Contribution to the Theme Section 'Global university rankings uncovered' (eds Stergiou KI, AC Tsikliras), Ethics Sci Environ Polit 13(2), 2013 (*http://www.int-res.com/journals/esep/esep-forthcoming/)*

Global university reputation and rankings: insights from culturomics

Konstantinos I. Stergiou*, Athanassios C. Tsikliras

Laboratory of Ichthyology, Department of Zoology, School of Biology, Aristotle University of Thessaloniki, UP Box 134, 541 24 Thessaloniki, Greece

* Corresponding author: K.I. Stergiou, email: kstergio@bio.auth.gr

Running head: University rankings and culturomics

ABSTRACT: In this study, we used culturomics (i.e. analysis of large electronic datasets for the study of human culture) in order to study the use of the names of various universities in the digitized corpus of English books. In particular, we used the Google Ngram viewer (available online: http://books.google.com/ngrams) to produce the frequencies of the names of 13 US, 5 UK and 4 Canadian universities in the English books and examined how these frequencies changed with time (1800-2008). We further used these frequencies to establish reputation rankings for these universities. Our results showed that Ngram is an easy-and-cheap-to-apply tool to approximate the reputation and 'intellectual' impact of universities over long time periods. Its reputation-generating capability, at least for top universities, is not worse than the within- and between-system capabilities of commercial tools (i.e. QS, THE and THE World Reputation Rankings). Ngram can, thus, be promising at least for students (and their families), who make choices that are affected by rankings, providing them with additional benefits (e.g. perception of the historical impact of a university) when compared to the short-term, volatile annual commercial rankings.

KEY WORDS: Global university rankings· University reputation· Google Ngram· Culturomics· QS· THE· Reputation rankings

INTRODUCTION

Global university rankings (GURs) are attracting increasing attention in the agenda of stakeholders directly or indirectly related to higher education (e.g. politicians, managers, administrators, policy makers, institutions, academia, students), and the number of agencies performing GURs is increasing with time (e.g. Harvey 2008,Williams 2008, Rauhvargers 2011, 2013, Jarocka 2012, Hazelkorn 2013). Available global ranking systems develop their annual league tables based generally on (e.g. Buela-Casal et al. 2007, Enserink 2007, Huang 2011, Federkeil 2009, Rauhvargers 2011, 2013, Hazelkorn 2013): (a) a variety of quantitative criteria and measures on which give different weights (e.g. number of papers, publications in Science/Nature, number of citations, number of Nobel Prize winners among their staff and alumni, faculty/student ratio); (b) web presence, visibility and access (such as Webometrics); and (c) reputation, such as, e.g., the World Reputation Rankings, henceforth called THER, produced since 2010 by Times Higher Education (THE), which are based on invitation-only survey of academic opinion. The degree of subjectivity of reputation rankings increases (see Federkeil 2009, Rauhvargers 2011, 2013, for an extensive discussion on reputation rankings and their shortcomings).

Fame, or reputation, is what is said or reported about a name. The reputation of a university may be defined as (van Vught 2008): *'The reputation of a higher education institution can be defined as the image (of quality, influence, trustworthiness) it has in the eyes of others. Reputation is the subjective reflection of the various actions an institution undertakes to create an external image. The reputation of an institution and its quality may be related, but they need not be identical. Higher education institutions try to influence their external images in many*

63 *ways, and not only by maximizing their quality.'*. University reputation, which has different
64 meanings for different groups and scientific fields, is '*a form of social*
65 *capital within the system of higher education that can be transformed into economic*
66 *capital, too*' (Federkeil 2009).

67 Although fame, on an individual perception basis, might be subjective, it can be
68 objectively measured by quantitatively estimating the frequency of the name appearing in
69 various sources, including books (Michel et al. 2010). The digitization of a millions of books
70 available online provides an important source and opportunity to study cultural trends (and
71 human behavior) based on the quantitative analysis of language and word usage in such digitized
72 texts; this new scientific field is known as culturomics (Michel et al. 2010).

73 Michel et al. (2010) constructed a corpus of digitized books (nowadays making up about
74 6% of all books ever printed: Lin et al. 2012) and, using the percentage of times a word/phrase
75 appears in the corpus of books (available in eight languages: English, Spanish, German, French,
76 Russian, Italian, Hebrew and Chinese), investigated cultural and other trends. Their approach
77 provides insights for different fields and issues (e.g. lexicography, collective memory, fame,
78 censorship, epidemiology) and gives rise to an important analytical tool for social sciences and
79 the humanities. Michel's et al. (2010) computational tool, later expanded (Lin et al. 2012), the
80 Google Ngram viewer (henceforth called Ngram), is available online
81 (http://books.google.com/ngrams). It has been recently applied in various fields, e.g. for tracking
82 emotions in novels (Mohammed 2011, Acerbi et al. 2013), for tracking poverty enlightenment
83 (Ravallion 2011), as a grammar checker (Nazar & Renau 2012), for studying the evolution of
84 computing (Soper & Turel 2012) and novel (Egnal 2013), in accounting (Ahlawat & Ahlawat
85 2012), in poetry (Diller 2013) and for analyzing drug literature (Montagne & Morgan 2013).

Herein, we used Ngram to investigate patterns in the use of university names (i.e. frequency of times appearing in the digitized books) and related such patterns with the rankings derived from three different commercial systems QS, THE and THEREP.

MATERIAL AND METHODS

Ngram estimates the usage of small sets of phrases and produces a graph the y axis of which shows how a phrase occurs in a corpus of books during a particular period relatively to all remaining phrases composed of same number of words (Lin et al. 2012). The analysis is available for 1800-2008 (Lin et al. 2012). A detailed account of the Ngram technique is provided in Michel et al. (2010) and Lin et al. (2012) whereas a step-by-step guide for its application using examples is available online (http://books.google.com/ngrams/info#advanced).

