Scientometrics (2015) 103:435–451
DOI 10.1007/s11192-015-1546-1

Nano language and distribution of article title terms according to power laws

Tomaz Bartol · Karmen Stopar

Received: 22 May 2014 / Published online: 28 February 2015
© Akadémiai Kiadó, Budapest, Hungary 2015

Abstract Scientometric evaluation of nanoscience/nanotechnology requires complex search strategies and lengthy queries which retrieve massive amount of information. In order to offer some insight based on the most frequently occurring terms our research focused on a limited amount of data, collected on uniform principles. The prefix nano comes about in many different compound words thus offering a possibility for such assessment. The aim is to identify the scatter of nanoconcepts, among and within journals, as well as more generally, in the Web of Science (WOS). Ten principal journals were identified along with all unique nanoterms in article titles. Such terms occur on average in half of all titles. Terms were thoroughly investigated and mapped by lemmatization or stemming to the appropriate roots—nanoconcepts. The scatter of concepts follows the characteristics of power laws, especially Zipf's law, exhibiting clear inversely proportional relationship between rank and frequency. The same three nanoconcepts are most frequently occurring in as many as seven journals. Two concepts occupy the first and the second rank in six journals. The same six concepts are the most frequently occurring in ten journals as well as full WOS database, representing almost two thirds of all nanotitled articles, in both instances. Subject categories don't play a decisive role. Frequency falls progressively, quickly producing a long tail of rare concepts. Drop is almost linear on the log scale. The existence of hundreds of different closed-form compound nanoterms has consequences for the retrieval on the Internet search engines (e.g. Google Scholar) which do not permit truncation.

Keywords Nanoscience · Bibliometrics · Lexical analysis · Power laws · Terminology · Subject categories · Search strategy · Compound words

Introduction

The field of nanoscience and nanotechnology has experienced a prolific growth of information in recent years. Many authors have attempted to assess the field by testing complex

T. Bartol (✉) · K. Stopar
Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, Ljubljana, Slovenia
e-mail: tomaz.bartol@bf.uni-lj.si

search strategies. Such composite queries involved many different terms, frequently requiring several stages of retrieval and different combinations of preceding queries. Researchers in the nanofield have usually been involved in order to outline the essential terminology. Such strategies are complex and not easy to replicate in different information systems. In addition to the many different terms the searches also involve representations of substances such as chemical compounds and formulae which present an additional problem in the retrieval. There is no universal opinion as to the description and definitions of this field. In addition, by including more and more terms to a query the quantity of articles grows to such magnitude that it becomes almost impossible to conduct a more detailed evaluation of some more particular topics. Transparency is quickly lost. It seems impossible to retrieve "all" relevant documents using search terms alone.

We used a more generalized methodology, which although simplified, should achieve good precision and recall and provide useful information on selected characteristics of some typical terms assembled on a basis of a consistent common denominator. We investigated terms that contain the prefix (word stem) *nano.* These terms result in high numbers even though they are not assigned to all relevant articles. In complex queries, the truncated nano retrieves between 70 and 90 % of relevant articles. These terms can serve as a streamlined measure for exploratory assessment of the occurrence of information in this field.

The ubiquitous *nano-* comes about in a variety of different grammatical forms and compound words—especially as a prefix appended to more specific concepts. Such combinations seem quite inexhaustible, building a long list of unique terms. We investigated the occurrence of such terms to discover which terms are the most frequent and what is their distribution in terms of ranking. If nanoterms occur with a consistently high frequency in article titles, the journal most likely represents an important source of information in this field. This serves as a basis for a more thorough and consistent evaluation of the terms.

To accurately determine the occurrence, each concept needs to be mapped to the common denominator. Once that has been established then particular patterns and specifics both in the ranks of these terms as well as frequency in specific journals as well as more generally will be determined. Our assumption is that just a few highly occurring terms account for a major share of all articles, in line with power laws which are typical of many information systems. Perhaps, few principal journals can reveal some consistent and informative patterns. We look to see if these patterns are reflected in the journal classification schemes. Finally, based on information derived from selected journals, we hope to identify the most frequent terms and their ranks in a more comprehensive sense that can serve as a model for this field in general such as in the global information systems.

Many articles relate research to the bibliometric (scientometric) aspects of nanoscience and nanotechnology. Different approaches are employed, such as based on lexical analysis, citations, and combinations of both. Authors, institutions, countries, publications (journals), different document types (articles, patents) were explored. In our analysis, we investigated the lexical characteristics so more emphasis will be placed on this approach. Different search fields were used by authors such as document title or journal title, and topics more in general (based on words in titles, abstracts and keyword fields). Complex Boolean search strategies were constructed, frequently developed on results by predecessors. Experts in the nano domain had been consulted in the construction of queries. The query by Noyons et al. (2003) included the truncated *nano* word root, and also contained several additional related terms. The authors investigated publication trends as well as countries and institutions. This search strategy was further expanded by Heinze (2004) and Heinze et al. (2007). Warris (2004) created a complex search strategy with a

comprehensive list of keywords for an assessment of a country's capability in nanotechnology. This strategy was also employed by some other authors. Calero et al. (2006) investigated research groups, employing selected terms for the identification of core publications. Several stages and possible traps in the lexical investigation of nanofield, such as the issues of natural language, definition of the field and construction of queries were addressed by Zitt and Bassecoulard (2006). Bassecoulard et al. (2007) tackled both lexical and citation characteristics. Citations patterns, and various links to several dozen nano-relevant journals, with the view of delineating a specific nano-set of journals were investigated by Leydesdorff and Zhou (2007). Principal authors, journals, and countries were explored in WOS by Kostoff et al. (2006) who employed complex search queries. An even more complex and inclusive query was constructed by Mogoutov and Kahane (2007) who also designated eight distinct subfields. Such queries retrieved a vast quantity of documents. Query by Porter et al. 2008 retrieved as much as 4.1 % of the total WOS database. A huge proportion of nanostudies was also reported by Grieneisen and Zhang (2011) who used the Topics field (TS = title, abstract and keywords) in WOS. A more simplified and more easily replicable query which, however, provides good precision and recall according to the authors, was proposed by Maghrebi et al. (2011). Porter's query was also employed by Wang et al. (2013) who co-linked top nano keywords with specific vocabularies and Milojević (2012) who identified title words in nano articles in order to cluster disciplinary components. Both papers identified nanoparticles as the most frequent among the nano-prefixed words. Principal journals in a more specialized field of nanocatalysis were ranked by Zibareva et al. (2014). A comprehensive review by Huang et al. (2011) evaluated and compared many different preceding search strategies, including the truncated nano-only, with regard to the identification of core nano journals. The review shows that each complex search strategy identifies different sets and ranks of core journals.

