Chapter 5: Reviewing the Reviewers: Training Neural Networks to Read Peer Review Reports

Martin Paul Eve, Birkbeck, University of London | Robert Gadie, University of the Arts, London | Victoria Odeniyi, University of the Arts, London | Shahina Parvin, Brandon University, Canada and Jahangirnagar University, Bangladesh.

Abstract

The study of academic peer review is often difficult owing to the confidentiality of reports. As an occluded genre of writing that nonetheless underpins scientific publication, relatively little is known about the ways that academics write and behave, at scale, in their reviewing practices. In this chapter, we describe for the first time the database of peer review reports at PLOS ONE, the largest scientific journal in the world, to which we had unique access. Specifically, we detail the approach that we took to training a multi-label, multi-class text classifier using the TenCent NeuralClassifier toolkit to examine the peer review reports. Although this resulted in a predictable failure to produce accurate levels of recall and precision, we argue that as these technologies further develop there are a range of uses – for both good and ill – that could be used to machine-read these archives.

1. Introduction - Reading Peer Review

Peer review is the system by which manuscripts are vetted for validity, appraised for originality, and selected for publication as articles in academic journals (serials) or as academic books (monographs).[1] Since an editor of an academic title cannot be expected to be an expert in every single area covered by a publication and since it appears undesirable to have a single person controlling the publication's flow of scientific and humanistic knowledge, there is a need for input from more people. Manuscripts submitted for consideration are shown to external expert advisers (peers) who deliver verdicts on the novelty of the work, criticisms or praise of the piece, and a judgement of whether to proceed to publication. A network of experts

1 Portions of this chapter are adapted from the openly licensed Martin Paul Eve et al, *Peer Review and Institutional Change in Academia*, Cambridge 2021.

with appropriate degrees of knowledge and experience within a field are coordinated to yield a set of checks and balances for the scientific and broader research landscapes. Editors are then bound, with some caveats and to some extent, to respect these external judgements in their own decisions, regardless of how harsh the mythical "reviewer 2" may be.[2]

The premise behind peer review may appear sound or even incontrovertible. Who could object to the best in the world appraising one another, nobly ensuring the integrity of the world's official research record? Yet, considering the system for even a few moments leads to several questions. What is a "peer" and who decides? What does it mean when a "peer" approves somebody else's work? How many "peers" are required before a manuscript can be properly vetted? What happens if "peers" disagree with each other? Does (or should) peer review operate in exactly the same fashion in disciplines as distinct as Neuroscience and Sculpture? Particle Physics and Social Geography? Math and Literary Criticism? When academics rely on publications for their job appointments and promotions, how does peer review interact with other power structures in universities? Do reviewers act with honour and integrity in their judgements within this system?

It is abundantly clear that the peer-review process is far from infallible. Every year, thousands of articles are retracted (withdrawn) for containing inaccuracies, for conducting unethical research practices, and for many other reasons.[3] On occasion, this has had devastating consequences in spaces such as public health. The

2 The age-old academic joke is that if the first reviewer loves an article, the second reviewer will hate it and be ultra-harsh in his or her judgement. Von Bakanic/Clark McPhail/Rita J. Simon, Mixed Messages: Referees' Comments on the Manuscripts They Review, in: *The Sociological Quarterly*, 30 (4/1989), 639-654, URL: http://www.jstor.org/stable/4121469; Lutz Bornmann/Hans-Dieter Daniel, The Effectiveness of the Peer Review Process: Inter-Referee Agreement and Predictive Validity of Manuscript Refereeing at *Angewandte Chemie*, in: *Angewandte Chemie International Edition*, 47 (38/2008), 7173-7178, doi:10.1002/anie.200800513; Louis Fogg/Donald W. Fiske, Foretelling the Judgments of Reviewers and Editors, in: *American Psychologist*, 48 (3/1993), 293-294, doi:10.1037/0003-066X.48.3.293; Stephen Lock, *A Difficult Balance: Editorial Peer Review in Medicine*, Philadelphia, PA, 1986; Richard E. Petty/Monique A. Fleming/Leandre R. Fabrigar, The Review Process at PSPB: Correlates of Interreviewer Agreement and Manuscript Acceptance, in: *Personality and Social Psychology Bulletin*, 25 (2/1999), 188-203, doi:10.1177/0146167299025002005; Robert J. Sternberg et al., Getting in: Criteria for Acceptance of Manuscripts in Psychological Bulletin, 1993-1996, in: *Psychological Bulletin*, 121 (2/1997), 321-323, doi:10.1037/0033-2909.121.2.321; Harriet Zuckerman/Robert K. Merton, Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System, in: *Minerva*, 9 (1/1971), 66-100, doi:10.1007/BF01553188.