We used Ngram for estimating the percentages of the names of the top US, Canadian and UK universities appearing in the corpus of English books during 1800-2008. For US and UK we selected all the universities found in the first 20 QS positions for 2012/13 (Table 1). For UK we also selected University of Edinburgh, which appeared in position 21. For Canada, we selected the first four universities appearing in the QS and THE lists (i.e. University of Toronto, McGill University, University of British Columbia and University of Alberta).

We consequently extracted the QS rankings of all of the US, UK and Canadian universities for all the years that are available (i.e. 2012/13, 2011, 2009, 2008; data are not available online for 2010) and estimated the mean annual rank for each of these universities (Table 1). We did the same using the THE and THER data for the available years (i.e. 2012/13, 2011/12, 2010/11) (Table 1). Based on the mean annual QS, THE and THER scores, we ranked the 13 US, 4 Canadian and 5 UK universities from 1 to 13, 1 to 4 and 1 to 5 (i.e. henceforth

called national lists), respectively, for each of the three systems. We used the recent Ngram frequencies (1980-2000) of the US, Canadian and UK universities to rank them in terms of reputation at the national level. Although we also present the frequencies for 2000-2008 we did not use them for the ranking because of technical differences between the data before and after 2000 (Michel et al. 2010). We then compared the Ngram national ranks with the national QS, THE and THER rankings estimated as described above. For this, we estimated the average difference between all combinations of the national QS, THE and THER ranks for all universities examined here. The average difference was 2 and was used as a reference point for comparing the Ngram reputation rankings with those of the three systems (i.e. we considered that differences in national rankings between Ngram and each of QS, THE and THER were important when they were greater than 2).

We also produced Ngram graphs for 10 European historical universities and compared their average of the lowest and highest frequency during 1980-2000 with the year of their establishment (taken from http://en.wikipedia.org/wiki/List_of_oldest_universities_in_continuous_operation).

RESULTS

The graphs produced with Ngram show trends in two (e.g. name-university: Stanford University) or three ngrams (e.g. university-of-name: University of Pennsylvania) during 1800 to 2008. The y-axis shows the percentage of the phrase selected when compared to all bigrams (or trigrams) contained in the corpus of the English books.

With respect to the top US universities (Fig. 1), the frequencies of all the university names examined here increased from 1800 to the 2000 with the exception of that for University

of Columbia, which peaked in the 1940s and declined thereafter, Stanford University, which peaked in 1970 and slightly declined thereafter, University of Michigan, which reached a peak in late 1970s and then declined, and University of Pennsylvania, which peaked in 1980 and remained stable thereafter. The frequencies for Harvard and University of Pennsylvania were higher than those of the remaining universities during 1800-1920. However, Columbia University[1] before 1896 was known as Columbia College, which had frequencies that increased up to 0.0001244 in 1895, being similar to those of University of Pennsylvania for the period up to the early 1870s (graph not shown). During 1920-1960 the frequencies for University of Columbia were higher than the remaining ones. After 1960, University of Chicago attained higher frequencies from all the remaining universities equaling those of Harvard for the years following 1980s (Fig.1). The frequencies of occurrences of the 13 US universities during 1980-2000 were higher than 0.00019, with the exception of that for California Institute of Technology, which was around 0.000045 (Fig. 1).

We also searched for many other US universities that appear in the first 200 positions (i.e. University of Washington, Rice University, Boston University, Purdue University, Ohio State University, University of Southern California, Northwestern University, Brown University, University of Minnesota, University of Florida; figure A1 online Appendix) all of which had, during 1980-2000, frequencies <0.00016, i.e. smaller than those of the 13 top US universities (but higher than that for California Institute of Technology) (Fig. 1). The only exception was the

[1] We also searched for Barnard College and Teachers College, both of which are affiliated with Columbia University (graphs not shown here). The frequencies of Barnard College during 1890-2008 were by 1 to 2 orders of magnitude smaller than the frequencies of Columbia University. In contrast, the frequencies of Teachers College increased exponentially from 1900 to a maximum in the early 1930s, with frequencies similar to those of Columbia University during 1927-1931, and since then declined exponentially to frequencies that were by 5 to 7 times lower than those of Columbia University during 1980-2000.

University of Minnesota with a frequency of 0.00036 in 2000 (around 0.00032 for 1980-2000), i.e. ranked similarly with the University of Pennsylvania during this period, and University of Washington, which had an average 1980-2000 frequency of about 0.00020, i.e. similar to that of Duke University. When we searched for University of California, its frequency in the corpus of English books was higher than those of the 13 US universities, reaching 0.0015 in 2000 (with an average of about 0.0014 for 1980-2000). This is clearly attributed to the fact that this university includes several universities in different cities (i.e. Berkley, San Diego, Santa Barbara, Los Angeles, San Francisco, Irvine) all of which had, however, frequencies <0.000004, with the exception of University of California, Los Angeles, which, when searched as "UCLA", its frequency climbed up to 0.00025 in 2000 (with an average 1980-2000 frequency of about 0.00024), thus positioned higher than Duke University and California Institute of Technology but lower than the remaining 11 universities. The frequencies of the remaining University of California sites also increased when we added the frequencies for their acronyms (i.e. UCSB, UCSD, UCI, UCB, UCSF) but all frequencies were <0.00004. This additional analysis showed that the 13 top US universities examined here are generally the dominant ones in terms of frequencies with which their names appear in the corpus of English books.

We ranked the 13 universities in terms of reputation based on their recent frequencies (1980-2000) (Table 2). These ranks were compared with the national QS, THE and THER ranks. With the exception of Harvard and MIT, for which all rankings provided the same results, the Ngram reputation rankings differed from the QS ones for 7 universities, with individual differences ranging from 3 to 4, from the THE rankings for 9 universities, with individual differences ranging from 3 to 8, and from the THER ones for 8 universities, with differences ranging from 3 to 9 (Table 2).