Other authors decided on a more hands-on approach that can be more easily tested in different information systems by employing a strategy based principally on the prefix nano in the article titles, excluding a few nano-terms which are not related to the field (Braun et al. 1997), Meyer and Persson (1998), Marinova and McAleer (2003), Guan and Ma (2007) although some important papers may not be retrieved by this method (Glänzel et al. 2003). Methodology based on the prefix nano was also employed in our research given that we were interested in identifying the characteristics of rank and frequency of only those terms that are based on the word root nano. In fact, nano can be prefixed to almost any other term to build compound words (Baird et al. 2004). Inferring from the complex queries in the previous paragraph, the truncated nano alone already retrieves between 70 and 90 % of all documents. At this point it is worth noting that many nano scientists are even not aware of truncation possibilities while searching for information in databases (Shiri 2011).

Since stemming and the closely related concept of lemmatization represent a complex field in computational linguistics, we offer only a very basic review of that research which is also related to nanosciences. In our research, we conducted manual identification of applicable nano terms in order to ascertain the frequency of terms that can be mapped to the same concept or idea. Milojević (2012) relied on computer lemmatization. Automatic lemmatization was used for the identification of nanotechnology subjects in national press (Veltri 2012). Lemmatization was also employed by Mogoutov (2007). Depending on the nature of research, some author's did not employ this procedure, for example in the identification of top technical words (Small 2011). The issue of stemming was also addressed in a more loose connection with nanosciences by Magerman et al. (2010). Some authors who also refer to lemmatization suggest the use of the expression 'hybrid word

family', i.e. words which are centred on an idea, as is also the example of nano-derived words (Thelwall and Price 2006).

Classical power laws (for example Bradford, Lotka, Zipf, Pareto) analyse the rank and frequency of events, in such cases when the occurrence of an event is inversely proportional to its rank, indicating that, on one hand, few selected events occur frequently whereas on the other hand most events occur rarely. A number of articles have been published in this specialized informetric field, directly linked to linguistics and events outside the domain of natural and technical sciences. For this reason we present only selected citations which have an association with our research. Power-law statistical distributions can be seen in a wide variety of natural and manmade phenomena (Newman 2005) and are characteristic of scientific networks Zitt and Bassecoulard (2006). Zipf's law (Zipf 1949), for example, deals with the rank and frequency of words in natural languages. The occurrence of words as outlined by Zipf's law has been mentioned in the context of nanotechnology by Tsuda et al. (2006) although the authors have not related these laws to terminology but rather to authorship networks. The applications of power laws (including Zipf's law) as well as nano sciences were reviewed in a wider context by Bar-Ilan (2008). Power laws (Lotka) were also employed by Milojević (2010) in the context of the number of papers per authors in the NanoBank database. The context of nanotechnology is referred to by Mogoutov et al. (2008) although the article's topic addresses the Zipf's law more in the context of biomedical terminology. Nanotechnology is presented as a field which produces a large number of scientific papers, thereby offering a good possibility for an analysis of language properties according to linguistic laws, such as Zipf's (Turenne 2010). A multi-word "nano materials" was singled out for testing along with some other words by Zhang et al. (2009) in the context of the Zipf's law as related to the assessment of new "hot topics" in technology. Zipf power laws have also been observed in some other science-and-technology fields, most notably chemistry. Distribution of chemical compounds and molecular representations was investigated by Lipkus et al. (2008), Benz et al. (2008), Karakoc et al. (2006). Holliday et al. (2011) also employed such laws in information retrieval (similarity searches) where a few observations (i.e. molecules) occur very frequently and the great majority occur only once. Yan et al. (2013) conclude that Zipf's law holds for the chemical language just as it does for many natural languages. Even though such laws are often presented as separate laws, they are frequently just different ways of looking at the same thing Adamic (2000).

Materials and methods

We selected the journals on consistent principles, identifying ten journals where the truncated nanoterms are present in the highest number of article titles. Systematic exclusion of non-relevant terms, such as *nanoseconds*, *nano2*, *nano3*, was not necessary since we investigated only nano-relevant journals and have in our more detailed analysis included only terms that come about with a higher frequency. Essentially, these terms denote the context of *nano* given their presence in nano-journals although some terms may hold a different connotation in other contexts, for example *nanometer*.

The truncated *nano* enables the retrieval of many different compound terms, for example in a closed form (nanoscience), hyphenated form (nano-science) as well as open form (nano science). In the continuation of the analysis we put special emphasis on closed-form nanoterms. This has a major impact on retrieval precision and recall in systems that do not offer the utility of right-hand truncation.

For evaluation purposes, we used Web of Science (WOS) Core Collection (previously better known as Citation Indexes: SCI-EXPANDED, SSCI, A&HCI). In this database we identified the ten journals that returned the highest number of articles containing a nanoterm in article titles (Table 1). We included the journals regardless of the WOS classification with the applicable *Nanoscience Nanotechnology* category (capitalized in Table 1) as four among the ten journals that contain the highest number of nanoarticles are not classified in those categories. In order to give our subsequent results some reference, we present this exploratory part of our analysis in this methodological section. Journal abbreviations in Table 1 are used in all subsequent figures and tables. We then refer to these data in the next section which presents the specific results.

The WOS category *Materials Science Multidisciplinary* is assigned to eight journals. In two journals (*Applied Physics Letters* and *Journal of Applied Physics*) both of which are not classified *nano-* nor *materials*-categories, 20 % of all article titles contain a nanoterm. Between one third and two thirds of articles contain at least one nanoterm in the title. The *Journal Of Materials Chemistry* contains the highest number of nanoarticles (1355) among all journals. Almost half of the articles (3322) contain a nanoterm. This journal, however, is not classified with the applicable *Nanoscience Nanotechnology* Category.