3 Jeffrey Brainard/Jia You, What a Massive Database of Retracted Papers Reveals about Science Publishing's "Death Penalty," in: *Science | AAAS*, 25.10.2018, URL: https://www.sciencemag.org/news/2018/10/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty [last accessed: Mar. 31, 2021]; Retraction Watch, URL: https://retractionwatch.com/ [last accessed: Mar. 31, 2021]. Björn Brembs/Katherine Button/Marcus Munafò,

at-the-time respected researcher Andrew Wakefield's notorious 1998 retracted paper claiming a link between the mumps, measles, and rubella (MMR) vaccine and the development of autism in children was published in perhaps the most prestigious medical journal in the world, *The Lancet*.[4] The work was undoubtedly subject to stringent single-blind pre-publication review and was cleared for publication. Yet the article was later retracted and branded fraudulent, having caused immense and ongoing damage to public health.[5] It is, alas, always easier to make an initial statement than subsequently to retract or to correct it. As a result, a worldwide anti-vaccination movement has seized upon this circumstance as evidence of a conspiracy. The logic uses the supposed initial validation of peer review and the prestige of *The Lancet* as evidence that Wakefield was correct and that he is the victim of a conspiratorial plot to suppress his findings. Hence, when peer review goes wrong, the general belief in its efficacy, coupled with the prestige of journals founded on the supposed expertise of peer review, has damaging real-world effects. Indeed, there are longstanding criticisms of the validity of peer review, exemplified in Franz J. Ingelfinger's notorious statement that the process is "only moderately better than chance" and Drummond Rennie's (the then deputy editor of the Journal of the American Medical Association) "if peer review was a drug it would never be allowed onto the market."[6]

Despite the aforementioned challenges, the role of peer review in improving the quality of academic publications and in predicting the impact of manuscripts through criteria of "excellence" is widely seen as essential to the research endeavour. As a term that first entered critical and popular discourse around 1960 but also as a practice that only became commonplace far later than most suspect, peer review is sometimes described as the "gold standard" of quality control and the majority of researchers consider it crucial to contemporary science.[7] Indeed, peer review is

---

Deep Impact: Unintended Consequences of Journal Rank, in: *Frontiers in Human Neuroscience*, 7 (2013), 1-12, doi:10.3389/fnhum.2013.00291.

4 A. J. Wakefield et al., RETRACTED: Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children, in: *The Lancet*, 351 (9103/1998), 637-641, doi:10.1016/S0140-6736(97)11096-0.

5 Fiona Godlee/Jane Smith/Harvey Marcovitch, Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent, in: *BMJ*, 342 (2011), c7452, doi:10.1136/bmj.c7452.

6 Franz J. Ingelfinger, Peer Review in Biomedical Publication, in: *The American Journal of Medicine*, 56 (1974), 686-692; Hans-Dieter Daniel, *Guardians of Science: Fairness and Reliability of Peer Review*, Weinheim 1993, see 4; Peter M. Rothwell/Christopher N. Martyn, Reproducibility of Peer Review in Clinical Neuroscience, in: *Brain*, 123 (9/2000), 1964-1969, doi:10.1093/brain/123.9.1964; Richard Smith, Peer Review: A Flawed Process at the Heart of Science and Journals, in: *Journal of the Royal Society of Medicine*, 99 (4/2006), 178-182; Richard Smith, Classical Peer Review: An Empty Gun, in: *Breast Cancer Research*, 12 (4/2010), S13, doi:10.1186/bcr2742.