The mean QS and THE university rankings differed for 5 universities, by 3 to 4 positions, whereas the THE and THER rankings differed for 6 universities by 3 to 5 positions and the QS and THER rankings for 7 universities by 3 to 6 positions (Table 2). Thus, the differences between the Ngram and the QS/THE/THER rankings were generally similar to the differences between ranking systems themselves.

With respect to the four Canadian Universities (Fig. 2), their frequencies in the English corpus increased up to 1980 and then remained stable. University of Toronto and McGill University enjoyed similar frequencies up to 1920. For the years following 1920, University of Toronto dominated, with its frequencies in the years after 1980 (i.e. 0.00026) being one order of magnitude higher than those of the remaining three universities (Fig. 2). From the latter, McGill had higher frequencies during 1920 - early 1970s whereas from then onwards the frequencies of the University of British Columbia surpassed those of McGill. University of Alberta was characterized by the lowest frequencies throughout the period (Fig. 2). The frequencies of University of British Columbia, McGill and University of Alberta during 1980-2000 ranged between 0.000093-0.00013, 0.000065-0.000049 and 0.000044-0.000049, respectively.

We also searched for other Canadian Universities that appear in various lists (i.e. Université de Montréal, University of Victoria, Dalhousie University, University of Western Ontario, McMaster University, Queen's University, University of Waterloo, University of Calgary; figure A2 online Appendix) and all had frequencies in 1980-2000 <0.000034, i.e. lower than the ones presented in figure 2. The only exception was Queen's University the frequency of which approached that of McGill in the early 1990s, and surpassed it in late 1990s by a small margin (i.e. 0.000062 and 0.000052, respectively). However, there are more than one Queen's Universities in the world. The Ngram rankings derived from the frequencies were exactly the

same with those of THE and THER whereas they differed from the QS ones, according to which McGill University is in the first place and University of Toronto in the second one (Table 1).

For the five UK universities (Fig. 3), University of Edinburgh had the highest frequencies during 1800-1910, which then slightly declined. The frequencies of Oxford and Cambridge, two of the oldest European universities, established in 1167 and 1209, respectively, were similar up to 1920 and increased exponentially after 1920 and 1970, respectively, with Oxford having higher frequencies than Cambridge since 1920. In 1980-2000, the frequencies of University of Edinburgh, Imperial College and University College London were by 2 orders of magnitude lower than those of Oxford and Cambridge (Fig. 3). We also searched for several other UK universities (figure A3 online Appendix) that appear in top lists (e.g. London School of Economics, University of Southampton, University of Essex, University of Glasgow, Durham University, University of Warwick, University of Lancaster) all of which had frequencies that were by 1 or 2 orders of magnitude lower than those of Cambridge and Oxford. These additional universities had also frequencies that during 1980-2000 were lower than those of University of Edinburgh (range: 0.000052-0.000061) and University College London (range: 0.000028-0.000088). The only exception was London School of Economics, which had frequencies ranging from 0.000086 to 0.00011, thus dominating the remaining universities after the mid 1940s but still 1 order of magnitude lower than those of Oxford and Cambridge in recent years (Fig. 3).The Ngram rankings differed by 1 or 2 positions than the other systems (Table 2) because Oxford is ranked first in Ngram and THE and second in QS and THER whereas the opposite is true of Cambridge.

Across countries, the frequencies of Oxford were higher than those of Harvard and Chicago after 1980 and of Cambridge after 1990. The frequencies of these two UK universities

220 in the last years are by 1.5 to 2 times higher than those of University of Chicago and Harvard
221 whereas the frequencies of the University of Toronto were by one order of magnitude lower than
222 those of the above four universities.
223 Overall, for all the 22 US, UK and Canadian universities examined here, the national
224 Ngram ranks were significantly correlated with the national QS (Fig. 4) and THER ones ($r=0.53$
225 and 0.46, $P<0.05$, respectively) but not with the THE ones ($r=0.32$, $P>0.05$).
226 The Ngram graphs for 10 of the oldest universities in the world are shown in figure 5.
227 Although the frequencies of these universities are by 2 to 3 orders of magnitude lower than those
228 of the US, UK and Canadian ones, this is expected given the use of the English corpus of books.
229 What is important here is that such historical universities do appear regularly in English books,
230 with percentages fluctuating with time.There is a positive relation between the age of the
231 university and its frequency in the corpus. Thus, the oldest university, University of Bologna,
232 generally displays the highest frequencies (except during 1950-1970 when University of Padua
233 attained higher frequencies), followed by the Universities of Padua, Salamanca, Naples,
234 Coimbra, Toulouse, Siena (its frequency increased exponentially since 1970), Valladolid, Murcia
235 and Macerata (established in 1290), which is not shown in figure 5 because of its very small
236 frequency when compared to the remaining ones. Indeed, the year of establishment of these
237 universities was negatively correlated ($r=-0.82$, $P<0.05$) (Fig. 6) with their average frequency
238 during 1980-2000 in the corpus of English books. It is worthy of mention, here, that from these
239 10 universities, only University of Bologna is found in the top 200 QS 2012/13 universities
240 whereas the Universities of Toulouse, Coimbra, Padua and Montpelier are among the top 500 QS
241 2012/13 (at positions from 278 to 386).
242

DISCUSSION

In this study, we used Ngram to produce the frequencies of the names of 22 US, UK and Canadian universities in the digitized corpus of English books, which is comprised by about half a trillion words (Lin et al. 2012), and studied how these frequencies changed with time (1800-2008). We further used the frequencies during 1980-2000 to establish reputation rankings for these universities. Naturally, books is only one such source that can be used to study reputation, with many other sources being also important and useful (e.g. newspapers, magazines, media: Michel et al. 2010; blogs and social networks: Dodds et al. 2011, Altmann et al. 2011, Ratkiewicz 2011).

Our results showed that the differences between the Ngram and the QS/THE/THER rankings for US universities are similar to the differences between the three ranking systems themselves, whereas the rankings for UK and Canadian universities were almost identical for the various systems (Table 2). This, together with the fact that Ngram and QS and THER national ranks were significantly correlated, clearly indicates that Ngram generally captures and reflects the reputation to the same extent that commercial rankings do, at least of the very top universities, in each country.