The analysis was performed on 2012 data. We downloaded all articles containing nanoterms in a separate experimental database. We identified all nanoterms per journal and conducted further analysis on those terms. Altogether, we identified 565 unique terms in all of the ten journals combined. As shown in Table 1, the journals under study published between 695 and 1355 nanoterm articles (9000 in total, in 2012) offering a good source for a more generalized information on the most frequent terms. At this stage of the analysis, we counted all nanoterms as distinctive occurrences. Singular, plural and other forms were counted as separate terms. The exact term *nano* (occurring in an open form, and also as hyphenated nano-, nano/etc.) occurs only 470-times in 9000 nanoarticles. Only a handful of additional articles could be retrieved with left-hand truncated terms so these were not included in the analysis. However, such a way of constructing the terms can also be followed for a possible identification of concepts–if the numbers of such compound terms become more significant.

In the next step, we identified the concepts by conflating the same-family terms based on a shared word root. We mapped all possible variants (typically nouns, verbs, adjectives) to the shared root–by removing the suffixes. First we sorted the terms alphabetically in order to identify the common shared concepts. The terms such as *nanostructure, nanostructures, nanostructural, nanostructured, nanostructuring,* for example, were consequently mapped to one term and truncated in the appropriate place (e.g. *nanostructur**).

The above preparatory part of the experiment served as basis for the subsequent analysis of the distribution of frequency and rank of such concepts. The relation between the frequency (occurrence) of concepts and their rank is a well-known topic in information science so we present only the general outline. Such a relationship is generally tested by the power laws, most notably the Pareto's law, and Zipf's law which are closely related (Adamic 2000). Our topic is applicable to the Zipf's law. Namely, Zipf investigated the occurrence or frequency of words in English texts in relation to the rank of the words. Zipf's law states that the frequency of a word is inversely proportional to its rank. Also, the most common term occurs, approximately, twice as frequently as the second most common term, etc. According to this law, the frequently occurring terms quickly subside followed by a long tail of terms which are rarely used. It has also been observed that power law's such as Zipf's can also occur for wholly novel words (Piantadosi 2014).

Table 1 Journals with the highest occurrence of nanoterms in article titles in 2012 (ti = nano*), all articles in this year, Web of Science categories of respective journals, and different nanoterms in article titles

Abbr.	Journal title	WOS categories	ti = nano*	All articles	Nanoterms
JMC	J. of Materials Chemistry	Chemistry Physical, Materials Sci. Multidiscip	1355	3322	181
JPC C	J. of Physical Chemistry C	Chemistry Physical, Materials Sci. Multidiscip., Nanosci. Nanotech.	1151	3342	178
APL	Applied Physics Letters	Physics Applied	1072	5102	182
JNN	J. of Nanoscience and Nanotechnology	Chemistry Multidiscip., Materials Sci. Multidiscip., Nanosci. Nanotech., Physics Applied, Physics Condensed Matter	886	1529	123
JAP	J. of Applied Physics	Physics Applied	870	4447	145
ACSN	ACS Nano	Chemistry Multidiscip., Chemistry Physical, Materials Sci. Multidiscip., Nanosci. Nanotech.	765	1247	176
NL	Nano Letters	Chemistry Multidiscip., Chemistry Physical, Materials Sci. Multidiscip., Nanosci. Nanotech., Physics Applied, Physics, Condensed Matter	701	1099	136
NNT	Nanotechnology	Materials Sci. Multidiscip., Nanosci. Nanotech., Physics Applied	748	1049	164
NNS	Nanoscale	Chemistry Multidiscip., Materials Sci. Multidiscip., Nanosci. Nanotech., Physics Applied	737	1032	140
ML	Materials Letters	Materials Sci. Multidiscip., Physics Applied	695	1634	104

We have decided to test the applicability of this rule in the context of the terms which begin with the prefix nano. Namely, we have previously observed that a few nanoterms occur frequently whereas most other terms rarely occur. Such distribution gave the impression that these terms might indeed follow some distinctive patterns, possibly generating specific scientific language or terminology where a few frequently used terms account for the majority of all uses. In order to assess the nature of this distribution more thoroughly we needed to calculate the rank of highly occurring terms and produce a figure which may reflect a possible distribution according to power laws. Zipf's distribution is usually presented as a curve. But not all curves are indicative of power laws. Only if a curve is nearly linear on a log scale can it be deduced that the frequency and rank distribution follow a power law such as Zipf's.

In order to ascertain the rank and frequency of principal nanoconcepts in the scope of the Zipf's law we first investigated the distribution of such concepts in each of the ten journals. In each journal we analysed, more particularly, the first three most frequently

occurring concepts. We also calculated the share of articles which could be retrieved with these concepts in each particular journal.

In the continuation, we investigated the position of nanoconcepts in a broader sense–by calculating the rank and frequency of these terms in all ten journals as a whole. This information served as a basis for the subsequent identification and assessment of these concepts in the WOS database more in general. In order to double-check the correctness of search results we searched in the WOS database by merging all ten journals into a single search (Boolean OR) verifying a long list of (truncated) concepts. By merging all items in a combined procedure we were able to identify the most frequently occurring concepts as corrected for a possible journal bias.

The concluding part of the analysis was performed on the whole WOS presenting the position of each most frequent concept as inversely proportional to its rank. This result is also presented on a log scale in order to better reflect the log-linear nature of these concepts according to the Zipfian distribution.

With a view to better describe the above procedures we present a summarised outline of the study:

(A) Preparatory procedures	(B) Analysing rank and frequency of nanoconcepts
(A.1) *Selection of materials*	(B.1) *Determining major nanoconcepts in each journal*
Selection of a database and identification of ten journals with the highest occurrence of nano-termed words (based on nano*) in article titles	Identifying the uppermost nanoconcepts in each journal Estimating the shares of the uppermost nanoconcepts
(A.2) *Establishing terms and concepts*	(B.2) *Identifying the rank and frequency of the principal nanoconcepts in a wider sense*
Extraction of unique nanoterms (by each journal) onto a separate list and assessing all such terms in each journal	Comparing the position of the principal nanoconcepts in the cumulative total of the ten journals with the position of these concepts in the whole WOS database
Characterisation of nanoconcepts (mapping the 'same-family' nanoterms to the common root, for example: n-structure(s), n-structured, n-structural, n-structuring–>n-structur*)	Visualisation of descending rank and frequency of nanoconcepts on a log scale in WOS

Results

Our initial exploratory analysis found that in 2012 there were as many as 2700 source titles (journals) in WOS which published documents containing a nanoterm in the title. Among the total of 2700 source titles, only 50 journals published more than 200 articles in 2012. The first ten most productive journals published almost 9000 nanoarticles, accounting for some 20 % of the total of articles in the WOS in this year. In order to verify the relevance of these journals we have also run a test search on the same principles in the Scopus database. The first ten most highly ranked journals were the same in Scopus as in WOS.