7 Melinda Baldwin, In Referees We Trust?, in: *Physics Today*, 70 (2/2017), 44-49, doi:10.1063/PT.3.3463; Melinda Baldwin, Scientific Autonomy, Public Accountability, and the Rise of "Peer Re-

much younger than many suspect. In 1936, for instance, Albert Einstein was outraged to learn that his unpublished submission to Physical Review had been sent out for review.[8] Yet, despite its relative youth, peer review has nonetheless become a fixture of academic publication. This raises the question, though, of why this might be the case. For surprisingly little evidence exists to support the claim that peer review is the best way to pre-audit work, leading Michelle Lamont and others to note the importance of ensuring that "peer review processes [… are] themselves subject to further evaluation."[9]

Research into peer review processes, however, can be difficult to conduct. Nevertheless, this has not prevented a burgeoning field from emerging around the topic.[10] Certainly, following the influential work of John Swales, there has been

---

view" in the Cold War United States, in: *Isis*, 109 (3/2018), 538-558, doi:10.1086/700070; Irene Hames, *Peer Review and Manuscript Management of Scientific Journals Guidelines for Good Practice*, Malden, MA, 2007, see 2; Bruce Alberts/Brooks Hanson/Katrina L. Kelner, Reviewing Peer Review, in: *Science*, 321 (5885/2008), 15, doi:10.1126/science.1162115; Samuel Moore et al., Excellence R Us: University Research and the Fetishisation of Excellence, in: *Palgrave Communications*, 3 (2017), doi:10.1057/palcomms.2016.105; Adrian Mulligan/Louise Hall/Ellen Raphael, Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers, in: *Journal of the American Society for Information Science and Technology*, 64 (1/2013), 132-161, doi:10.1002/asi.22798; Aileen Fyfe et al., Managing the Growth of Peer Review at the Royal Society Journals, 1865-1965, in: *Science, Technology, & Human Values* 45 (3/2019), 405-429, doi:10.1177/0162243919862868; David Shatz, *Peer Review: A Critical Inquiry*, Issues in Academic Ethics, Lanham, MD, 2004, see 1.

8 Baldwin, Scientific Autonomy, 542.

9 Michèle Lamont, *How Professors Think: Inside the Curious World of Academic Judgment*, Cambridge, MA, 2009, see 247. See also Marcel C. LaFollette, *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing*, Berkeley, CA, 1992.

10 For just a selection, see Vladimir Batagelj/Anuška Ferligoj/Flaminio Squazzoni, The Emergence of a Field: A Network Analysis of Research on Peer Review, in: *Scientometrics*, 113 (1/2017), 503-532, doi:10.1007/s11192-017-2522-8; Jonathan Tennant/Tony Ross-Hellauer, The Limitations to Our Understanding of Peer Review, in: *SocArXiv*, 2019, doi:10.31235/osf.io/jq623; Erwin O. Smigel/H. Laurence Ross, Factors in the Editorial Decision, in: *The American Sociologist*, 5 (1/1970), 19-21; Charles M Bonjean/Jan Hullum, Reasons for Journal Rejection: An Analysis of 600 Manuscripts, in: *PS*, 11 (1978), 480-483; Elizabeth Ehrhardt Mustaine/Richard Tewksbury, Reviewers' Views on Reviewing: An Examination of the Peer Review Process in Criminal Justice, in: *Journal of Criminal Justice Education*, 19 (3/2008), 351-365, doi:10.1080/10511250802476178; Richard Tewksbury/ Elizabeth Ehrhardt Mustaine, Cracking Open the Black Box of the Manuscript Review Process: A Look Inside, in: *Justice Quarterly*, *Journal of Criminal Justice Education*, 23 (4/2012), 399-422, doi:10.1080/10511253.2011.653650; Omar Sabaj Meruane/Carlos González Vergara/Álvaro Pina-Stranger, What We Still Don't Know About Peer Review, in: *Journal of Scholarly Publishing*, 47 (2/2016), 180-212, doi:10.3138/jsp.47.2.180; Francisco Grimaldo/Ana Marušić/Flaminio Squazzoni, Fragments of Peer Review: A Quantitative Analysis of the Literature (1969-2015), in: *PLOS ONE*, 13 (2/2018), e0193148, doi:10.1371/journal.pone.0193148; Francisco Grimaldo/Mario Paolucci/Jordi Sabater-Mir, Reputation or Peer