The within- and between-systems differences in rankings can generally be high albeit less so for the very top ones (e.g. Dichev 2001, Marginson 2007, Usher & Savino 2007, Federkeil 2009, Huang 2011, Chen & Liao 2012).The same was also true of the QS, THE and THER rankings for the years used here. For instance, from Table 1 is evident that with the exception of Harvard, MIT, Johns Hopkins, University of Michigan and Oxford for which the differences in mean annual ranks between QS and THE are <1, for all remaining universities the differences were from 2.6 to 31 positions. Thus, one has to wonder about the usefulness of the exact annual

266 rank of a university (e.g. McGill University: position 18 or 32; University of Alberta: position 85
267 or 116) (Table 1), that reflects noise rather than news (Dichev 2001), as opposed to some index
268 referring to a relatively long period.

269 Our results showed that Ngram is an easy-and-cheap-to-apply tool to approximate the
270 reputation and 'intellectual' impact of universities over long time periods. Its reputation-
271 generating capability, at least for top national universities, is not worse than the within- and
272 between-systems capabilities of the commercial tools, which are generally regarded as providing
273 'reliable' information. However, if the reputation ranking of universities can be obtained by just
274 typing their names in Ngram and checking their frequencies, then there is probably no need to
275 resource to the very expensive procedures of the commercial reputation ranking systems, which
276 take into account a large number of variables and their reputation scores of universities are
277 practically meaningless for universities below the top 50 (Rauhvargers 2013). In addition,
278 contrary to various indicators used in commercial ranking systems that can be 'manipulated' by
279 institutes for climbing up the rank (e.g. see Table 1 in Hazelkorn 2009), Ngram cannot. Ngram
280 can, thus, be promising at least for students (and their families), who make choices that are
281 affected by rankings in an increasing degree (e.g. Sauder & Lancaster 2006, Bowman & Bastedo
282 2009, Hazelkorn 2009) and pay particular attention to reputation (Federkeil 2009). Naturally,
283 student decisions on selecting a university are a multidimensional process that depends also on
284 other factors (e.g. other reputation and prestige indicators such as tuition fees and instructional
285 expenditure for liberal arts: Bowman & Bastedo 2009; student's economic status: Clarke 2007).
286 Students might have additional 'educational' benefits by using the Ngram tool. For instance, they
287 will also have a perception of the historical impact of a university, something that it is not true
288 when the short-term, volatile rankings are concerned (the earliest GUR system is available since

2003), which might mislead students when making their choice. Indeed, the ten old universities examined here might not appear in top 100 lists but historical universities have undoubtedly driven the evolution of modern universities and higher education in general. This contribution and historical perspective can be felt when someone is visiting their campuses and especially their libraries (e.g. University of Coimbra, University of Salamanca, Trinity College in Dublin).

In general, one might expect that references to old universities will decrease in the last several decades, because more, newer, institutions are now competing for reputation. However, with few exceptions (e.g. Columbia University, Stanford University, University of Michigan, University of Salamanca, University of Padua: Figs 1 to 3, 5) for which the frequencies consistently declined for an extended period, the frequencies of the universities examined here generally increased with time during the last 100 years. This is most probably explained by the fact that the increase in the number of universities competing for reputation parallels a global large increase in the references to universities.

Although people are becoming more famous nowadays than before, they are also forgotten more rapidly (Michel et al. 2010). In contrast, as mentioned above, universities are generally characterized by rather continually increasing fame, which must be attributed to the fact that universities are there forever and their fame is accumulated from generation to generation. This agrees with the positive relation between Ngram frequency and age of universities. As universities are the productive units of scientific knowledge, this fame accumulation certainly reflects the accumulation of knowledge and thus the continually growing importance of science to the well being and future of our societies.

Our work suffers from certain biases in the estimations of frequencies. For instance, when searching of university names using their acronyms, Ngram might be counting the

312 frequency of acronyms that also refer to other entities. For example, when searching for
313 University of California, Berkeley, as 'UCB', the corpus will obviously provide the sum of the
314 frequencies of all the occurrences of this one ngram acronym (e.g. University of Colorado at
315 Boulder, United Christian Broadcasters, if they are occurring), irrespectively of its actual
316 reference. Thus, there is a risk of having a bias in the frequency count. One might need to use
317 very sophisticated disambiguation algorithms to determine the correct reference of an acronym in
318 a given context, and, with a limited context window of one ngram, this can be rather hard. This
319 problem of ambiguity also applies to the case of universities that are also publishing houses. In
320 this case, part (ranging from relatively small, e.g. University of Michigan, to large, e.g.
321 Cambridge and Oxford) of the frequency count of the names of these universities will be because
322 of the citations of the books by this publisher. Although the frequencies related to university
323 publishing houses are most probably part of a university's reputation, one would need to measure
324 the impact of works published by authors affiliated to other universities and printed by other
325 publishing houses to make up for that extra bonus that is given to the universities with publishing
326 houses. In that sense, this is also a source of bias that needs more complex statistical procedures,
327 algorithms and analyses applied on the downloaded whole dataset in order to be controlled (see,
328 e.g., Acerbi et al. 2013).

