Extracting and quantifying eponyms in full-text articles

Guillaume Cabanac

Extracting and quantifying eponyms in full-text articles

Guillaume Cabanac

Abstract Eponyms are known to praise leading scientists for their contributions to science. Some are so widespread that they are even known by laypeople (e.g., Alzheimer's disease, Darwinism). However, there is no systematic way to discover the distributions of eponyms in scientific domains. Prior work has tackled this issue but has failed to address it completely. Early attempts involved the manual labelling of all eponyms found in a few textbooks of given domains, such as chemistry. Others relied on search engines to probe bibliographic records seeking a single eponym at a time, such as Nash Equilibrium. Nonetheless, we failed to find any attempt of eponym quantification in a large volume of full-text publications. This article introduces a semi-automatic text mining approach to extracting eponyms and quantifying their use in such datasets. Candidate eponyms are matched programmatically by regular expressions, and then validated manually. As a case study, the processing of 821 recent *Scientometrics* articles reveals a mixture of established and emerging eponyms. The results stress the value of text mining for the rapid extraction and quantification of eponyms that may have substantial implications for research evaluation.

Keywords Eponymy · Text mining · Regular expressions · Academic publications

> *I have long worshiped the eponym as one of the last vestiges of humanism remaining in an increasingly numeralized and computerized society.*
>
> (Robertson 1972)

G. Cabanac (✉)
Computer Science Department, IRIT UMR 5505 CNRS, University of Toulouse,
118 route de Narbonne, 31062 Toulouse Cedex 9, France
e-mail: guillaume.cabanac@univ-tlse3.fr

Introduction

In his thought-provoking essay on eponymy, Garfield (1983, p. 393) stressed that "Eponyms remind us that science and scholarship are the work of dedicated people." Four decades earlier, Merton (1942, p. 121) had acknowledged the role of this "mnemonic and commemorative device" in the social structure of science. He further defined it as "the practice of affixing the name of the scientist to all or part of what he has found, as with the Copernician system, Hooke's law, Planck's constant, or Halley's comet" (Merton 1957, p. 643). Since then, the topic of eponymy has been extensively discussed. It is even gaining an increasing attention nowadays (Fig. 1). Although a comprehensive review of the literature falls beyond the scope of this article, let us outline a few outstanding contributions before considering the following issue: How to systematically extract eponyms from full-text articles?

Merton (1942, 1957) highlighted the prominent role of eponymy in the reward system of science. From the perspective of the history of science, Beaver (1976) studied the rate of eponymic growth. He discussed a puzzling observation: Although the number of scientists increased exponentialy during the twentieth century, the practice of eponymy remained constant in time. Garfield (1983) commented several features of eponymy, such as its twofold definition,[1] the various purposes of eponyms, and their debated use—especially in medicine. From a psychological and historic perspectives, Simonton (1984) discussed the relation between eponymy and ruler eminence in studying European hereditary monarchs. Further scientometric studies investigated the development of eponymy (Thomas 1992) and its relation with research evaluation (Száva-Kováts 1994) through *non-indexed eponymal citedness*, as the use of an eponym without any proper citation of the original work.

Several research articles have discussed the history and developments of specific eponyms in virtually all fields of science. Some tackle a single eponym, such as the 'Shpol'skii fluorimetry' in analytical chemistry (Braun and Klein 1992), 'Southern blotting' in molecular biology (Thomas 1992), the 'Nash Equilibrium' in mathematics (McCain 2011), and the 'Henry V sign' in medicine (Shanahan et al 2013). In scientometrics, Braun et al (2010) discussed two eponyms (i.e., Garfield's law of concentration and Garfield's constant) coined by and after E. Garfield in the *festschrift* dedicated to his 85th anniversary. Whilst most studies dealt with eponyms based on person names, McCain (2012) studied the use of 'evolutionary stable strategies,' as a non-eponymous case of obliteration by incorporation.[2] On a different note, various eponyms are known for failing to acknowledge the actual discoverers (Stigler 1980). In addition, in some extreme cases, the scientific community called for eponym retraction. For instance, Wallace and Weisman (2000) argued the case against the use of 'Reiter's syndrome' honouring a war criminal and eventually, Panush et al (2007) recommended its retraction and suggested the use of 'reactive arthritis' instead.

[1] "In our day-to-day lives, we frequently encounter places and things named after people. […] The term for a person so honored is "eponym." Thus, Rudolf Diesel is the eponym of the diesel engine. […] In addition to designating the namesake of a word, eponym has a second meaning—a term or phrase *derived from* a person's name. By this definition, diesel engine is also an eponym. The second usage seems to be gaining ascendancy and clearly predominates in the literature consulted for this essay." (Garfield 1983, p. 384)

[2] McCain (2011, p. 1413) traces back this concept to (Merton 1965, pp. 218–219) and, in one of his famous footnotes, Merton (1988, p. 622) also mentions "short proleptic discussions" of it in (Merton 1968b, pp. 25–38). Obliteration by incorporation (OBI) is "the obliteration of the sources of ideas, methods, or findings by their being anonymously incorporated in current canonical knowledge." (Merton 1988, p. 622).