We identified 565 different unique terms beginning with a word root *nano*. An occasional term contained a spelling mistake, such as *nanoparticies* (instead of *nanoparticles*). This was the exact form as occurring in the original WOS. These terms are only retrievable as such so we did not conduct any 'corrections'. In any case, these single errors had no significant role among the almost 9000 titles containing a nanoterm. Article titles in the

Journal Applied Physics Letters comprise as many 182 different nanoterms, followed by the *Journal of Materials Chemistry* (*181 term*), and the *Journal of Physical Chemistry C* with 178 different terms. The last journal on the list—*Materials Letters*—comprises 104 different terms. The fact that as many as 565 different terms can be found in the total of the ten journals indicates that many terms occur very rarely. Dozens of terms come about only once, for example *nanocalorimetry, nanoionics, nanofoam, nanobionics.*

Figure 1 presents the results for the three most frequent concepts in each journal which were ascertained on the basis of the truncated word stems. For the purposes of practicality, we will now refer to these concepts in the plural form.

There is a fairy high similarity in the use of concepts among the journals. N*anoparticles* as well as *nanotubes* are always represented. The place of *nanoparticles* is not surprising as this is a very general term. *Nanoparticles* thus occupy the first position in seven journals. *Nanotubes* most frequently occupy the second place (also in seven journals). *Nanowires*, which come about on seven instances, occupy the first position three times. The only exception to this pattern are *nanocomposites*, *nanocrystals* and *nanostructures*, each occurring only once—rank three. The comparison of total numbers among the journals must be viewed in relation to the number of all relevant articles in a particular journal which is provided in the Table 1. However, the distribution of ranks within each journal is similar for all journals, and respective concepts. This is evident in Fig. 2.

What is worth noting is that there is no difference between the ranks and frequency in the WOS *Nanoscience Nanotechnology* journals [indicated by a single letter (*n*) in parentheses after the name of the journal in the legend in Fig. 1] and journals which are not classified with this category. Also, the other WOS categories don't seem to have much influence either. For example, the journal *Materials Letters* which is classified with a *physics* category (but no *chemistry*) shows virtually the same rank-and-frequency patterns as *ACS Nano*, which is classified with *Chemistry Physical* (but no *physics* as such).

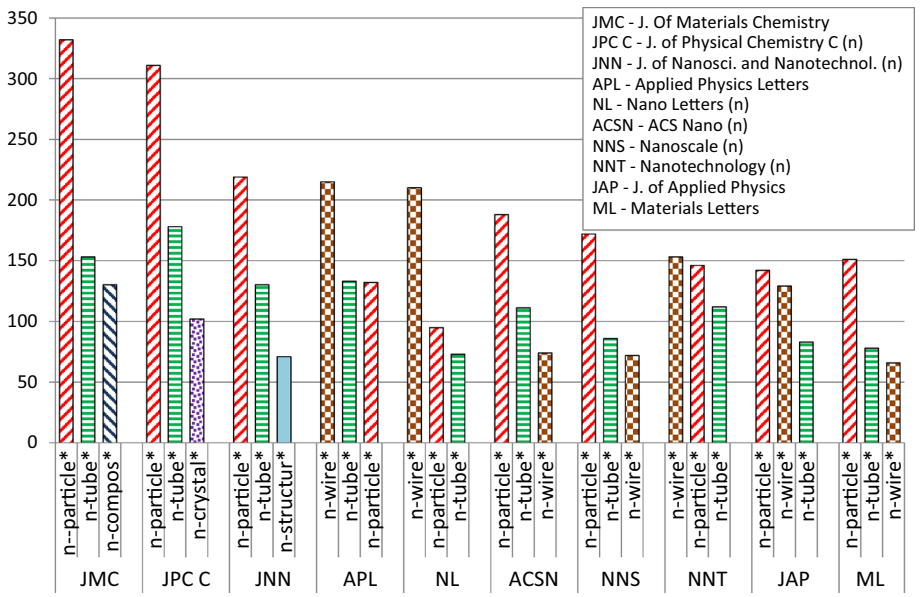

Fig. 1 Rank and frequency of the most frequent nanoconcepts in selected ten journals in 2012

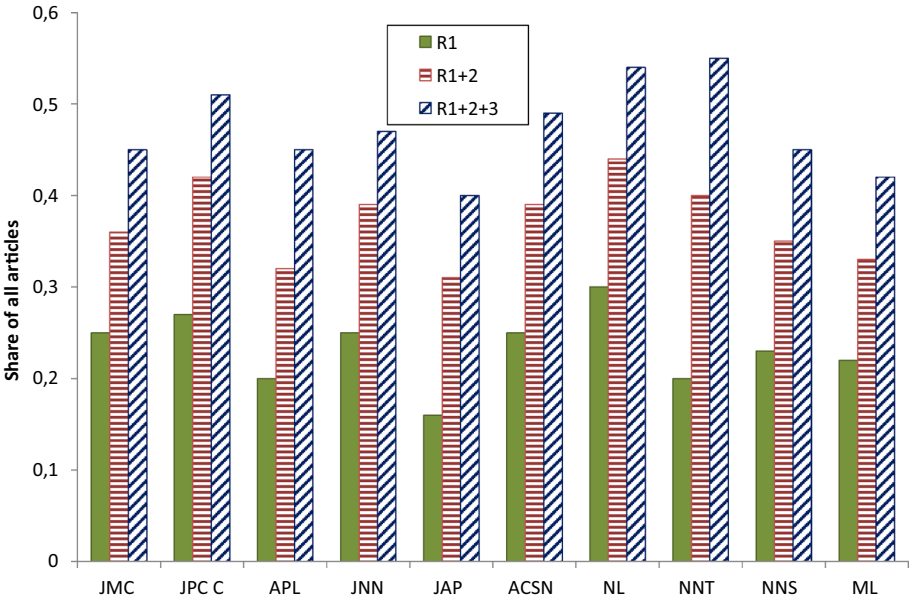

Fig. 2 Frequency of one (R1), two (R1 + 2, and three (R1 + 2 + 3) highest ranked nanoconcepts in article titles of ten principal journals as a share of all articles in 2012