329 The analysis presented here might also have important cultural and historical
330 implications, which, however, are outside the scope of this work. For instance, the frequencies of
331 the 10 old European universities displayed characteristic periodicities of about 20 years that
332 might reflect important historical and cultural events (and see Gao et al. 2012, for analyzing
333 long-range correlations in ngram frequencies). The same is also true of the alternating patterns in
334 terms of frequency dominance between universities (e.g. Universities of Coimbra and Toulouse:

during1800-1870 and 1940-today, University of Coimbra has higher frequencies than University of Toulouse whereas the opposite is true of 1870-1940). Another interesting issue is the relation between the increasing frequencies of the University of Bologna in the last years (Fig. 5) and the *Magna Charta Universitatum Europaeum* that was proposed by the University of Bologna in 1986 and the Bologna Declaration of 1999 towards the reform of Higher Education in Europe. Finally, the prominent declining pattern in the frequency for the Columbia University after 1940 (Fig. 1) may be related to particular historical facts that might have affected its reputation (e.g. atom research and the Manhattan Project in the 1940s, intense student activism in the 1960s resulting in the President's resignation, links between the university and the Vietnam War, Columbia College did not admit women until 1983: http://en.wikipedia.org/wiki/Columbia_University, section Columbia University (1896–present); assessed 19 August 2013).

Acknowledgments. The authors wish to thank C. Apostolidis and two anonymous reviewers for their valuable comments and suggestions.

LITERATURE CITED

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) The expression of emotions in 20th Century Books. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

Ahlawat S, Ahlawat S (2012) An innovative decade of enduring accounting ideas as seen through the lens of culturomics: 1900-1910. American Institute of Higher Education – The 7th International Conference Williamsburg, VA – March 7 – 9, pp 8-19

357 Altmann EG, Pierrehumbert JB, Motter AE (2011) Niche as a determinant of word fate in online
358 groups. PLoS One 6(5): e19009. doi:10.1371/journal.pone.0019009
359 Bowman NA, Bastedo MN (2009) Getting on the front page: Organizational reputation, status
360 signals, and the impact of U.S. News and World Report on student decisions. Res High
361 Educ 50: 415–436
362 Buela-Casal G, Gutiérrez-Martínez O, Bermúdez-Sánchez MP, Vadillo-Muñoz O (2007)
363 Comparative study of international academic rankings of universities. Scientometrics 71:
364 349–365. doi: 10.1007/s11192-007-1653-8
365 Chen K, Liao P (2012) A comparative study on world university rankings: a bibliometric survey.
366 Scientometrics 92: 89-103
367 Gao J, Hu J, Mao X, Perc M (2012) Culturomics meets random fractal theory: insights into long-
368 range correlations of social and natural phenomena over the past two centuries. J. R. Soc.
369 Interface 9: 1956–1964 doi:10.1098/rsif.2011.0846
370 Clarke M (2007) The Impact of higher education rankings on student access, choice, and
371 opportunity. Higher Educ Europe 32: 59-70
372 Dichev I (2001) News or noise? Estimating the noise in the U.S. News university rankings. Res
373 Higher Educ 42: 237-266
374 Diller HJ (2013) Culturomics and genre: Wrath and anger in the 17 th Century. In: McConchie R
375 W et al. (ed), Selected Proceedings of the 2012 Symposium on New Approaches in English
376 Historical Lexis (HEL-LEX 3), 54-65. Somerville, MA: Cascadilla Proceedings Project.
377 Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of
378 happiness and information in a global social network: Hedonometrics and twitter. PLoS
379 one 6(12): e26752. doi:10.1371/journal.pone.0026752

380 Egnal M (2013) Evolution of the novel in the United States - The statistical evidence. Social
381 Science History 37: 231-254. doi: 10.1215/01455532-2074429
382 Enserink M (2007) Who ranks the university rankers? Science 317: 1026-1028.
383 Federkeil G (2009) Reputation indicators in rankings of higher education institutions. In: Kehm
384 BM, Stensaker B (eds), University Rankings, diversity, and the new landscape of higher
385 education, pp 19–33, Sense Publishers, Rotterdam/Boston/Taipei
386 Harvey L (2008) Rankings of higher education institutions: A critical review. Qual Higher Educ
387 14: 187-207
388 Hazelkorn E (2009) Rankings and the battle for world-class excellence: Institutional strategies
389 and policy choices. Higher Educ Managem Policy 21: 1-22
390 Hazelkorn E (2013) How rankings are reshaping higher education. In: Climent V, Michavila F,
391 Ripolles M (eds), Los Rankings Univeritarios: Mitos y Realidades, Ed. Tecnos
392 Huang M-H (2011) A comparison of three major academic rankings for world universities: from
393 a research evaluation perspective. J Lib Inf Stud 9: 1-25
394 Jarocka M (2012) University ranking systems – From league table to homogeneous groups of
395 universities. World Academy of Science, Engineering and Technology 66: 800-805
396 Lin Y, Michel J-B, Aiden EL, Orwant J, Brockman W, Petrov S (2012) Syntactic annotations for
397 the Google Books Ngram corpus. Proceedings of the 50th Annual Meeting of the
398 Association for Computational Linguistics Volume 2: Demo Papers (ACL '12)
399 Marginson S (2007) Global university rankings: Implications in general and for Australia.
400 Journal of Higher Education Policy and Management 29: 131-142
401 Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Brockman W, The Google Books Team,
402 Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL

(2010) Quantitative analysis of culture using millions of digitized books. Science 331: 176-182. DOI: 10.1126/science.1199644

Mohammad S (2011) From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp 105–114, Portland, OR, USA, 24 June 2011

Montagne M, Morgan M (2013) Drugs on the internet, Part IV: Google's Ngram Viewer analytic tool applied to drug literature. Substance Use and Misuse 48: 415-419. doi:10.3109/10826084.2013.763493

Nazar R, Renau I (2012) Google books N-gram corpus used as a grammar checker. Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing, pp 27–34, Avignon, France, April 23, 2012

Ratkiewicz J, Conover MD, Meiss M, Goncalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp 297-304

Rauhvargers A (2011) Global university rankings and their impact. European University Association, Brussels, 81 pp (electronic version available at www.eua.be)

Rauhvargers A (2013) Global university rankings and their impact – Report II. European University Association, Brussels, 87 pp (electronic version available at www.eua.be)

Ravallion M (2011) The two poverty enlightenments: Historical insights from digitized books spanning three centuries. Poverty and Public Policy 3: 1-46

Sauder M, Lancaster R (2006) Do rankings matter? The effects of U.S. News &World Report rankings on the admissions process of Law Schools. Law Soc Rev 40: 105-134

426 Soper DS, Turel O (2012) An n-Gram analysis of communications 2000–2010. Communications
427 of the ACM 55(5): 81-87
428 Usher A, Massimo S (2007) A global survey of university ranking and league tables. Higher
429 Educ Europe 32: 5-15
430 Van Vught F (2008) Mission diversity and reputation in higher education. Higher Educ Pol 21:
431 151–174
432 Williams R (2008) Methodology, meaning, and usefulness of rankings. Australian Universities'
433 Review 50: 51-58
434

Table 1. Annual and mean annual rankings for different top US, Canadian and UK universities according to QS, Times Higher Education (THE) and THE World Reputation Rankings (THER).

