09	1.14	0.99	1.14	1.06	1.07
Catalysis	0.91	1.00	1.00	1.00	1.06	0.93	1.06	0.99	0.92	0.98	0.95	0.96
Electrochemistry	0.95	0.89	0.99	0.99	1.02	1.00	1.12	1.02	0.94	1.01	0.92	0.95
Biological Psychiatry	1.05	0.98	0.94	1.03	0.96	0.98	1.06	1.00	0.99	1.03	1.02	1.00
Fuel Technology	1.18	1.06	1.15	1.01	1.14	0.94	1.10	1.19	0.95	1.08	0.99	0.98
Animal Science and Zoology	0.92	0.98	1.07	1.09	0.99	0.97	1.06	1.04	0.98	1.04	1.02	0.96
Astronomy and Astrophysics	1.12	1.01	0.95	1.03	0.99	0.97	0.97	0.99	1.04	1.05	1.01	1.00
Computers in Earth Sciences	1.11	0.95	1.05	1.01	1.05	0.92	1.03	1.02	0.99	1.09	0.99	1.00
Ecology	1.02	1.11	1.07	1.09	1.00	0.88	1.08	1.09	1.01	1.02	0.99	0.96
Analysis	1.04	0.97	0.95	1.03	1.06	0.93	1.05	1.00	0.99	1.07	1.00	0.97
Automotive Engineering	1.01	0.99	1.01	1.02	1.12	0.81	1.07	1.04	0.90	1.05	1.06	0.93
Average	**1.03**	**1.01**	**1.04**	**1.03**	**1.06**	**0.95**	**1.07**	**1.06**	**0.99**	**1.07**	**1.03**	**1.01**

Two non-function words were also present in all subject areas, *analysis* and *effects* (Table 4). Ignoring the Assessment and Diagnosis outlier, the term *analysis* associates with lower impact both overall and in most subjects. *Analysis* seems to be particularly unvalued in electrochemistry (0.84) and catalysis (0.87), perhaps because these terms suggest a lack of empirical data. Conversely, in Complementary and Manual Therapy (1.27) it might associate with meta-analyses, which tend to be highly cited. There are also disciplinary variations in the average impact associated with the term *effects*, although the reason is unclear.

Table 4. The average citation impact of the terms *analysis* and *effects* in the titles of all subject areas.

Subject	**analysis**	**effects**
Assessment and Diagnosis	1.67	2.04
Ceramics and Composites	0.92	0.95
Computational Theory and Mathematics	1.16	1.12
Biochemistry	0.93	1.04
Immunology	0.92	0.89
Pharmaceutical Science	0.80	0.91
Food Animals	0.93	1.05
Complementary and Manual Therapy	1.27	1.36
Catalysis	0.87	0.90
Electrochemistry	0.84	0.93
Biological Psychiatry	1.02	1.06
Fuel Technology	1.01	1.12
Animal Science and Zoology	0.90	1.09
Astronomy and Astrophysics	0.89	0.94
Computers in Earth Sciences	0.98	1.05
Ecology	0.93	0.90
Analysis	1.14	1.00
Automotive Engineering	0.97	0.93
Average	**1.01**	**1.07**

The specialism with the lowest average citation impact for term frequency 1 is Automotive Engineering (Figure 4). The cause in this case is the presence of trade magazines, such as *Public Transport International* and *Automotive Industries AI,* that contain rarely cited articles and news about specific localities, or industry events. These include “Public transport in Vienna: Popular, accepted, high quality” and “LG Chem to supply GM with battery cells and electronic components for Chevrolet Volt”. Locality, product and company names provide a collection of low frequency uncited terms (similar to the organism specificity issue within ecology: Fox & Burns, 2015). These are obscure topics in the sense of being highly specific to a company, event or locality rather than focusing on a topic that would be part of the general knowledge for a discipline. To illustrate this, it seems unlikely that many future articles would need to cite information about the electronics supplier for the Chevrolet Volt.

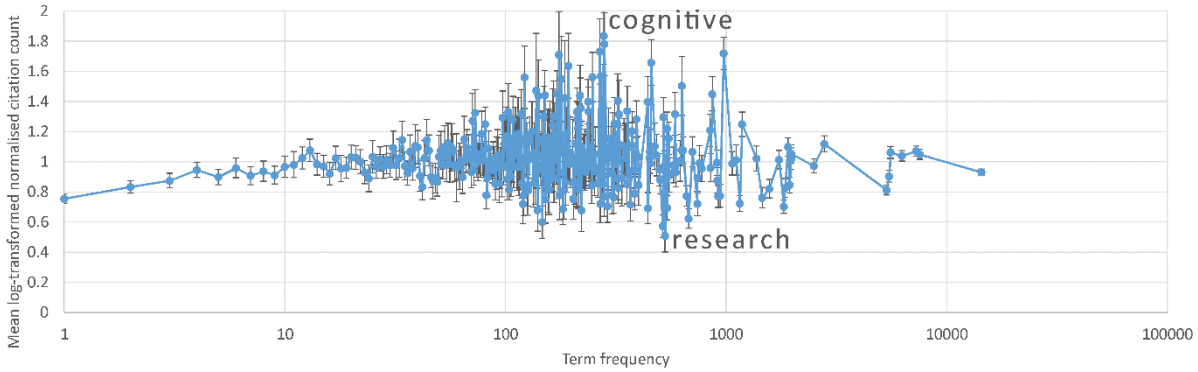

Figure 4. The year-normalised, log-transformed average citation count of title words by frequency for Automotive Engineering 2009-2015. Error bars show estimated 95% confidence intervals. The highest and lowest impact points are annotated.