Review? The Role of Outliers, in: *Scientometrics*, 116 (3/2018), 1421-1438, doi:10.1007/s11192-018-2826-3; Ann C. Weller, Editorial Peer Review: Its Strengths and Weaknesses, in: *Journal of the Medical Library Association*, 90 (1/2002); Lutz Bornmann, Peer Review and Bibliometrics: Potentials and Problems, in: Jung Cheol Shin/Robert K. Toutkoushian/Ulrich Teichler (eds.), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*, Dordrecht, 2011, 145-164, doi:10.1007/978-94-007-1116-7; Nyssa J. Silbiger/Amber D. Stubler, Unprofessional Peer Reviews Disproportionately Harm Underrepresented Groups in STEM, in: *PeerJ*, 7 (2019), e8247, doi:10.7717/peerj.8247; Flaminio Squazzoni, Peering Into Peer Review, in: *Sociologica*, 3 (2010), 1-27, doi:10.2383/33640; Cassidy R. Sugimoto and Blaise Cronin, Citation Gamesmanship: Testing for Evidence of Ego Bias in Peer Review, in: *Scientometrics*, 95 (3/2013), 851-862, doi:10.1007/s11192-012-0845-z; Steven N. Goodman, Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine, in: *Annals of Internal Medicine*, 121 (1/1994), 11, doi:10.7326/0003-4819-121-1-199407010-00003; Jean-Pierre E.N. Pierie/Henk C. Walvoort/A. John P.M. Overbeke, Readers' Evaluation of Effect of Peer Review and Editing on Quality of Articles in the Nederlands Tijdschrift Voor Geneeskunde, in: *The Lancet*, 348 (9040/1996), 1480-1483, doi:10.1016/S0140-6736(96)05016-7; Michael J. Mahoney, Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System, in: *Cognitive Therapy and Research*, 1 (2/1977), 161-175, doi:10.1007/BF01173636; Richard L. Kravitz et al., Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care?, in: *PLOS ONE*, 5 (4/2010), e10072, doi:10.1371/journal.pone.0010072; Daniel M. Herron, Is Expert Peer Review Obsolete? A Model Suggests That Post-Publication Reader Review May Exceed the Accuracy of Traditional Peer Review, in: *Surgical Endoscopy*, 26 (8/2012), 2275-2280, doi:10.1007/s00464-012-2171-1; Ferric C. Fang/Anthony Bowen/Arturo Casadevall, NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity, in: *ELife*, 5 (2016), e13323, doi:10.7554/eLife.13323; Amber E. Budden et al., Double-Blind Review Favours Increased Representation of Female Authors, in: *Trends in Ecology & Evolution*, 23 (1/2008), 4-6, doi:10.1016/j.tree.2007.07.008; Margaret E. Lloyd, Gender Factors in Reviewer Recommendations for Manuscript Publication, in: *Journal of Applied Behavior Analysis*, 23 (4/1990), 539-543, doi:10.1901/jaba.1990.23-539; Tom Tregenza, Gender Bias in the Refereeing Process?, in: *Trends in Ecology & Evolution*, 17 (8/2002), 349-350, doi:10.1016/S0169-5347(02)02545-4; E. Ernst/T. Kienbacher, Chauvinism, in: *Nature*, 15.08.1991, 560, doi:10.1038/352560b0; Ann M. Link, US and Non-US Submissions: An Analysis of Reviewer Bias, in: *JAMA*, 280 (3/1998), 246-247, doi:10.1001/jama.280.3.246; Paolo Dall'Aglio, Peer Review and Journal Models, in: *ArXiv*, 2006, URL: http://arxiv.org/abs/physics/0608307 [last accessed: Mar. 31, 2021]; Gilbert W. Gillespie/Daryl E. Chubin/George M. Kurzon, Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change, in: *Science, Technology, & Human Values*, 10 (3/1985), 44-54, doi:10.1177/016224398501000306; Stephen J. Ceci/Douglas P. Peters, Peer Review: A Study of Reliability, in: *Change*, 14 (6/1982), 44-48; Douglas P. Peters/Stephen J. Ceci, Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again, in: *Behavioral and Brain Sciences*, 5 (2/1982), 187-195, doi:10.1017/S0140525X00011183; Blaise Cronin, Vernacular and Vehicular Language, in: *Journal of the American Society for Information Science and Technology*, 60 (3/2009), 433-433, doi:10.1002/asi.21010; Joseph S. Ross et al., Effect of Blinded Peer Review on Abstract Acceptance, in: *JAMA*, 295 (14/2006), 1675-1680, doi:10.1001/jama.295.14.1675; G. D. L. Travis/H. M. Collins, New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System, in: *Science, Technology, & Human Values*, 16 (3/1991), 322-341, doi:10.1177/016224399101600303; Daryl E. Chubin/Edward