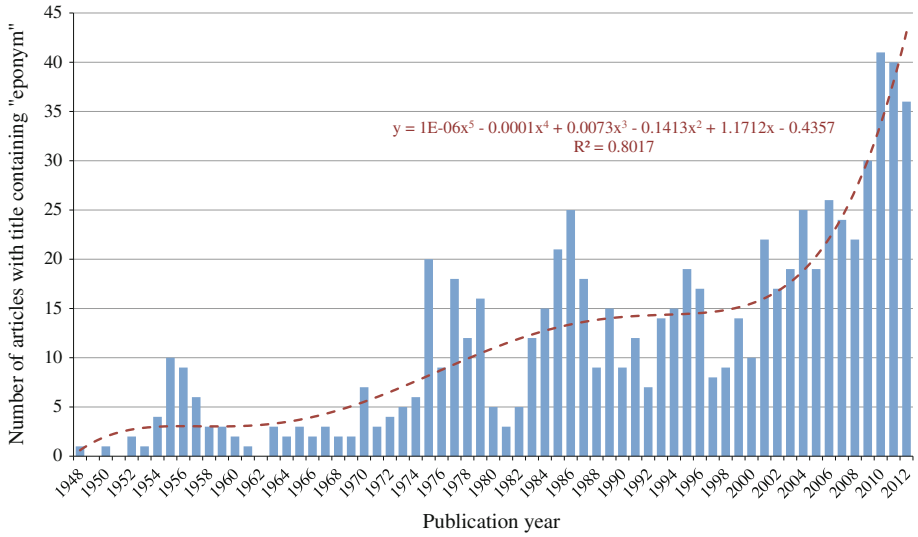

Fig. 1 The number of articles published each year about eponyms has been increasingly growing, with a sharper increase during the last decade. These 743 records were retrieved from the *Web of Knowledge* on April 1st, 2013 by searching for "eponym*" in the Title field. Note that this figure underestimates the literature of eponyms, as articles pre-dating 1948 or lacking this word in their titles were not retrieved (e.g., Merton 1942)

In various fields, editorials and reviews have discussed a few prominent eponyms, such as in chemistry (Cintas 2004), and in forensic pathology (Nečas and Hejna 2012). There is indeed even a study with a narrower scope, addressing the eponyms in the field of education that were named after Spanish people (Fernández-Cano and Fernández-Guerrero 2003). There are also several dictionaries of eponyms of general interest (e.g., Ruffner 1977; Freeman 1997), as well as specialized ones (e.g., Zusne 1987; Trahair 1994).

A more long-standing debate was one dissecting the merits and flaws of eponyms in medicine, where they are widely practiced (e.g., see Boring 1964; Robertson 1972; Kay 1973). The climax of this debate is illustrated by the double-page spread in the *British Medical Journal* that featured a supporter (Whitworth 2007) and two opponents (Woywodt and Matteson 2007) of eponyms head-to-head. It seems that the matter is still not settled, though.

In the early days of the *Science Citation Index*, Garfield (1965, p. 189) reflected on the feasibility of inferring *implicit* references present in documents, evoking the case of an "eponymic concept or term." It turns out that the impact of a research contribution is underestimated when considering *explicit* references only (Garfield 1973). Száva-Kováts (1994) notably raised this point to dispute the views of Cole and Cole (1972), who refuted an anti-elitist theory that they dubbed the *Ortega Hypothesis*.[3] Száva-Kováts (1994, p. 60) concluded that "the data of citedness based on citation indexes are quite inadequate to indicate the real measure of actual citedness of scientists in scientific articles."

Some researchers have attempted to extract and quantify every eponym used in various fields of science. Several manual methods have been devised, targeting various materials. For instance, Diodato (1984) looked at the titles of articles in psychology and mathematics.

[3] See *Scientometrics*, 12(5–6), 1987 for further discussions about the *Ortega Hypothesis*.

Braun and Pálos (1989, 1990) perused the subject indexes of chemistry textbooks.[4] Besides the inspection of various dictionaries and source books in psychology (Roeckelein 1995), Roeckelein (1972, 1974, 1995) reported a series of line-by-line content analysis of introductory psychology textbooks. These daunting tasks were performed with the help of an army of student volunteers, who tediously marked the textbooks "without knowing why," as acknowledged in the footnotes of (Roeckelein 1972, 1995).

The present article is concerned with this latter line of research, as we tackle the following question: How can we improve eponym extraction and quantification from full-text articles? It is possible that computing capabilities can provide an affordable, fast, reliable, and replicable method for extracting and quantifying eponyms in scientific articles.

Method

We designed a semi-automatic text mining approach to extract eponyms from any collection of documents, such as the articles published in an academic journal. The proposed approach relies on following two steps. *First*, each document is processed by a computer program using regular expressions to detect candidate eponyms in the text (e.g., Bradford's bibliographical law). *Second*, these candidate eponyms are manually validated and labelled with the underlying person's name (e.g., Bradford). Eventually, a list of names is produced ranked by frequency of appearance. We detail these two steps in the following sections.

Step 1: automatic extraction of candidate eponyms with regular expressions

Building on a standard text mining technique, we rely on regular expressions to identify eponyms in texts. A regular expression defines a pattern of text that is to be matched in a given document. For instance, the pattern `[A-Z][A-Za-z]+ian` matches any character string that starts with a capital letter (`[A-Z]`) followed by at least one letter (`[A-Za-z]+`) and ends with the three letters `ian`. As a result, this pattern matches the word "Mertonian" in the title "What is Mertonian sociology of science?" of (Hargens 2004), for instance. The interested reader is referred to (Friedl 2006) for a thorough coverage of regular expressions.

In Step 1, each document is processed by the computer program provided in Listing 1 (see Appendix) in search of eponyms. This program parses textual contents with the regular expression that is illustrated in Fig. 2. Here we tackle two kinds of eponyms:

- Adjectival eponyms are matched in Part 1 of the regular expression. Such eponyms include, for example, Hippocratic medicine, Aristotelian logic, Euclidean geometry, Boolean algebra, and Keynesian economics. This list found in (Merton 1957, p. 643) is certainly not comprehensive, but we used it as a cue to match the suffixes -ean, -ian, and -tic in Part 1. Still, this list can easily be tailored in Listing 1.
- Nominal eponyms appear in various expressions mixing up the name(s) of person(s), bibliographic references, and the target of the eponym (e.g., a law, a distribution). Here are some examples of eponyms matched by our regular expression:

[4] Besides extracting and counting eponyms manually, Braun and Pálos (1989) also plotted the distribution of eponyms according to their presumed origin. Their results seem to confirm Beaver's (1976) observation in that the beginning of the twentieth century might have marked "an alteration in the structure of the reward system in science, a movement towards increased anonymity."