Considering the number of nanoarticles in each journal (Table 1) and the most frequently occurring nanoconcepts (Fig. 1) it is evident that a few selected terms retrieve a substantial percent of all articles. For the three frequent concepts presented in Fig. 1 we also calculated the share of articles that can be retrieved using the first, first two, and first three highly ranked concepts (Fig. 2). The figure again shows clear similarities among the journals: a single leading concept (R1) retrieves 20 % of all articles. The three principal highest ranking concepts (R 1 + 2 + 3) retrieve, on average, half of all articles.

Second part of the experiment compared the position of nanoconcepts in ten journals and WOS. Altogether, some 57.000 records could be retrieved in WOS in 2012 using a truncated *nano** title term. Figure 3 represents the frequency of the same terms in the ten journals as well as in WOS. The concept of *nanoparticles* exhibits the highest occurrence in both cases. In WOS, 14,682 articles can be retrieved, and 1888 in the ten journals. We present the comparison of ranks and frequencies on the same scale in Fig. 3, by presenting the values for ten journals on the lower primary *x*-axis, and WOS on the upper secondary *x*-axis. The long "Zipfian tail" of infrequent concepts begins very soon. Therefore, in Fig. 3 we present only those terms which occur at least 40 times in the ten journals combined. We observe that the terms are positioned similarly, with an interesting exception of *nanowires.* Some less important difference can also be noticed in *nanocomposites*, and *nanomaterials* much further down the list. *Nanowires* are ranked very clearly on the second place in the ten journals. In WOS, they occupy the fifth place. This can be attributed to a stronger coverage of this particular subject in a few selected journals: *Applied Physics Letters, Nano Letters, Nanotechnology*, and *J. of Applied Physics*. This special concept seems to be less frequently represented in other not so "nano-oriented" journals. In the ten journals, the same leading six concepts account for as many as 65 % of all nanoarticles. This is very similar to WOS where the same concepts account for 63 % of all nanoarticles. The terms

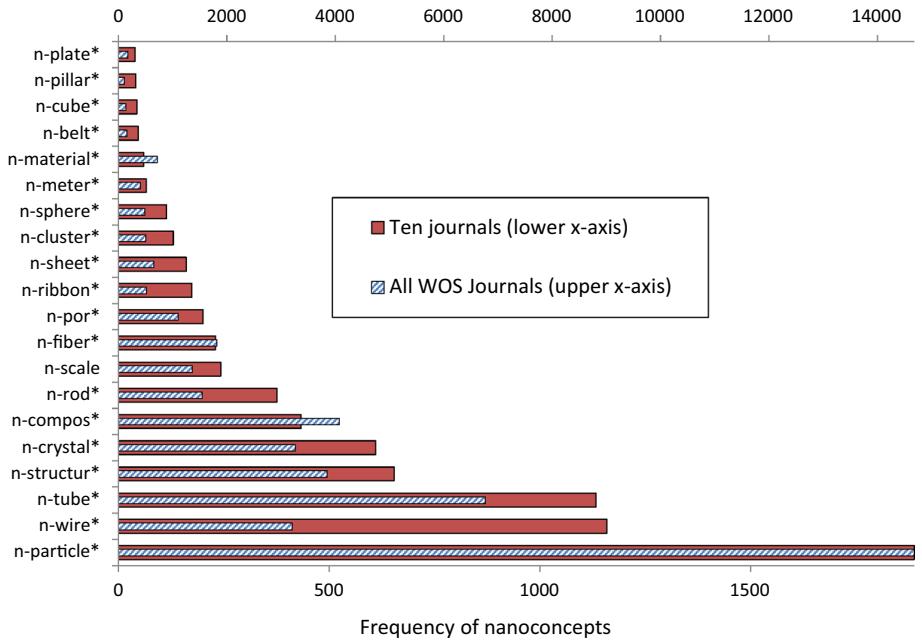

Fig. 3 Ranks of the most frequent nanoconcepts in article titles of ten journals (n) and WOS ($n \times 7.8$) in 2012

further down the list quickly drop in frequency, producing a long tail of rarely occurring terms, from *nanocalorimetry* to *nanowelding* and *nanoxerography*.

We also provide an extended list of frequently occurring concepts (Table 2) in the ten journals ranked in descending order, from nanoparticle* (1.888 records) to nanoflower* (15 records). As has been shown before, the first few concepts are by far the most important. The first six concepts retrieve almost two thirds of all nano-termed titles. Some of these concepts may also have meanings outside the scope of nanoscience and technology, however, all these concepts are occurring with a substantial frequency in nano journals thus indicating relevance in this field.

By identifying the most frequent nanoconcepts in ten major journals we were then able to identify the most frequently occurring concepts in the full WOS database by inferring that the major truncated terms in ten journals would invariably retrieve very frequent terms in WOS. This information thus served as a basis for the final part of our analysis—the identification of the rank and frequency of these terms in research articles more in general, taking WOS database to be a comprehensive global source of scientific information. We identified all articles in WOS with nanoconcepts occurring at least 50 times (Fig. 4). There were 38 concepts, ranking in a descending order—from *nanoparticles* (Rank 1:14,682 occurrences) to *nanomembranes* (Rank 38:50 occurrences). We again stemmed the terms belonging to the same concept in order to conflate all possible variants into the same root. Figure 4 shows these concepts as plotted by frequency against the rank (primary y-axis; left side of the chart). In order to provide additional information regarding the ranking according to power-laws we then calculated the log 10 rank-frequency for the same data set. This is shown on the secondary y-axis (right side of the chart). The relationship on the