Discussion

The main limitation of the word frequency analysis reported here is that an individual term can have different meanings (polysemy) and there are also different terms that mean the same thing (synonyms), so word frequency comparisons are simplifications. In addition, a word can be used in different typical contexts that alter its meaning. For instance, the term *analysis* is part of the name of an area of maths (functional analysis), and a specific method (social network analysis) as well as being a general term for generating knowledge and understanding. Another important limitation is the absence of the arts and humanities as well as all core social sciences so that the findings only relate to natural, formal and life sciences, health sciences and engineering. Since only a minority of subjects in these areas have been examined, there may be other fields that follow a different pattern. A technical limitation is that gathering articles over a longer time span would have increased the term frequency counts of some of the words in each category, but probably would not have affected the overall patterns and findings.

The evidence clearly points to unique terms in article titles associating with lower citation impact in all disciplines. This suggests that rarely researched topics tend to attract fewer citations. Although this seems to be the most likely reason, there are alternative explanations. It is possible that authors that describe their topics in an unusual way (e.g., due to a lack of language skills or extreme language proficiency leading to obscure word choices) that alienates potential readers. Authors may also fail to incorporate the generality of the findings into the title, missing out on part of their audience. More seriously, weaker researchers may fail to adequately generalise their findings or may pick narrow topics (e.g., Finberg, 2015) and so their overly specific research has lower impact. The different citation associations of function words undermine the findings somewhat by showing that even the presence of specific neutral words in titles (e.g., *and*) can associate with higher (or lower) average citation impact in different subjects. Since words that are not content words can associate with differences in expected citation rates, the low citation impact of articles with rarely used title words could also be due to causes other than the topics of the articles.

The results also show the same basic pattern in the term frequency graphs for each subject, but with clear disciplinary differences in the citation impact associated with individual terms. It is perhaps surprising that individual function words, such as *the*, can associate with higher impact research in some fields but lower impact research in others. This could be due to different styles adopted within high and low impact journals, the

presence of *the* within phrases associated with a high or low impact sub-fields, the scarcity of definite articles from translated documents in some languages, or the tendency of the definite article to denote a more specific topic.

The almost universally higher citation impact association of the term *and* (i.e., articles with titles containing *and* tend to be more cited) is surprising since the presence of a conjunction seems to connote a longer, more complex title (although three words can easily include "and"), whereas most previous research (reviewed in the Introduction) has found longer titles to associate with fewer citations.

Conclusions

Focusing on the end of the time period examined, the data suggests that in all subject areas examined except one, if a new article is published with a title that includes at least one term that has not been used in a title in the subject area within the previous six years then this article can be expected to receive fewer citations than average for its subject and year. Assuming, with some support from Table 1 and the surrounding discussion, that the cause of this association is that articles with unique title terms tend to be describing obscure topics, then a generalisation of this is that new articles on obscure topics will tend to attract fewer citations than average for their subject and year.

A simple conclusion from this research is that, except perhaps in the arts, humanities and social sciences, researchers should avoid creating titles that make their research seem obscure (i.e., rarely researched) because they may not be read. It seems likely that researchers should also attempt to generalise their studies as far as possible and to highlight this generality when writing their titles. This strategy should lead to research that is more useful to more people and may result in more citations. This advice should be incorporated into the guidelines given to beginning researchers about writing articles (e.g., Hartley, 2005, 2008). Ultimately, the purpose of most research publishing is to attract an audience and composing article titles should be a key part of a strategy to achieve this. Of course, this is only general advice and researchers should not be deterred from attempting to conduct unusual research if they believe that it will attract an audience anyway.

A secondary tentative conclusion, which is a by-product of the research rather than part of the aims, derives from the higher citation association of the term *and* in almost all subjects, which presumably stems from more complex titles since it is a conjunction. It seems that authors should not be afraid to mention multiple things within their article titles as this may show more comprehensive research or may relate to more researchers' topics of interest. This is a tentative conclusion, however, since title lengths do not have a clear association with citation counts. Similarly, the inclusion of *for* within a title suggests a purpose for the research, which seems to be a logical way to attract readers. For future research, it would be useful to investigate the citation association of function words in more detail.

References

Anthony, L. (2001). Characteristic features of research article titles in computer science. IEEE Transactions on Professional Communication, 44(3), 187-194.