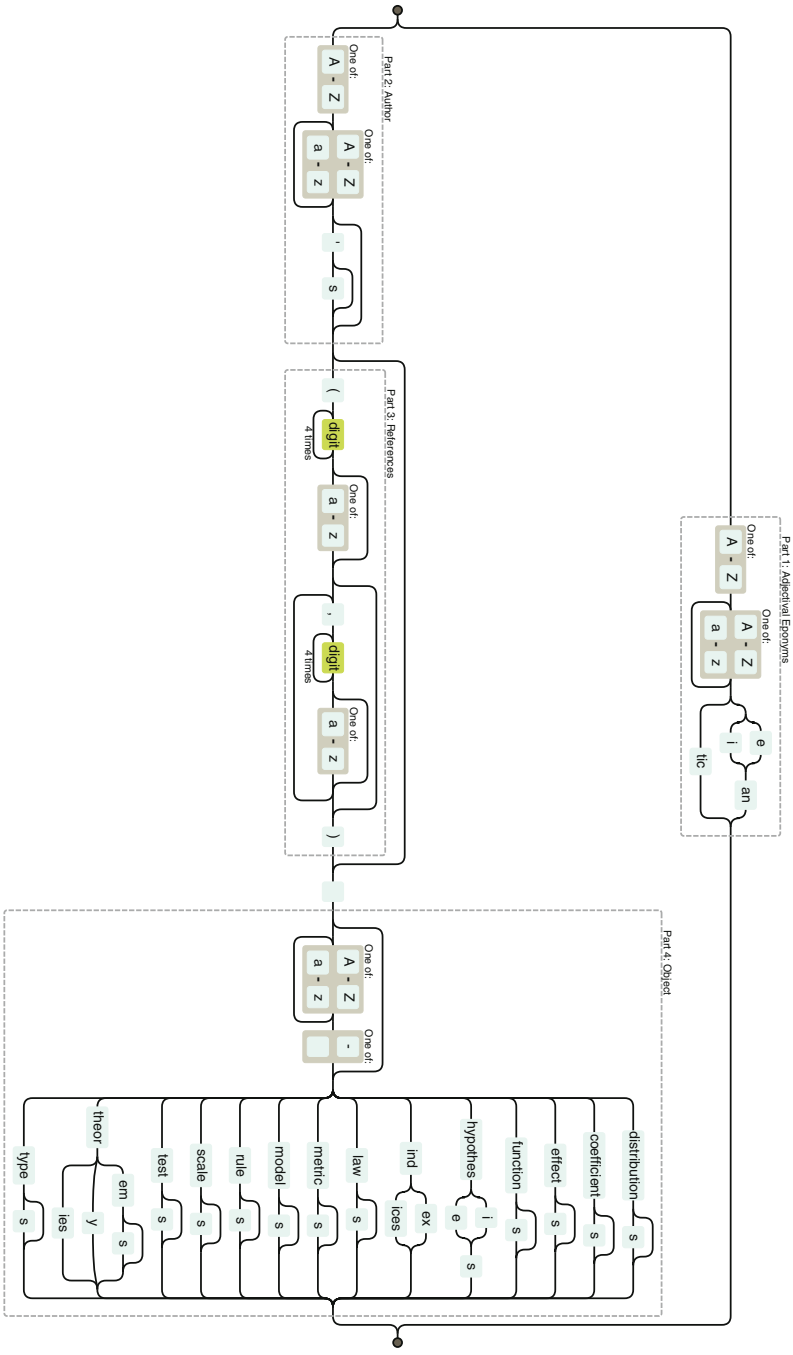

Fig. 2 Syntax diagram of the regular expression used in Listing 1 to extract eponyms from text. The upper sub-expression (i.e., Part 1) matches adjectival eponyms (e.g., 'Mertonian'), whilst the lower sub-expression (i.e., Parts 2–4) matches nominal eponyms, such as 'Vinkler's (2010a, 2013) π_v-index.' This diagram was produced by http://www.regexper.com

- 'Bradford's Law' is matched by Parts 2 and 4.
- '[The] Hirsch *h*-index' is matched by Parts 2 and 4.
- 'Vinkler's (2010a, 2013) π_v-index' is matched by Parts 2–4.

Eponyms are known to appear in both possessive and non-possessive forms (e.g., consider 'Lotka's law' *versus* 'Lotka law').[5] The regular expression in Part 1 matches both.

Note that Parts 1–2 match capitalised eponyms only. Nonetheless, Garfield (1983, p. 384) noted that many listed in (Ruffner 1977) are no longer capitalized once absorbed in everyday language (e.g., diesel engine, saxophone). Here we rely on capital letters as a clue to eponymy, at the expense of such absorbed, non-capitalised eponyms. In addition, the proposed regular expressions consider author-date referencing style, as recommended by the American Psychological Association (APA 2010, Chap. 6), amongst others. Note that the proposed regular expression also handles other less complex referencing styles, such as numeric referencing (e.g., 'Vinkler's π_v-index [3, 6]'). Eventually, the set of words used in Part 4 of the regular expression (i.e., distribution, coefficient) should be reviewed and tailored regarding the domain under study.

The outcome of Step 1 is a list of candidate eponyms weighted by their number of occurrences in the processed documents. Let us stress that any eponym found several times in a given document contributes only one point to its weight in the list. Thus, there is no over-representation of a given eponym just because it was used a large number of times in a few papers. The weight of an eponym in the result list is thus correlated with its acceptance in the considered research community.

Step 2: manual validation and labelling of eponyms

The proposed regular expression is designed to match capitalised eponyms. Unfortunately, other non-eponymic expressions are matched too (e.g., 'Average *h* index'). Such 'false positive' expressions have to be identified and discarded manually.

In Step 2, each eponymic expression extracted during Step 1 is manually assessed. If the expression does not correspond to an eponym, it is dropped. Otherwise, it is labelled with the target person's name. The human assessor may use any available dictionary of eponyms (e.g., Ruffner 1977; Freeman 1997) in conjunction with online materials for this task. Eventually, all eponymic expressions associated with the same person's name are grouped and their weight are summed up. For instance, 'Hirschian' (found in 5 papers) and 'Hirsch's *h*-index' (found in 45 papers) are grouped under the 'Hirsch' label with weight 50.