Table 2 List of nano-truncated terms occurring most frequently in the ten journals

Nanoparticle*	Nanorod*	Nanocluster*	Nanopillar*	Nanohybrid*	Nanopowder*	Nanoindent*
Nanowire*	Nanoscale	Nanosphere*	Nanoplate*	Nanoplatelet*	Nanomechanic*	Nanoflower*
Nanotube*	Nanofiber*	Nanometer*	Nanopattern*	Nanoantenna*	Nanocapsule*	
Nanostructur*	Nanopor*	Nanomaterial*	Nanodot*	Nanoring*	Nanoshell*	
Nanocrystal*	Nanoribbon*	Nanobelt*	Nanoflake*	Nanoimprint*	Nanogap*	
Nanocompos*	Nanosheet*	Nanocube*	Nanosize*	Nanosecond*	Nanomembrane*	

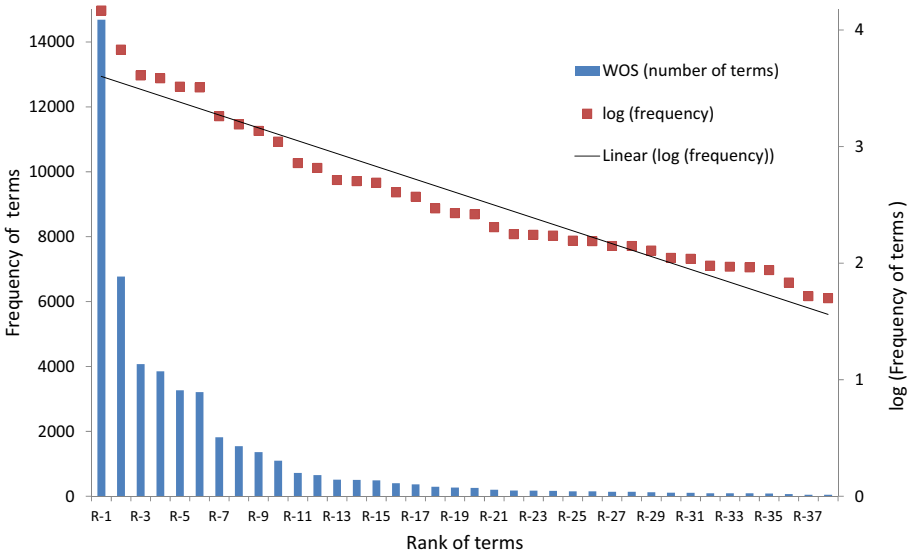

Fig. 4 Rank of 38 title-nanoterms in Web of Science (2012) ordered by frequency and on a log scale

log plot is almost linear thus exhibiting clear Zipfian characteristics. The more frequently occurring terms are soon followed by a long tail of rare concepts, many of which come about only once.

Discussion

Authors have systematically investigated many aspects of the ever expanding field of nanosciences by using complex queries, trying to capture as many nano-related articles as possible. Given the enormous scope of the field, however, this endeavour seems quite impossible. Our approach was to thoroughly evaluate only article title terms that stem from the prefix nano, in order to ascertain particular characteristics in the frequency and rank of these terms. The many different search strategies reviewed in this paper are certainly more comprehensive and return greater number of records. However, authors frequently acknowledge that there is no universal agreement on the choice of the most appropriate terms. Therefore, such complex search strategies return substantially dissimilar results. Sometimes it is also difficult to replicate queries which had been suggested. The queries include formulae and symbols which further complicate the searches. Namely, the retrieval depends on the way a particular information system harvests its data. On the other hand, the choice of nano-prefixed terms is quite transparent and can readily be reproduced. And transparency and reliability are an indispensable condition in nano search strategies (Mogoutov and Kahane 2007).

Four of the ten journals identified by our methods are currently not classified with the applicable *Nanoscience Nanotechnology* category. Other authors were also unable to delineate nano-journals clearly from other journals, such as disciplinary journals in chemistry and physics as categorized by the US Library of Congress classification system (Leydesdorff and Zhou 2007). The most productive journal in our query (*J. of Materials*

Chemistry) is not classified as a nanojournal in WOS although in this journal more than 40 % of all article titles contain a nanoterm. In the review of several different search strategies by Huang et al. (2011) based on 2006 data, and WOS *topic* field as opposed to our WOS *article title* field, this journal achieves the rank six in one strategy but lower ranks in other strategies. Our analysis thus also shows that the results do not only depend on the design of a query but are also only valid at the very time of an analysis. Strotmann and Zhao (2010) state that in scientometric studies, where Zipf-type exponential distributions are common, even small errors in the data can result in a significant error in an analysis. We suggest that the time of observation will also be essential.

Altogether, we identified as many as 565 different nanoterms in ten journals alone in 2012. In individual journals fewer terms were employed indicating that many terms were only used once. In these journals, as many as 9000 among the 24,000 articles contained such a term. The terms come about as a noun (singular or plural), or some other grammatical form. Working with large lexical corpora, some authors employed computer lemmatization (Milojević 2012) in order to identify distinct concepts. We chose to conduct a more accurate 'manual' lemmatization by conflating all interrelated terms into a single concept. Only then was it possible to ascertain the real rank and frequency of distinct topics (concepts) and not only mere unique terms. For example, *nanopattern* and *nanopatterns* account for less than 25 % of the truncated *nanopattern**. Most other terms come about in other forms (*nanopatterned, nanopatterning* …).

Three concepts (*nanoparticle-, nanotube-*, and *nanowire-*) are always present among the first three concepts in as many as seven journals. *Nanoparticle-* and *nanotube-* are present in all ten journals. Moreover, these two concepts occupy the first and the second respective rank in as many as six journals. Again, WOS category *Nanoscience Nanotechnology* doesn't seem to play a decisive role in this distribution. Neither do other categories: *chemistry*-classified journal *ACS Nano* and a *physics*-classified journal *Materials Letters* exhibit virtually the same ranks of the most frequent terms. The nano records are pretty evenly distributed across several different WOS-categories (Grieneisen and Zhang 2011). Some authors have thus used some other international classification schemes, for example by Frascati (in a study not related to nanosciences), in order to offset a possible bias in the delineation of fields of science (Bartol et al. 2014).