Bartol, T., & Stopar, K. (2015). Nano language and distribution of article title terms according to power laws. Scientometrics, 103(2), 435-451.

Boettger, R. K., & Friess, E. (2014). What are the most common title words in technical communication publications? In 2014 IEEE International Professional Communication Conference (IPCC) (pp. 1-7). Los Alamitos, CA: IEEE Press.

Buter, R. K., & van Raan, A. F. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. Journal of Informetrics, 5(4), 608-617.

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In: K. Fiedler (Ed.). Social Communication, New York, NY: Psychology Press (pp. 343-359).

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of Informetrics, 7(4), 861-873.

Fairclough, R., & Thelwall, M. (2015). More precise methods for national research citation impact comparisons. Journal of Informetrics, 9(4), 895-906. doi: 10.1016/j.joi.2015.09.005

Finberg, H. (2015). Journalism professionals, academics debate the value of research. Poynter. http://www.poynter.org/2012/academic-food-fight-over-the-value-of-research/178750/

Fox, C. W., & Burns, C. S. (2015). The relationship between manuscript title structure and success: editorial decisions and citation performance for an ecological journal. Ecology and evolution, 5(10), 1970-1980.

Gamboa, C. (2015). Connecting with the Community: Matt Owens on obscure research that makes a big impact. Sage Connection – Inisight. http://connection.sagepub.com/blog/industry-news/2015/07/15/connecting-with-the-community-matt-owens-on-obscure-research-that-makes-a-big-impact/

Goodman, N. (2012). Familiarity breeds: clichés in article titles. Br J Gen Pract, 62(605), 656-657.

Guo, S., Zhang, G., Ju, Q., Chen, Y., Chen, Q., & Li, L. (2015). The evolution of conceptual diversity in economics titles from 1890 to 2012. Scientometrics,102(3), 2073-2088.

Hallock, R. M., & Dillner, K. M. (2016). Should title lengths really adhere to the American Psychological Association's twelve word limit? American Psychologist, 71(3), 240-242.

Hartley, J. (2005). To attract or to inform: what are titles for? Journal of Technical Writing and Communication, 35(2), 203-213.

Hartley, J. (2008). Academic writing and publishing. London, UK: Routledge.

Hudson, J. (in press). An analysis of the titles of papers submitted to the UK REF in 2014: authors, disciplines, and stylistic details. Scientometrics. doi:10.1007/s11192-016-2081-4

Jacques, T. S., & Sebire, N. J. (2010). The impact of article titles on citation hits: an analysis of general and specialist medical journals. JRSM Open, 1(1), 2. doi:10.1258/shorts.2009.100020

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

James, C. R. (2014). Science unshackled: how obscure, abstract, seemingly useless scientific research turned out to be the basis for modern life. Baltimore, MD: JHU Press.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1), 9-26.

Lundberg, J. (2007). Lifting the crown—citation z-score. Journal of informetrics, 1(2), 145-154.

McGowan, J., & Tugwell, P. (2005). Informative titles described article content. Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada, 26(3), 83-84.

Mexal, S. (2010). The unintended value of the humanities. The Chronicle of Higher Education. http://chronicle.com/article/The-Unintended-Value-of-the/65619

Nagano, R. L. (2009). Lexical comparison of journal article titles in soft disciplines. Porta Lingua 2009, 111-117.

Nair, L. B., & Gibbert, M. (2016). What makes a 'good' title and (how) does it matter for citations? A review and general model of article title attributes in management science. Scientometrics, 107(3), 1331-1359.

Paiva, C. E., Lima, J. P. D. S. N., & Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. Clinics, 67(5), 509-513.

Rostami, F., Mohammadpoorad, A., & Hajizadeh, M. (2014). The effects of the characteristics of title on citation rates of articles. Scientometrics, 98(3), 2007–2010.

Sagan, D. (2013). Cosmic apprentice: Dispatches from the edges of science. University of Minnesota Press

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. Journal of Information Science, 34(5), 680-687.

Sahragard, R., & Meihami, H. (2016). A diachronic study on the information provided by the research titles of applied linguistics journals. Scientometrics, 108(3), 1315-1331.

Selkirk, E. (1996). The prosodic structure of function words. In: J. L. Morgan, K. Demuth (eds.) Signal to syntax: Bootstrapping from speech to grammar in early acquisition. New York, NY: Psychology Press (pp. 187-214).

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

Tenopir, C., Wilson, C. S., Vakkari, P., Talja, S., & King, D. W. (2010). Cross country comparison of scholarly e-reading patterns in Australia, Finland, and the United States. Australian Academic & Research Libraries, 41(1), 26-41.

Thelwall, M. & Maflahi, N. (2015). How important is computing technology for library and information science research? Library and Information Science Research, 37(1), 42–50.

Thelwall, M. (2016). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. Journal of Informetrics, 10(2), 336-346.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics, 5(1), 37-47.