an ever-increasing number of studies that examine the language and mood of published academic articles, grant proposals, and editorials.[11] This is not surprising.After all, as Peter van den Besselaar, Hélène Schiffbaenker, Ulf Sandström, and Charlie Mom note, "[l]anguage embodies normative views about who/where we communicate about, and stereotypes about others are embedded and reproduced

J. Hackett, *Peerless Science: Peer Review and U.S. Science Policy*, Albany, NY, 1990; Mahoney, Publication Prejudices; A. H. Bardy, Bias in Reporting Clinical Trials, in: *British Journal of Clinical Pharmacology*, 46 (2/1998), 147-150, doi:10.1046/j.1365-2125.1998.00759.x; Kay Dickersin et al., Publication Bias and Clinical Trials, in: *Controlled Clinical Trials*, 8 (4/1987), 343-353, doi:10.1016/0197-2456(87)90155-3; Kay Dickersin/Yuan-I. Min/Curtis L. Meinert, Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards, in: *JAMA*, 267 (3/1992), 374-378, doi:10.1001/jama.1992.03480030052036; Daniele Fanelli, Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data, in: *PLoS ONE*, 5 (4/2010), doi:10.1371/journal.pone.0010271; John P. A. Ioannidis, Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials, in: *JAMA*, 279 (4/1998), 281-286, doi:10.1001/jama.279.4.281; Dalmeet Singh Chawla, Thousands of Grant Peer Reviewers Share Concerns in Global Survey, in: *Nature*, 15.10.2019, doi:10.1038/d41586-019-03105-2; Tony Ross-Hellauer, What Is Open Peer Review? A Systematic Review, in: *F1000Research*, 6 (2017), 588, doi:10.12688/f1000research.11369.2; Kanu Okike et al., Single-Blind vs Double-Blind Peer Review in the Setting of Author Prestige, in: *JAMA*, 31 (12/2016), 1315-1316, doi:10.1001/jama.2016.11014; Stanley Fish, No Bias, No Merit: The Case against Blind Submission, in: *PMLA*, 103 (5/1988), 739-748; Dakota Murray et al., Author-Reviewer Homophily in Peer Review, in: *BioRxiv* 400515 (2019), doi:10.1101/400515.