Some further specifics is revealed in our research: the position of concepts, in all journals, shows a clear inverse proportional relationship between the rank and frequency, pretty much in accordance with the Zipf's law which holds that a minority of frequently occurring words—*nanoconcepts* in our case, accounts for the majority of all occurrences in a text—article titles in our case. Indeed, as few as three major concepts account, roughly, for half of all nanotitled articles, in each journal. The frequency of concepts falls progressively, eventually producing a long tail of very rare concepts, many occurring only once. These results show a typical distribution of events governed by classical power laws, such as those by Lotka (referring to authors), Bradford (referring to publications), and in our case Zipf (referring to words), which are characteristic of scientific networks Zitt and Bassecoulard (2006). Our exploratory analysis and selection of journals revealed that the journals containing a nanoterm in article titles also show some typical characteristics of power laws (Bradford's law). This was outside the scope of this paper, however, so this observation may serve as material for our further research in the field.

Further examination of the most highly ranking concepts in the Boolean union of all ten principal journals reveals that the same six concepts are also the most frequently occurring concepts in the full WOS database, representing almost two thirds of all nanotitled articles. The only difference is found in the rank of *nanowire-*. This concept is ranked more highly

in the ten journals, indicating an important dominance of just a few journals in this particular subject. Again, no special role can be attributed to WOS classification. Two journals with the highest *nanowire-* occurrence are classified with the WOS *nano*-category and two are not. Most other concepts exhibit very similar ranks both in WOS and the ten journals. Among less frequently occurring concepts there is also *nanometer** which may also refer to a context unrelated to nano and was excluded by some other authors using correction in the queries. However, given that this term was derived from article titles in nano-journals we have nevertheless counted this term as well. In any event, the exclusion of this particular term would not change the linear relationships on the log scale.

In total, almost 57,000 articles in WOS contained a nano-truncated term in 2012. 14,682 thereof can be attributed to *nanoparticl-* (Rank 1) and 6773 to *nanotube-* (Rank 2). In the strict Zipfian sense the highly occurring words possess low semantic content (Melz et al. 2005). In our case, this "low semantic content" is evidenced by the concept of *nanoparticles* where the meaning is indeed fairly general so they logically occupy the highest rank among the terms. We identified and ranked all such nano concepts which occurred at least 50 times. The final log plot in our analysis demonstrates a manifest linear relationship, displaying very evident Zipfian characteristics of such "nanolanguage". This Zipfian trait was also observed in chemical language by Yan et al. (2013). The present analysis is based on 2012 data. It would be interesting, in further research, to also assess more specifically the individual terms over a longer period of time.

It is also necessary to comment on the role of the exact *nano* term which typically comes about as hyphenated or open-form term. Such representations account for much less than 10 % of all relevant occurrences. The vast majority of concepts is retrievable only in an exact closed form belonging to different grammatical classes. This should not be a problem in databases such as WOS which offer right-hand truncation although some authors report that many nano searchers make no use of truncation (Shiri 2011). Even experienced end-users, however, will face serious challenges on the Internet search engines, most notably Google Scholar, where hundreds of different compound terms will only be retrievable in an exact form. In addition, WOS generally offers the utility of a stemmer (lemmatization). This utility seems to ignore most compound nanoterms so caution must be exercised when mining for nano information—also in more standardized information retrieval systems and databases.

Conclusions

Nano terms which come about in article titles can serve as a good generalized indicator of the subject content. In many journals, almost half of all article-titles contain such a term. The terms come about in many different forms. The frequency of distinct concepts can only be ascertained by conflating all terms into a consistent concept. Such concepts exhibit a very obvious inverse proportional relationship between the rank and frequency, typical of power laws—most notably the Zipf's law: minority of frequently occurring nanoconcepts accounts for the majority of all occurrences. Moreover, such a relationship is almost linear on logarithmic scale. Among the principal journals, there is little difference in the patterns of rank and frequency of the most frequent concepts. The relationships do not seem to be strongly influenced by journal subject categories. Additionally, the relationship between the rank and frequency of the most frequent concepts in the ten principal journals is very similar to that in the entire WOS database. The vast majority of terms come about as nanoprefixed closed-form compound words. This can only be offset by an appropriate

truncation (wildcard). Such a representation of nanoconcepts has significant implications for the retrieval on the Internet scholarly search engines which offer no truncation possibilities. The Boolean OR cannot effectively compensate for this shortcoming as there exist many hundreds of such terms. The rules of the Zipf's law typically apply to language corpora so it seems that the field of nanosciences and nanotechnolgy is gradually building a veritable nanolanguage. Given the constant developments in the nanofield it is expected that the long tail of concepts will continue to grow, especially as nano can be appended to a seemingly indefinite number of terms. Observation of these ranks, over some period of time, may detect some new trends and developments in the field. The few most frequent terms, however, will probably keep a more stable rank, with a few exceptions perhaps.

Acknowledgments This work was supported by the Slovenian Research Agency, Research Programme P4-0085 (D).

References

Adamic, L. A. (2000). Zipf, power-laws, and pareto: A ranking tutorial. Xerox Palo Alto Research Center, Palo Alto. http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html. Accessed 20 April 2014.

Baird, D., Nordmann, A., & Schummer, J. (2004). Introduction. *Discovering the nanoscale* (pp. 1–8). Amsterdam: IOS Press.

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

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics, 98*(2), 1491–1504.

Bassecoulard, E., Lelu, A., & Zitt, M. (2007). Mapping nanosciences by citation flows: A preliminary analysis. *Scientometrics, 70*(3), 859–880.

Benz, R. W., Swamidass, S. J., & Baldi, P. (2008). Discovery of power-laws in chemical space. *Journal of Chemical Information and Modeling, 48*(6), 1138–1151.

Braun, T., Schubert, A., & Zsindely, S. (1997). Nanoscience and nanotecnology on the balance. *Scientometrics, 38*(2), 321–325.

Calero, C., Buter, R., Cabello Valdés, C., & Noyons, E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. *Scientometrics, 66*(2), 365–376.

Glänzel, W., Meyer, M., Du Plessis, M., Thijs, B., Magerman, T., Schlemmer, B., et al. (2003). *Nanotechnology: Analysis of an emerging domain of scientific and technological endeavour (Report).* Leuven: K.U. Leuven, Steunpunt O&O Statistieken.