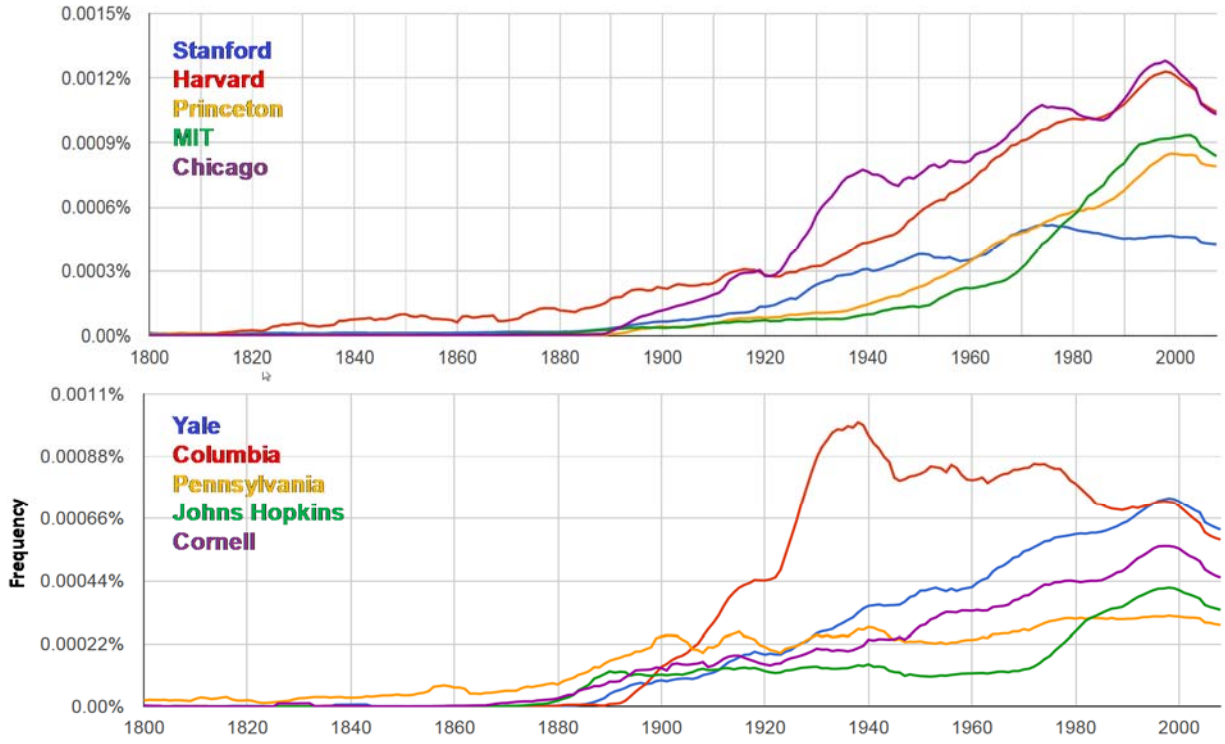
Country/University	Annual world university rankings										Mean annual ranking		
	QS				THE			THER			QS	THE	THER
	2012	2011	2009	2008	2012	2011	2010	2012	2011	2010	2008-12	2010-12	2010-12
US													
Harvard	3	2	1	1	4	2	1	1	1	1	1.8	2.3	1.0
MIT	1	3	9	9	5	7	3	2	2	2	5.5	5.0	2.0
Yale	7	4	3	2	11	11	10	10	10	9	4.0	10.7	9.7
CalTech	10	12	10	5	1	1	2	11	11	10	9.3	1.3	10.7
Chicago	8	8	7	8	10	9	12	14	14	15	7.8	10.3	14.3
Princeton	9	13	8	12	6	5	5	7	7	7	10.5	5.3	7.0
Stanford	15	11	16	17	2	2	4	6	4	5	14.8	2.7	5.0
Columbia	11	10	11	10	14	12	18	14	15	23	10.5	14.7	17.3
Pennsylvania	12	9	12	11	15	16	19	18	19	22	11.0	16.7	19.7
Johns Hopkins	16	16	13	13	16	14	13	19	18	14	14.5	14.3	17.0
Cornell	14	15	15	15	18	24	14	17	16	16	14.8	18.7	16.3
Michigan	17	14	19	18	20	18	15	12	12	13	17.0	17.7	12.3
Duke	20	19	14	13	23	22	24	31	33	36	16.5	23.0	33.3
Canada													
Toronto	19	23	29	41	21	19	17	16	16	17	28.0	19.0	16.3
McGill	18	17	18	20	34	28	35	31	25	29	18.3	32.3	28.3
British Columbia	45	51	40	34	30	22	30	31	25	31	42.5	27.3	29.0
Alberta	108	100	59	74	121	100	127				85.3	116.0	
UK													
Oxford	5	5	5	4	2	4	6	4	6	6	4.8	4.0	5.3
Cambridge	2	1	2	3	7	6	6	3	3	3	2.0	6.3	3.0
University College	4	7	4	7	17	17	22	20	21	19	5.5	18.7	20.0
Imperial College	6	6	5	6	8	8	9	14	13	11	5.8	8.3	12.7
Edinburgh	21	20	20	23	32	36	40	46	49	45	21.0	36.0	46.7

439 Table 2. National ranks developed from the mean annual ranks of QS, Times Higher Education
440 (THE) and THE World Reputation Rankings (THER) (see Table 1) and from Ngram analysis for
441 1980-2000.