Data

As a case study, the method was applied to the *Scientometrics* journal. All available full-text articles were considered: 41 issues numbered 82(1) to 95(2) published from 2010 to 2013 were assessed containing 821 articles. These were downloaded in HTML format from

[5] MacAskill and Anderson (2013) reviewed the pros and cons voiced in neurology about these two forms. For instance, Smith (1975) recommended the discontinuation of the possessive form for naming morphologic defects since "the author neither had nor owned the disorder, e.g., Down syndrome." And Haines and Olry (2003) ironically stressed that "James Parkinson did not die of his own personal disease … he died of a stroke." Jana et al (2009) claimed that the inconsistency in the use of the two forms hinders literature search. Garfield (1983, pp. 389–390) supports the possessive form for non-clinical eponyms, as "in any science, an original theory is an individual's intellectual invention."

SpringerLink. Eventually, the formatting instructions were dropped by stripping out HTML tags. This data cleaning process resulted in one file of plain text for each of the 821 articles.

Results

The 821 full-text files were processed by the computer program showed in Listing 1. Step 1 resulted in 3,457 candidate eponyms (see Online Supplementary Material). In Step 2, only 493 of these candidate eponyms passed manual validation. Eventually, there were 226 distinct person names targeted by these validated eponyms, and Fig. 3 shows the most frequent names cited in the eponyms of the processed articles. A graphical view of this Hall of Fame is shown in Fig. 4, where name sizes are proportional to their frequency as reported in Fig. 3. To the best of my knowledge, this is the first attempt at semi-automatic eponym extraction and quantification from full-text articles.

The most frequently eponymised person in the corpus is Jorge E. Hirsch, professor of physics at the University of California, San Diego. He is acknowledged for inventing the *h*-index (Hirsch, 2005), which gauges the impact of an author's research according to his/her number of publications and citation rate. The *h*-index has soon attracted a great deal of attention in the community of scientometrics, as dozens of articles have discussed it and proposed extensions to it (Schreiber et al 2012). In this study, 15% of the 821 articles featured an eponym such as 'Hirsch index,' 'Hirsch's *h*-index,' and the adjective 'Hirschian.' Here we need to bare in mind the following two empirical observations about eponyms:

> "First, names are not given to scientific discoveries by historians of science or even by individual scientists, but by the community of practicing scientists (most of whom have no special historical expertise). Second, names are rarely given, and never generally accepted unless the namer (or accepter of the name) is remote in time or place (or both) from the scientist being honored." (Stigler 1980, p. 148)

Hirsch's (2005) article was published in November 15, 2005 with *PNAS*, but a preprint[6] was already available online as of August 3, 2005. Hirsch does not appear to have coined himself the eponym 'Hirsch index,' since the only occurrence of his name appears in the article's byline. A few days after the preprint was posted on arXiv (August 18), Ball's (2005) *Nature* article publicized the *h*-index. Following a quick framing of Hirsch and his invention, the second paragraph of this article starts with "His 'h-index' depends on" without introducing any eponymic version, though. Eventually, one of the first uses of the eponym in reference to the *h*-index (i.e., 'Hirsch-type index') seems to be due to Braun et al's (2005) paper dated November 21, 2005. This eponym spread like wildfire in *Scientometrics*, starting with van Raan's (2006) article received on December 1, 2005, and subsequent others published in 2006 onward (e.g., Egghe and Rousseau 2006; Banks 2006; Braun et al 2006)—note the presence of the eponym in the title of these articles!

Still regarding the Hirsch index, let us go back to Stigler's (1980, p. 148) aforementioned observations. The first one obviously applies, since leading scientometricians introduced and publicized the use of this eponym. The second observation, however, does not seem to fit here: only one week[7] separated the *PNAS* publication and the first

[6] http://arxiv.org/abs/physics/0508025v1

[7] This immediateness has to be contrasted with Thomas's (1992) question and the clues from the literature that she recalled: "How long does it take for an eponymous event to achieve eponymy? There have been suggestions ranging from one or two years (Diodato 1984) to 61 years (Stigler 1980)."