Grieneisen, M. L., & Zhang, M. (2011). Nanoscience and nanotechnology: Evolving definitions and growing footprint on the scientific landscape. *Small (Weinheim an der Bergstrasse, Germany), 7*(20), 2836–2839.

Guan, J., & Ma, N. (2007). China's emerging presence in nanoscience and nanotechnology: A comparative bibliometric study of several nanoscience "giants". *Research Policy, 36*(6), 880–886.

Heinze, T. (2004). Nanoscience and nanotechnology in Europe: Analysis of publications and patent applications including comparisons with the United States. *Nanotechnology Law & Business, 1*(4), 427–447.

Heinze, T., Shapira, P., Senker, J., & Kuhlmann, S. (2007). Identifying creative research accomplishments: Methodology and results for nanotechnology and human genetics. *Scientometrics, 70*(1), 125–152.

Holliday, J. D., Kanoulas, E., Malim, N., & Willett, P. (2011). Multiple search methods for similarity-based virtual screening: Analysis of search overlap and precision. *Journal of Cheminformatics, 3*(1), 1–15.

Huang, C., Notten, A., & Rasters, N. (2011). Nanoscience and technology publications and patents: A review of social science studies and search strategies. *Journal of Technology Transfer, 36*(2), 145–172.

Karakoc, E., Sahinalp, S. C., & Cherkasov, A. (2006). Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *Journal of Chemical Information and Modeling, 46*(5), 2167–2182.

Kostoff, R. N., Lau, C. G. Y., Tolles, W. M., & Murday, J. S. (2006). The seminal literature of nanotechnology research. *Journal of Nanoparticle Research, 8*(2), 193–213.

Leydesdorff, L., & Zhou, P. (2007). Nanotechnology as a field of science: Its delineation in terms of journals and patents. *Scientometrics, 70*(3), 693–713.

Lipkus, A. H., Yuan, Q., Lucas, K. A., Funk, S. A., Bartelt, W. F., Schenck, R. J., & Trippe, A. J. (2008). Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *The Journal of Organic Chemistry, 73*(12), 4443–4451.

Magerman, T., Looy, B. V., & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics, 82*(2), 289–306.

Maghrebi, M., Abbasi, A., Amiri, S., Monsefi, R., & Harati, A. (2011). A collective and abridged lexical query for delineation of nanotechnology publications. *Scientometrics, 86*(1), 15–25.

Marinova, D., & McAleer, M. (2003). Nanotechnology strength indicators: International rankings based on US patents. *Nanotechnology, 14*(1), R1. doi:10.1088/0957-4484/14/1/201.

Melz, R., Biemann, C., Böhm, K., Heyer, G., & Schmidt, F. (2005). Real-time analysis of speech streams and their representation as conceptual structures. In *Proceedings of HCI-05. Las Vegas, Nevada, USA: HCI International.*

Meyer, M., & Persson, O. (1998). Nanotechnology-interdisciplinarity, patterns of collaboration and differences in application. *Scientometrics*, *42*(2).

Milojević, S. (2010). Power law distributions in information science: Making the case for logarithmic binning. *Journal of the American Society for Information Science and Technology, 61*(12), 2417–2425.

Milojević, S. (2012). Multidisciplinary cognitive content of nanoscience and nanotechnology. *Journal of Nanoparticle Research, 14*(1), 1–28.

Mogoutov, A., Cambrosio, A., Keating, P., & Mustar, P. (2008). Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics, 2*(4), 341–353.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy, 36*(6), 893–903.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351.

Noyons, E. C. M., Buter, R. K., van Raan, A. F., Schmoch, U., Heinze, S., Hinze, S., & Rangnow, R. (2003). *Mapping excellence in science and technology across Europe: Nanoscience and Nanotechnology* (Final report No. EC-PPN CT-2002-0001). Leiden: Leiden University.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 1–19. doi:10.3758/s13423-014-0585-6.

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research, 10*(5), 715–728.

Shiri, A. (2011). Revealing interdisciplinarity in nanoscience and technology queries: A transaction log analysis approach. *Knowledge Organization, 38*(2), 135–153.

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

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics, 4*(2), 194–200.

Thelwall, M., & Price, L. (2006). Language evolution and the spread of ideas on the Web: A procedure for identifying emergent hybrid word family members. *Journal of the American Society for Information Science and Technology, 57*(10), 1326–1337.

Tsuda, K., Rinaldo, F. J., Kryssanov, V. V., & Thawonmas, R. (2006). The structure of patent authorship networks in Japanese manufacturing companies. In *ICE-B* (pp. 289–293). International Conference on E-Business, Setubal, Portugal. http://www.ice.ci.ritsumei.ac.jp/~ruck/PAP/ice-b06.pdf. Accessed 20 April 2014.

Turenne, N. (2010). Modelling noun-phrase dynamics in specialized text collections. *Journal of Quantitative Linguistics, 17*(3), 212–228.

Veltri, G. A. (2012). Viva la Nano-Revolución! A semantic analysis of the Spanish national press. *Science Communication, 35*(2), 143–167.

Wang, L., Notten, A., & Surpatean, A. (2013). Interdisciplinarity of nano research fields: A keyword mining approach. *Scientometrics, 94*(3), 877–892.

Warris, C. (2004). *Nanotechnology benchmarking project* (p. 45). Australian Academy of Science. http://www.sciencearchive.org.au/policy/nano-report.pdf. Accessed 20 April 2014.

Yan, S., Spangler, W. S., & Chen, Y. (2013). Chemical mame extraction based on automatic training data generation and rich feature set. *IEEE-ACM Transactions on Computational Biology and Bioinformatics, 10*(5), 1218–1233.

Zhang, W., Yoshida, T., & Tang, X. (2009). Distribution of multi-words in Chinese and English documents. *International Journal of Information Technology & Decision Making, 8*(2), 249–265.

Zibareva, I. V., Vedyagin, A. A., & Bukhtiyarov, V. I. (2014). Nanocatalysis: A bibliometric analysis. *Kinetics and Catalysis, 55*(1), 1–11.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Management, 42*(6), 1513–1531.