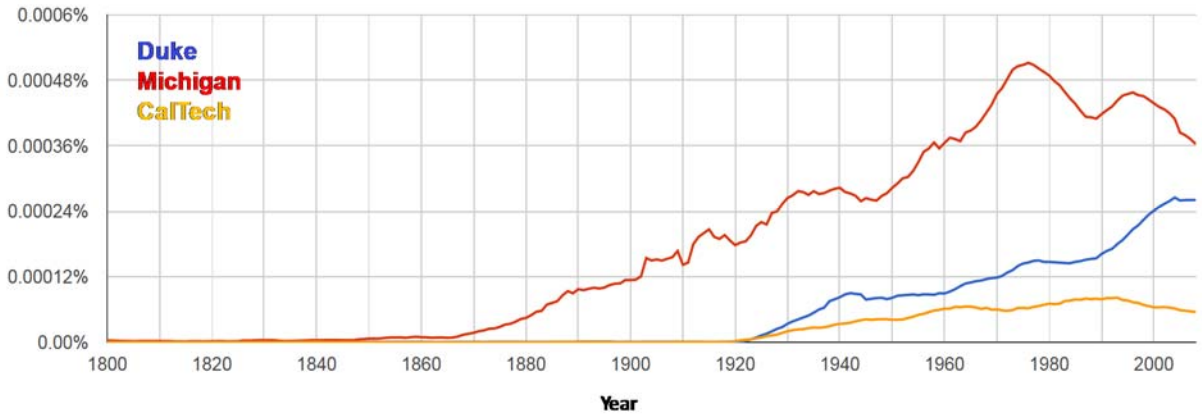
442

University	National ranks			
	QS	THE	THER	Ngram
US				
Harvard	1	2	1	1
MIT	3	4	2	2
Yale	2	6	5	2
CalTech	5	1	6	7
Chicago	4	5	8	1
Princeton	6	4	4	2
Stanford	7	3	3	4
Columbia	6	8	11	2
Pennsylvania	6	9	12	5
Johns Hopkins	7	7	10	4
Cornell	7	11	9	3
Michigan	8	10	7	4
Duke	8	12	13	6
Canada				
Toronto	2	1	1	1
McGill	1	2	3	3
British Columbia	3	3	2	2
Alberta	4	4	4	4
UK				
Oxford	2	1	2	1
Cambridge	1	2	1	2
University College	3	4	4	4
Imperial College	3	3	3	5
Edinburgh	4	5	5	3

List of Figures



467



468

469

470 Fig. 1

471

472

473

474

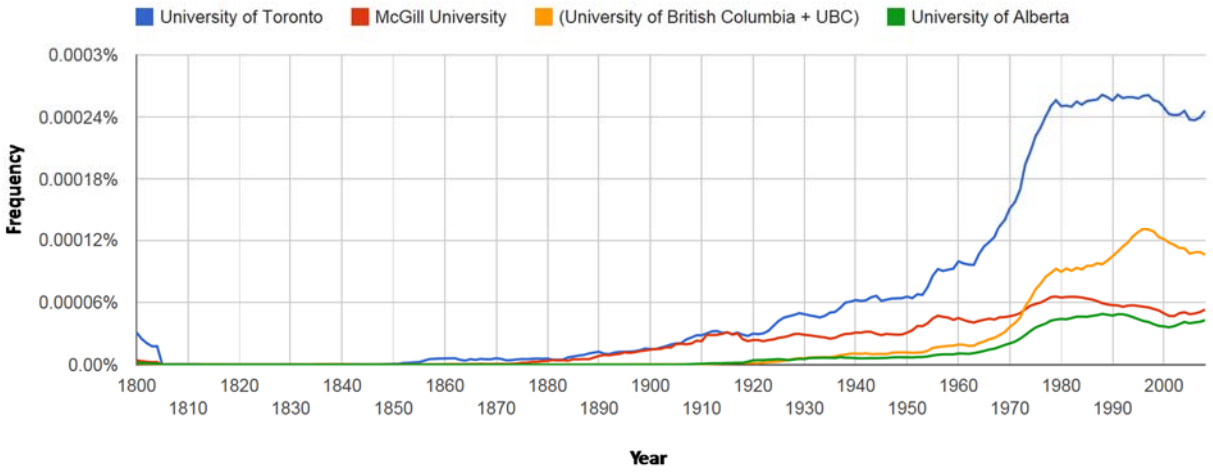

University of Toronto
McGill University
(University of British Columbia + UBC)
University of Alberta
0.0003%
0.00024%
0.00018%
0.00012%
0.00006%
0.00%
Frequency
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Year

Fig. 2.