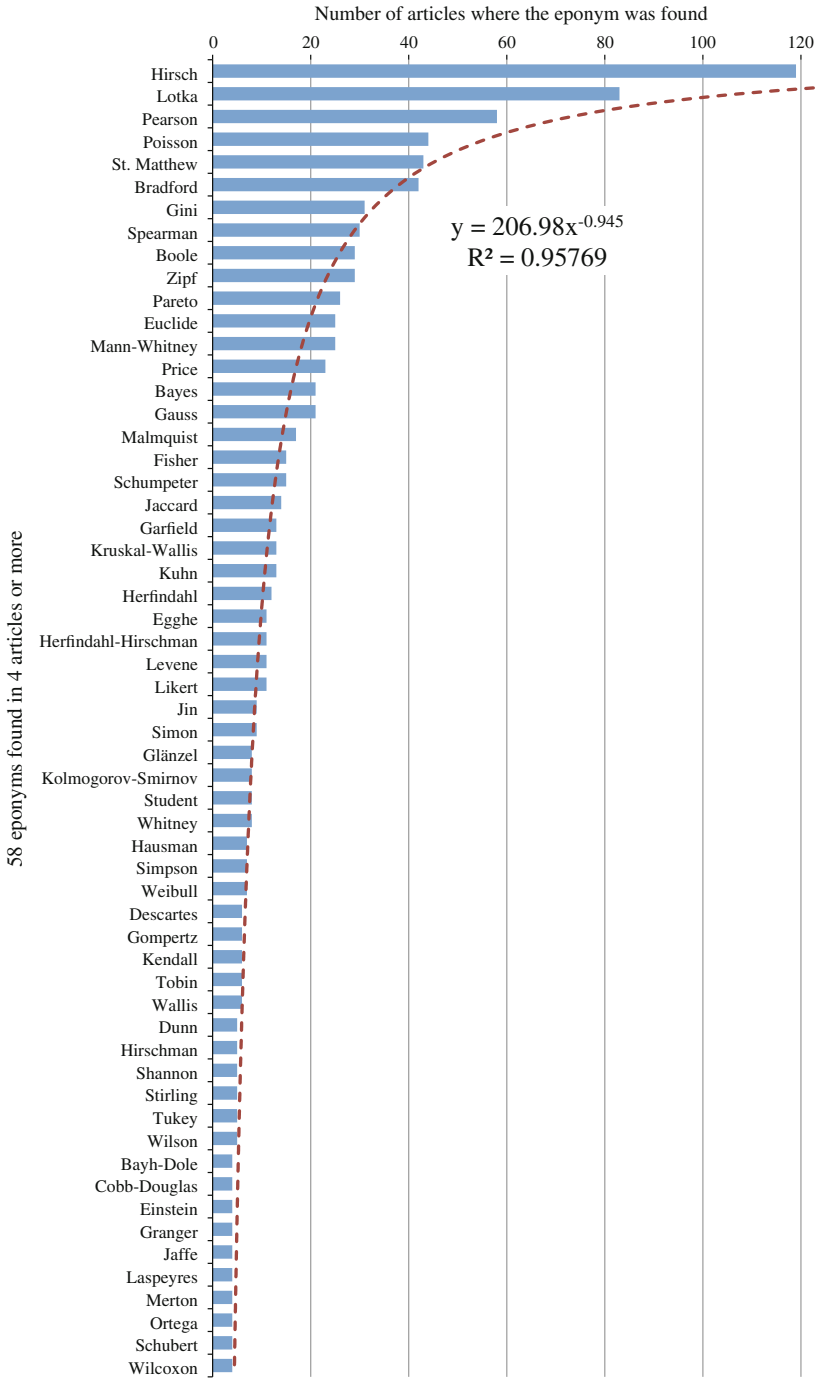

Fig. 3 The 58 most frequent person names cited in 821 *Scientometrics* articles published in 2010–2013. The distribution of decreasing name occurrences fits a power law ($R^2 = 0.9577$)

Fig. 4 Hall of Fame for the eponimised persons extracted from 821 *Scientometrics* articles published in 2010–2013 (see Fig. 3). This word cloud was produced by http://www.wordle.net

occurrence of the eponym 'Hirsch-type index.' Is this to say that Hirsch's (2005) *h*-index was so extraordinary that it challenged and defeated the venerable mechanics of eponymy? Or are we facing a case of *self-suggested* eponymy? Had Hirsch picked up another letter than *h* (which incidentally turns out to be his initial), would we be commenting the ubiquity of eponyms named after him? Let us leave this question to historians of science.

A number of influential scholars are praised in various highly cited eponyms in Fig. 3. Some acknowledge the work of discoverers of famous laws and distributions that are of particular interest to scientometrics, such as Bradford, Gauss, Lotka, Pareto, Poisson, and Zipf (see e.g., Bar-Ilan 2008; De Bellis 2009). Many other eponyms praise the inventors of statistical procedures and tests,[8] such as Fisher, Gini, Kolmogorov–Smirnov, Kruskal–Wallis, Mann–Whitney, Pearson, Spearman, and Student (see e.g., Kotz et al 2005). Prominent scholars who published with *Scientometrics* were also eponymised: de Solla Price, Garfield, and Merton, all three being early gatekeepers to serve the journal (Beck et al 1978; Braun 2004). We also note the emergence of eponyms coined after three 'younger' scientometricians sitting in the board of the journal: Egghe, Glänzel, and Schubert.

[8] These are prone to *non-indexed eponymal citedness* (Száva-Kováts 1994), as some manuals of style advise not to cite the associated references. For instance, the *Publication Manual of the American Psychological Association* states: "Do not give a reference for statistics in common use; this convention applies to most statistics used in journal articles. Do give a reference when (a) less common statistics are used, especially those that have appeared so recently that they can be found only in journals; (b) a statistic is used in an unconventional or a controversial way; or (c) the statistic itself is the focus of the article." (APA 2010, p. 116).

Finally, the perspicacious reader might have noted that St. Matthew ranks highly in the list (fifth). This is due to him starring in the 'Matthew effect' eponym coined by Merton (1968a). It is well established, however, that "St. Matthew did not discover the Matthew Effect" (Stigler 1980, p. 148), thus illustrating Stigler's purposely ironic Law of Eponymy: "No specific discovery is named after its original discoverer."[9]

Discussion

The main issue in this paper is how well this methodology worked in terms of standard criteria used in the Information Retrieval (IR) community: efficiency and effectiveness (see Kelly 2009, pp. 116–119).

Efficiency relates to the performance of an IR system in minimizing execution time and space needed (i.e., storage). The computer program involved in Step 1 processed the 821 articles in about 30 seconds using a regular laptop of year 2011. In contrast, Step 2 took longer as it required the manual validation of the 3,457 candidate eponyms. Note that there are affordable ways to reduce the duration of this task. For instance, crowdsourcing would allow the distribution of this manual validation task (with potential redundancy to satisfy with quality concerns) to several people working for a small amount of money. This opportunity has proved effective in IR, where so-called 'workers' assess the relevance of documents with respect to a given textual query (see Lease and Yilmaz 2013).

Effectiveness relates to the performance of an IR system in producing quality results. Here, two measures are usually considered. On the one hand, *precision* measures the extent to which the retrieved eponyms are relevant. The method yields 100% precision, as candidate eponyms are manually validated. On the other hand, *recall* measures the extent to which all relevant eponyms (in the whole corpus of articles) are retrieved. It is not possible to measure recall in this case study because we do not know the set of all eponyms used—this would require the manual labelling of all the eponyms in the 821 articles. We leave to future work the suggestion of creating such 'ground truth' for benchmarking eponym extraction approaches.

All in all, the proposed method for eponym extraction and quantification yields precise results but does not guarantee completeness. It should be stressed here that the regular expression used in Listing 1 (especially Part 4 showed in Fig. 2) should be tailored to the scientific domain under study. For instance, it should be extended to match eponyms such as 'Brownian motion' and 'Schrödinger equation' in physics.

