Journal of Informetrics 5 (2011) 608–617

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact

R.K. Buter*, A.F.J. van Raan

Centre for Science and Technology Studies, Leiden University, The Netherlands

ARTICLE INFO

Article history:
Received 22 September 2010
Received in revised form 3 May 2011
Accepted 31 May 2011

Keywords:
Impact
Occurrence
Special characters
Scientific publications
Titles
Bootstrap
Heatmaps

ABSTRACT

We investigated the occurrence of non-alphanumeric characters in a randomized subset of over almost 650,000 titles of scientific publications from the Web of Science database. Additionally, for almost 500,000 of these publications we correlated occurrence with impact, using the field-normalised citation metric CPP/FCSm. We compared occurrence and correlation with impact both at in general and for specific disciplines and took into account the variation within sets by (non-parametrically) bootstrapping the calculation of impact values. We also compared use and impact of individual characters in the 30 fields in which non-alphanumeric characters occur most frequently, by using heatmaps that clustered and reordered fields and characters. We conclude that the use of some non-alphanumeric characters, such as the hyphen and colon, is common in most titles and that *not* including such characters generally correlates negatively with impact. Specific disciplines on the other hand, may show either a negative, absent, or positive correlation. We also found that thematically related science fields use non-alphanumeric characters in comparable numbers, but that impact associated with such characters shows a less strong thematic relation. Overall, it appears that authors cannot influence success of publications by including non-alphanumeric characters in fields where this is not already commonplace.

1. Introduction

Every day, the inbox of a modern researcher readily fills up with emails from friends, colleagues, and even complete strangers. Even more, at stated intervals, emails arrive that contain titles of interesting publications which have recently been added to databases such as Pubmed, Scopus, or the Web of Science. Furthermore, personal messages, electronic forums, web sites, and social networks all require attention and time. Evidently, new scientific literature is only one stream of information that nowadays flows towards a researcher—albeit a rather pivotal one for the profession at hand. Already some time ago, Meadows (1974) estimated that an average researcher had to scan through roughly 3000 titles per year. We assume that this has only become more, and that the increased information burden leaves even less time to deal with them. Clearly, to get attention of potential readers, it is crucial that a publication is presented effectively to a researcher. In many cases, the title is the way to accomplish this (Soler, 2007). Of course, an author could try a tactic employed by writers of certain emails BEGGING FOR attention. Yet, there is a good chance that this will annoy and subsequently put off potential readers, and since being read is an important factor in the professional success of authors, this is evidently not desirable. As writing and publishing is a communal effort, readers are used to certain topics and styles. Authors can use this to their benefit, by using

* Corresponding author at: Wassenaarseweg 62A, P.O. Box 905, 2300 AX Leiden, The Netherlands. Tel.: +31 71 527 3909.
E-mail addresses: buter@cwts.leidenuniv.nl (R.K. Buter), vanraan@cwts.leidenuniv.nl (A.F.J. van Raan).

1751-1577/$ – see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.joi.2011.05.008

familiar ways of phrasing a title in order to facilitate quick reading and to use signal words that are expected to trigger the interest of an audience. Yet, phrasing a title too general can bore: a title has to *stand out* too. Standing out can be accomplished by phrasing differently, for example by using a well-known (but within science not common) literary template such as "to X or not to X" (and filling at the X the particular topic of interest). Alternatively, it could be as simple as using particular, non-alphanumeric characters in a title.

Specific non-alphanumeric characters and title characteristics have been the subject of previous research. Early studies by Dillon (1981, 1982) showed that the colon (":") has become a standard character in titles of scientific publications. Lewison and Hartley (2005) also studied the colon and found differences in title length and colon usage, both over time and over disciplines. Hartley (2007) combined a meta-analysis with new results and showed that colons are preferred by students because they improve the structure of a title, but are not necessarily appreciated by their fellow academics, who make up the intended audience of most scientific publications. However, studies cited by Hartley (2007) failed to find significant differences between the number of citations for publications with and without colons in their title, although the scope of this result was limited to a single journal. Beside the colon, Ball (2009) showed that the question mark has become a frequently appearing in titles in *Medicine* and (to a lesser extend) in *Physics*. We generalize these previous studies on specific aspects of titles and investigate both use of specific characters in publication titles and correlation with impact in a broad and extensive sense. By this, we mean that we do not focus on a particular (non-alphanumeric) character nor limit our investigation to specific journals or science fields.

Our main research question is: given the importance of readership in the success of scientific publications, could something simple as using a particular type of character "boost" the success of a publication. Our hypothesis is that the effect of non-alphanumeric characters on the success of publications is constrained by conventions regarding readability and form. Consequently, if such characters occur and exhibit a positive correlation with the success of publications, those characters usually have a known function or are accepted elements. We investigate this by posing the following research questions. First, what non-alphanumeric characters exist in scientific publications? Then, can we see a difference in the success of publications with and without such characters? Also, are such effects global, or can we see differences over disciplines? Additionally, what is the effect of frequently occurring characters? Finally, how does the use and impact of characters compare over fields?

2. Method

To investigate non-alphanumeric characters in titles, we extracted publications from all research fields available in the Web of Science database[1] (WoS) published in the period 1999–2008. However, the number of publications available in the WoS for that period is large (almost 13 million), which makes exhaustive analyses too time-consuming and we therefore took a representative, 5% random sample from the WoS, reducing the number of publications to almost *650,000*.

To extract the non-alphanumeric characters from the titles of the publications in this set, we used simple regular expression[2] (Aho, 1990). By matching titles with this expression, we got for every publication a (possibly empty) list of non-alphanumeric characters. If this list was empty, we regarded a title as "alphanumeric", and "non-alphanumeric" otherwise.

To express the success of a publication we chose citation counts. Obviously, success expressed in citations is not the same as success expressed in readers (Moed, 2005). Still, we consider this metric appropriate in the context of scientific success, as well as (not unimportantly) generally easier to obtain than number of readers. To calculate the citation rates we proceeded as follows. For the (almost 500,000) articles, letters, notes, and reviews in our sample, we counted the citations from the (almost 13 million) other publications in the WoS published in the same period. Unfortunately, absolute citation counts cannot be used to compare publications published in different fields, because the (average) length of a reference list differs from field to field and hence the expected number of citations. Therefore, we used the CWTS ***CPP/FCSm*** indicator, a field-normalised citation metric that employs the WoS journal subject categories (JSCs) as proxies for fields (Moed, De Bruin, & Van Leeuwen, 1995; Van Raan, 1996). By using this metric, we can compare citation counts over fields. We acknowledge that JSCs have well-known problems with respect to the delineation of related research, and that they provide a rather high-level classification with only 243 categories to represent all scientific research. Nevertheless, they are a readily available categorisation of related publications, with a definition that is standardized in a transparent way and quite stable over time. We measured correlation between the use of non-alphanumeric characters and impact both in general (whole of science) and at the discipline level (such as "mathematics" or "clinical medicine"). Because we regard disciplines as aggregations of related science fields, we also used aggregations of JSCs to represent disciplines. These aggregations have been developed at CWTS (NOWT, 2008) and are actively maintained.[3]

[1] This WoS database is available to the CWTS under license from its publisher Thomson Scientific and contains publications published from 1980 onwards.

[2] We used the Perl-like regular expression `[^\w\s]` which has the following meaning: the square brackets indicate a character set consisting of the character families `\w` (all alphanumeric characters) and `\s` (all whitespace characters), which is negated by the caret (`^`); so this expresses a match of characters which are *neither* alphanumeric nor whitespace.

[3] Maintenance is needed because the JSCs are not stable, but gradually change over time.