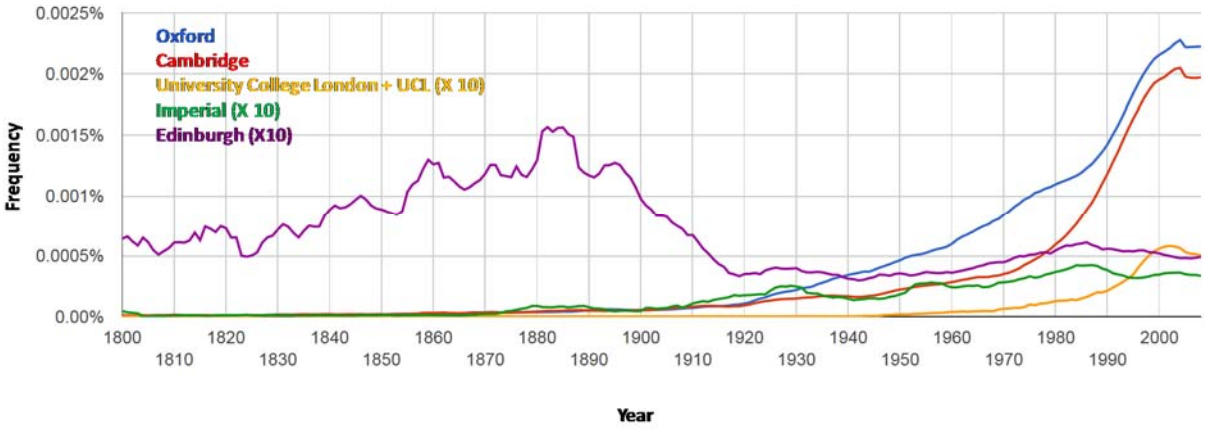

Oxford
Cambridge
University College London + UCL (X 10)
Imperial (X 10)
Edinburgh (X10)
Frequency
0.0025%
0.002%
0.0015%
0.001%
0.0005%
0.00%
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Year

Fig. 3

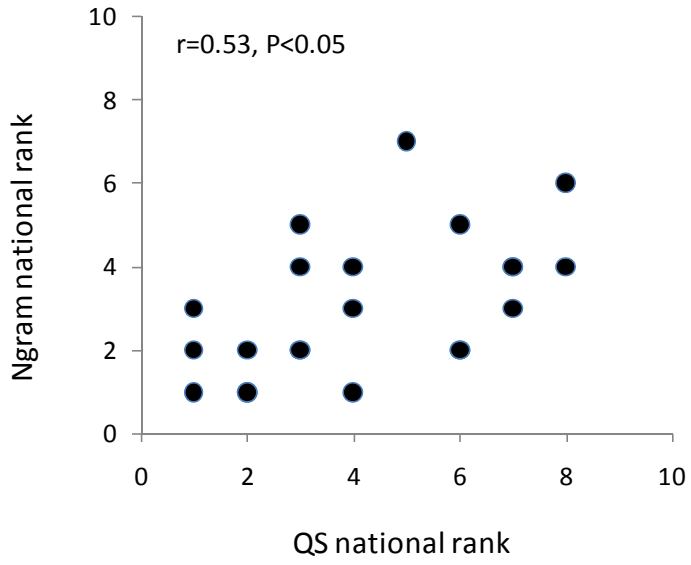


Fig. 4

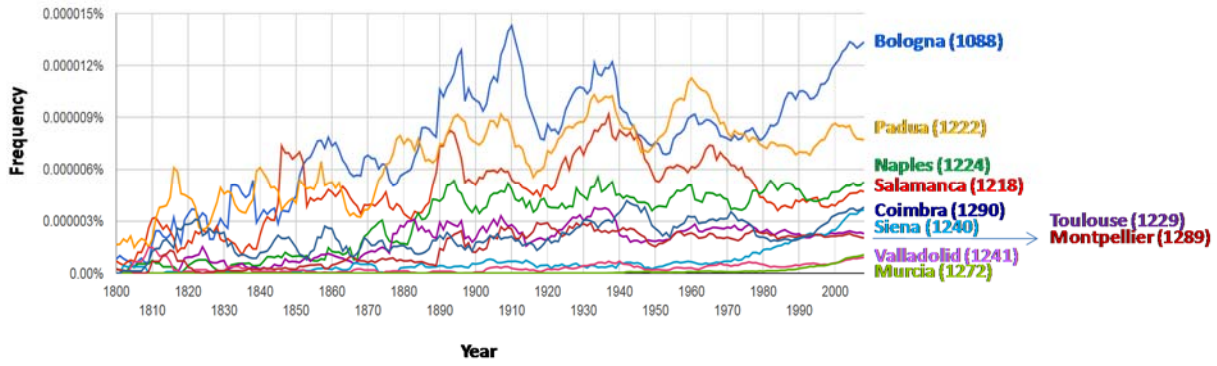

Frequency
0.000015%
0.000012%
0.000009%
0.000006%
0.000003%
0.00%
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Year
Bologna (1088)
Padua (1222)
Naples (1224)
Salamanca (1218)
Coimbra (1290)
Siena (1240)
Valladolid (1241)
Murcia (1272)
Toulouse (1229)
Montpellier (1289)

Fig. 5

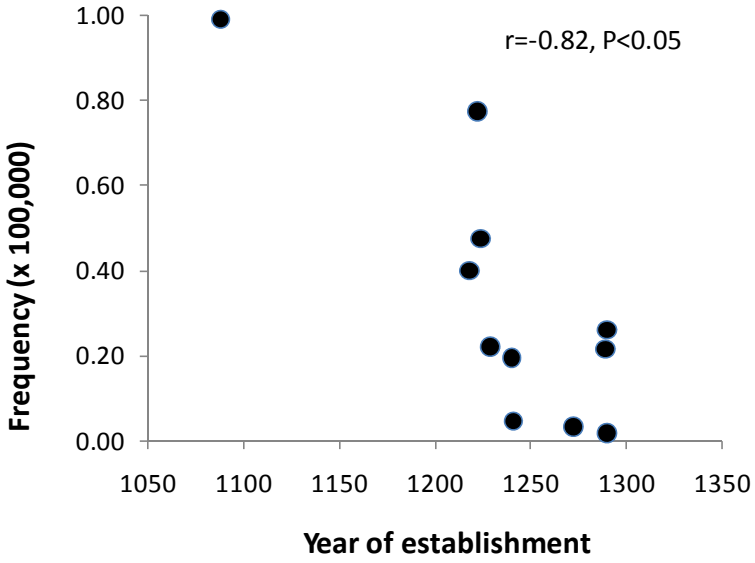

r=-0.82, P<0.05
Frequency (x 100,000)
1.00
0.80
0.60
0.40
0.20
0.00
1050
1100
1150
1200
1250
1300
1350
Year of establishment

Fig. 6.