Another point worth discussing relates to *non-indexed eponymal citedness*, as commented by Száva-Kováts (1994, pp. 68–69):

> "this phenomenon is a very frequent and long-standing feature in the journal literature of physics, with a permanent and growing importance. It demonstrates that not 'a handful' of the most eminent scientists who are creating paradigms of science, but roughly 2,000 eponymous scientists have the chance to be mentioned in the text of recent articles on physics with their scientific achievements in eponymal form, that is, without any formal bibliographical reference. Hence, this mass of scientific people faces the possibility of losing indexed citations this way. It points out that the stock of non-indexed eponymal citations amounts to a third of the indexed ones at both

[9] It happens, however, that Kennedy (1972) had coined Boyer's Law earlier: "Mathematical formulas and theorems are usually not named after their original discoverers."

ends of a 30-year period (1939–1969) representing two ages in the history of science, in two representative source journals of physics."

It must be acknowledged here that the present method extracts eponyms from texts, without any consideration of them being indexed or not in the reference sections. As it stands, it is thus an optimistic approximation of *non-indexed eponymal citedness*.

Finally, the present method overlooks two subtle manifestations of acknowledgements in the same vein as eponyms. On the one hand, some units of measurement are of eponymic nature *per se* (e.g., 1N for one newton, 2J for two joules). On the other hand, there are papers featuring the name of a scientist as a keyword, such as the mathematicians Galton and Pearson in (Stigler 1989), or the (younger) psychologist Hartley in (Zhang and Liu 2011). It might be worth devising a further text mining approach to unveil such implicit citations of an author's œuvre in keywords, but this must be left to future work.

Conclusion

In his take on the reward system of science, Merton (1957, p. 642) suggested that "heading the list of immensely varied forms of recognition long in use is eponymy." Prior work reported results of painstaking manual extractions of eponyms from various materials (Diodato 1984; Braun and Pálos 1989, 1990; Roeckelein 1995). These authors operated on the titles of research articles, as well as on the subject indexes appearing in textbooks.

In the present study a semi-automated text mining approach was introduced to extract and quantify eponyms from full-text articles. The method relies on a computer program processing text with regular expressions (Listing 1), followed by manual validation of the candidate eponyms found. This approach was tested on a corpus of 821 articles published in *Scientometrics* from 2010 to 2013. Thus the findings stress the distribution of eponyms named after prominent scientists in the fields of mathematics and scientometrics. To the best of my knowledge, this is the first attempt of eponym quantification from full-text articles.

Such a text mining approach may be applied to unveil the most prominently eponymised scientists in any field of science. The results may also contribute to spotting new research trends (e.g., the *h*-bubble as coined by Rousseau et al (2013)), and updating existing dictionaries of eponyms (e.g., Ruffner 1977). It only requires that full-text articles are available.[10] The method may well serve as an umpteenth illustration of the potentials of text mining for understanding the developments of science (Nature 2012; Van Noorden 2012).

Acknowledgements I am indebted to Prof. Tibor Braun, who brought to my attention his attempts to unveil eponyms from chemistry textbooks (Braun and Pálos 1989, 1990; Braun and Klein 1992) and suggested the use of computing capabilities for eponym mining. I also acknowledge the feedback of Prof. James Hartley and Dr. Gilles Hubert on an earlier version of this article.

[10] Although considering all full-text articles currently available online, the present study only covers years 2010 to 2013. As a result, Figs. 3 and 4 show the use of eponyms in the contemporary literature of *Scientometrics*. It would have been interesting, however, to extract and quantify the eponyms used since the inception of the journal. As a follow-up to the seminal work on authorship by de Solla Price and Gürsey (1975), a longitudinal study may reveal the dynamics of eponyms in terms of transience and continuance.

Appendix: computer program for eponym extraction and quantification

Listing 1 Bash script processing a corpus of full-text articles to extract and quantify eponyms

```
#!/bin/bash
# Extracts Eponymic Expressions from Textual Documents (e.g., Hirsch's h index)
# Requires GNU Coreutils software (/opt/local/libexec/gnubin in the path)
# License: Creative Commons Attribution-ShareAlike 3.0 Unported License.
# (see http://creativecommons.org/licenses/by-sa/3.0)
# @author Guillaume Cabanac (guillaume.cabanac@univ-tlse3.fr)
# @version 14-MAY-2013

# Parts of the regular expression matching eponyms ($RE_EPONYMS)
RE_ADJECTIVE="[A-Z][A-Za-z]+(?:(?:e|i)an|tic)"
RE_AUTHOR="[A-Z][A-Za-z]+(?:s?)?"
RE_REFERENCE="(?:␣\(\d{4}[a-z]?(?:,␣\d{4}[a-z]?)*\))?"
RE_OBJECT="(?:[A-Za-z]+[-␣])?(?i:distributions?|coefficients?|effects?|functions?| ✂
     hypothes(?:i|e)s|ind(?:ex|ices)|laws?|metrics?|models?|rules?|scales?|tests?| ✂
     theor(?:ems?|y|ies)|types?)"
RE_EPONYMS="$RE_ADJECTIVE|$RE_AUTHOR$RE_REFERENCE␣$RE_OBJECT"

# Matches non-eponyms
RE_NO_EPONYMS="(?:All|An|Her|His|In|Its|The|These|This|Thus|Two|Our|We|When|While)\s ✂
     +"

# Create a temp file where eponymic expressions found in articles are concatenated
EXTRACTED_EPONYMS=$(mktemp)

# Extract eponymic expressions from each article. Concatenate in $EXTRACTED_EPONYMS
for ARTICLE in $(ls ScimPapers/s11192-*.txt) ; do
  sed -r "s/[[:blank:]]+/␣/g" $ARTICLE | # Drop multiple whitespaces
  sed    "s/[^A-z0-9(),␣-]//g"         | # Drop accents (e.g., Glänzel -> Glnzel)
  grep -Pho "$RE_EPONYMS"              | # Search for eponyms
  grep -Pv  "$RE_NO_EPONYMS"           | # Skip non-eponymic
  grep -vif nonEponymicAdjectives.txt  | # Skip non-eponymic adj. (e.g., Belgian)
  sort                                 | # Sort eponyms (mandatory before uniq)
  uniq -i >> $EXTRACTED_EPONYMS          # Drop duplicated entries & append to file
done

# Rank eponymic expressions wrt their number of occurrences in distinct articles
cat $EXTRACTED_EPONYMS           | # Output all eponimic expressiojns
sed -r "s/[^A-z]+/␣/g;s/␣s␣/␣/g" | # Only keep letters (Pearson's = Pearson)
sort                             | # Sort (mandatory before uniq)
uniq -ic                         | # Drop duplicates & count number of occurrences
sort -rk 1                         # Rank the list of eponymic expressions found

rm $EXTRACTED_EPONYMS              # Remove temporary file
```

References

APA. (2010). *Publication manual of the American Psychological Association*. Washington, DC: American Psychological Association.

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature, 436*(7053), 900. doi:10.1038/436900a.

Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics, 69*(1), 161–168. doi:10.1007/s11192-006-0146-5.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: A review. *Journal of Informetrics, 2*(1), 1–52. doi:10.1016/j.joi.2007.11.001.

Beaver, D. d. (1976). Reflections on the natural history of eponymy and scientific law. *Social Studies of Science, 6*(1), 89–98. doi:10.1177/030631277600600105.

Beck, M. T., Dubrov, G. M., Garfield, E., & de Solla Price, D. (1978). Editorial statements. *Scientometrics, 1*(1), 3–8. doi:10.1007/BF02016836.

Boring, E. G. (1964). Eponym as placebo. *Acta Psychologica, 23*, 9–23. doi:10.1016/0001-6918(64)90072-1.

Braun, T. (2004). Editorial foreword. *Scientometrics, 60*(1), 9. doi:10.1023/B:SCIE.0000027301.70522.4c.

Braun, T., & Klein, A. (1992). Shpol'skii fluorimetry: The anatomy of an eponym. *Trends in Analytical Chemistry, 11*(6), 200–202. doi:10.1016/0165-9936(92)80042-5.
Braun, T., & Pálos, A. (1989). Textbook trails of eponymic knowledge in analytical chemistry. *Trends in Analytical Chemistry, 8*(5), 158–160. doi:10.1016/0165-9936(89)85033-2.
Braun, T., & Pálos, A. (1990). The name of the game is fame: Eponyms and eponymy in chemistry. *New Journal of Chemistry, 14*(8–9), 595–597.
Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals [Letter]. *The Scientist, 19*(22), 8.
Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals [Short communication]. *Scientometrics, 69*(1), 169–173. doi:10.1007/s11192-006-0147-4.
Braun, T., Glänzel, W., & Schubert, A. (2010). The footmarks of Eugene Garfield in the journal *Scientometrics*. *Annals of Library and Information Studies, 57*(3), 177–183.
Cintas, P. (2004). The road to chemical names and eponyms: Discovery, priority, and credit. *Angewandte Chemie International Edition, 43*(44), 5888–5894. doi:10.1002/anie.200330074.
Cole, J. R., & Cole, S. (1972). The Ortega Hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. *Science, 178*(4059), 368–375. doi:10.1126/science.178.4059.368.
De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Lanham: Scarecrow.
Diodato, V. (1984). Eponyms and citations in the literature of psychology and mathematics. *Library & Information Science Research, 6*(4), 383–405.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics, 69*(1), 121–129. doi:10.1007/s11192-006-0143-8.
Fernández-Cano, A., & Fernández-Guerrero, I. M. (2003). Eponymy for research evaluation: Spanish cases from the educational field. *Research Evaluation, 12*(3), 197–203. doi:10.3152/147154403781776591.
Freeman, M. S. (1997). *A new dictionary of eponyms*. New York: Oxford University Press.
Friedl, J. E. F. (2006). *Mastering regular expressions, 3rd edn*. Sebastopol: O'Reilly.
Garfield, E. (1965). Can citation indexing be automated?. In M. E. Stevens, V. E. Giuliano, L. B. Heilprin (eds.), *Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation* (pp. 189–192). Washington, DC: National Bureau of Standards. Miscellaneous Publication 269.
Garfield, E. (1973). Uncitedness III—the importance of *Not* being cited. *Current Contents, 8*, 5–6.
Garfield, E. (1983). What's in a name? The eponymic route to immortality. *Current Contents, 47*, 5–16.
Haines, D. E., & Olry, R. (2003). "James Parkinson did not die of his own personal disease … he died of a stroke" eponyms: Possessive or nonpossessive?. *Journal of the History of the Neurosciences, 12*(3), 305–307. doi:10.1076/jhin.12.3.305.16678.
Hargens, L. L. (2004). What is Mertonian sociology of science? *Scientometrics, 60*(1), 63–70. doi:10.1023/B:SCIE.0000027309.30756.6c.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16,569–16,572. doi:10.1073/pnas.0507655102.
Jana, N., Barik, S., & Arora, N. (2009). Current use of medical eponyms—a need for global uniformity in scientific publications. *BMC Medical Research Methodology, 9*(1), 18. doi:10.1186/1471-2288-9-18.
Kay, H. E. M. (1973). In praise of eponyms [Points of view]. *The Lancet, 302*(7840), 1256. doi:10.1016/S0140-6736(73)90988-4.
Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundation and Trends in Information Retrieval, 3*(1–2), 1–224. doi:10.1561/1500000012.
Kennedy, H. C. (1972). Who discovered Boyer's Law?. *The American Mathematical Monthly, 79*(1), 66–67. doi:10.2307/2978134.
Kotz, S., Balakrishnan, N., Read, C. B., & Vidakovic, B. (eds.) (2005). *Encyclopedia of statistical sciences, 2nd edn*. New York: Wiley. doi:10.1002/0471667196.
Lease, M., & Yilmaz, E. (2013). Crowdsourcing for information retrieval: Introduction to the special issue. *Information Retrieval, 16*(2), 91–100. doi:10.1007/s10791-013-9222-7.
MacAskill, M. R., & Anderson, T. J. (2013). Whose name is it anyway? Varying patterns of possessive usage in eponymous neurodegenerative diseases. *PeerJ, 1*, e67. doi:10.7717/peerj.67.
McCain, K. W. (2011). Eponymy and obliteration by incorporation: The case of the "Nash Equilibrium". *Journal of the American Society for Information Science and Technology, 62*(7), 1412–1424. doi:10.1002/asi.21536.
McCain, K. W. (2012). Assessing obliteration by incorporation: Issues and caveats. *Journal of the American Society for Information Science and Technology, 63*(11), 2129–2139. doi:10.1002/asi.22719.
Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology, 1*(1), 115–126. doi:2027/mdp.39015008014428.

Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review, 22*(6), 635–659. doi:10.2307/2089193.
Merton, R. K. (1965). *On the shoulders of giants. A Shandean postscript*. New York: Free.
Merton, R. K. (1968a). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. doi:10.1126/science.159.3810.56.
Merton, R. K. (1968b). *Social theory and social structure*. New York: Free.
Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis, 79*(4), 606–623.
Nature (2012). Gold in the text? [Editorial]. *Nature, 483*(7388), 124. doi:10.1038/483124a.
Nečas P., & Hejna, P. (2012). Eponyms in forensic pathology. *Forensic Science, Medicine, and Pathology, 8*(4), 395–401. doi:10.1007/s12024-012-9328-z.
Panush, R. S., Wallace, D. J., Dorff, R. E. N., & Engleman, E. P. (2007). Retraction of the suggestion to use the term "Reiters syndrome" sixty-five years later: The legacy of Reiter, a war criminal, should not be eponymic honor but rather condemnation. *Arthritis and Rheumatism, 56*(2), 693–694. doi:10.1002/art.22374.
van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502. doi:10.1556/Scient.67.2006.3.10.
Robertson, M. G. (1972). Fame is the spur the clear eponym doth raise [Letter to the editor]. *Journal of the American Medical Association, 221*(11), 1278. doi:10.1001/jama.1972.03200240052019.
Roeckelein, J. E. (1972). Eponymy in psychology. *American Psychologist, 27*(7), 657–659. doi:10.1037/h0033259.
Roeckelein, J. E. (1974). Contributions to the history of psychology: XVI. Eponymy in psychology: Early versus recent textbooks. *Psychological Reports, 34*(2), 427–432. doi:10.2466/pr0.1974.34.2.427.
Roeckelein, J. E. (1995). Naming in psychology: Analyses of citation counts and eponyms. *Psychological Reports, 77*(1), 163–174. doi:10.2466/pr0.1995.77.1.163.
Rousseau, R., García-Zorita, C., & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics, 7*(2), 294–300. doi:10.1016/j.joi.2012.11.012.
Ruffner, J. A. (eds.) (1977). *Eponyms dictionaries index*. Detroit: Gale Research.
Schreiber, M., Malesios, C., & Psarakis, S. (2012). Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators. *Journal of Informetrics, 6*(3), 347–358. doi:10.1016/j.joi.2012.02.001.
Shanahan, F., Houlihan, C., & Marks, J. C. (2013). In praise of the literary eponym—Henry V sign. *Quarterly Journal of Medicine, 106*(1), 93–94. doi:10.1093/qjmed/hcs210.
Simonton, D. K. (1984). Leaders as eponyms: Individual and situational determinants of ruler eminence. *Journal of Personality, 52*(1), 1–21. doi:10.1111/j.1467-6494.1984.tb00546.x.
Smith, D. W. (1975). Classification, nomenclature, and naming of morphologic defects. *The Journal of Pediatrics, 87*(1), 162–164. doi:10.1016/S0022-3476(75)80111-9.
de Solla Price, D., & Gürsey, S. (1975). Studies in scientometrics I: Transience and continuance in scientific authorship. *Ciência da Informação, 4*(1), 27–40.
Stigler, S. M. (1980). Stigler's law of eponymy. In T. F. Gieryn (Eds.) *Transactions of the New York Academy of Sciences, vol 39* (pp 147–157). doi:10.1111/j.2164-0947.1980.tb02775.x, Robert K. Merton Festschrift Volume.
Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science, 4*(2),73–79. doi:10.1214/ss/1177012580.
Száva-Kováts, E. (1994). Non-Indexed Eponymal Citedness (NIEC): First fact-finding examination of a phenomenon of scientific literature. *Journal of Information Science, 20*(1), 55–70. doi:10.1177/016555159402000107.
Thomas, K. S. (1992). The development of eponymy; A case study of the Southern blot. *Scientometrics, 24*(3), 405–417. doi:10.1007/BF02051038.
Trahair, R. C. S. (1994). *From Aristotelian to Reaganomics: A dictionary of eponyms with biographies in the social sciences*. Westport: Greenwood.
Van Noorden, R. (2012). Trouble at the text mine. *Nature, 483*(7388), 134–135. doi:10.1038/483134a.
Wallace, D. J., & Weisman, M. (2000). Should a war criminal be rewarded with eponymous distinction? The double life of Hans Reiter (1881–1969). *Journal of Clinical Rheumatology, 6*(1), 49–54.
Whitworth, J. A. (2007). Should eponyms be abandoned? No. *British Medical Journal, 335*(7617), 425. doi:10.1136/bmj.39308.380567.AD.
Woywodt, A., & Matteson, E. (2007). Should eponyms be abandoned? Yes. *British Medical Journal, 335*(7617), 424. doi:10.1136/bmj.39308.342639.AD.

Zhang, C., & Liu, X. (2011). Review of James Hartley's research on structured abstracts. *Journal of Information Science, 37*(6), 570–576. doi:10.1177/0165551511420217.
Zusne, L. (1987). *Eponyms in psychology: A dictionary and biographical sourcebook*. New York, Westport, and London: Greenwood.