11 John Swales, *Genre Analysis: English in Academic and Research Settings*, Cambridge 1990; Rahime Nur Aktas/Viviana Cortes, Shell Nouns as Cohesive Devices in Published and ESL Student Writing, in: *Journal of English for Academic Purposes*, 7 (1/2008), 3-14, doi:10.1016/j.jeap.2008.02.002; Nigel Harwood, "I Hoped to Counteract the Memory Problem, but I Made No Impact Whatsoever": Discussing Methods in Computing Science Using I, in: *English for Specific Purposes*, 24 (3/2005), 243-267, doi:10.1016/j.esp.2004.10.002; Nigel Harwood, "Nowhere Has Anyone Attempted … In This Article I Aim to Do Just That": A Corpus-Based Study of Self-Promotional I and We in Academic Writing across Four Disciplines, in: *Journal of Pragmatics*, Focus-on Issue: Marking Discourse, 37 (8/2005), 1207-1231, doi:10.1016/j.pragma.2005.01.012; W. Shehzad, How to End an Introduction in a Computer Science Article: A Corpus-Based Approach, in E. Fitzpatrick (ed.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, Amsterdam 2015, 227-241, URL: https://brill.com/view/title/30110 [last accessed: Mar. 31, 2021]; Ulla Connor/Anna Mauranen, Linguistic Analysis of Grant Proposals: European Union Research Grants, in: *English for Specific Purposes*, 18 (1/1999), 47-62, doi:10.1016/S0889-4906(97)00026-4; Davide Simone Giannoni, Medical Writing at the Periphery: The Case of Italian Journal Editorials, in: *Journal of English for Academic Purposes*, 7 (2/2008), 97-107, doi:10.1016/j.jeap.2008.03.003. These examples are drawn from Theresa Lillis/Mary Jane Curry, The Politics of English, Language and Uptake: The Case of International Academic Journal Article Reviews, in: *AILA Review*, 28 (2015), 127-150, doi:10.1075/aila.28.06lil.

in language."[12] Indeed, a number of existing studies have examined the linguistic properties of peer review reports written by the authors themselves.[13]

2. Scaling our Understanding of Peer Review Using Neural Networks

As part of our Andrew W. Mellon Foundation-funded project, "Reading Peer Review," we were granted access to the archive of peer review reports at the world's largest, trans-disciplinary scientific journal, *PLOS ONE*.[14] This title has a radical policy on peer review that is different to many other journals. As Catriona J. MacCallum put it, "[t]he basis for […] decisions" in conventional pre-publication peer review "is inevitably subjective. The higher-profile science journals are consequently often accused of 'lottery reviewing,' a charge now aimed increasingly at the more specialist literature as well. Even after review, papers that are technically sound are often rejected on the basis of lack of novelty or advance."[15]

PLOS wanted to work differently. The peer-review procedure at *PLOS ONE* is predicated on the idea of "technical soundness" in which papers are judged according to whether or not their methods and procedures are thought to be solid, rather than on the basis of whether their contents are judged to be important.[16] This is a model in which reviewers should not accept or reject a paper for its novelty or significance, but should only assess its scientific validity. Hence, *PLOS ONE* will accept replication studies, null results (where an experiment didn't work), and other forms of scientific output that might not be published elsewhere. Thus, *PLOS ONE* was designed to "initiate a radical departure from the stifling constraints of

12 Peter van den Besselaar et al., Explaining Gender Bias in ERC Grant Selection – Life Sciences Case, in: *STI 2018 Conference Proceedings*, Leiden University, 2018, 314. See also Christian Burgers/Camiel J. Beukeboom, Stereotype Transmission and Maintenance Through Interpersonal Communication: The Irony Bias, in: *Communication Research*, 43 (3/2016), 414-441, doi:10.1177/0093650214534975; Camiel J. Beukeboom/Christian Burgers, Linguistic Bias, in: *Oxford Research Encyclopedia of Communication*, 2017, doi:10.1093/acrefore/9780190228613.013.439.

13 Peter Woods, *Successful Writing for Qualitative Researchers*, London 2006, 140-146; David Coniam, Exploring Reviewer Reactions to Manuscripts Submitted to Academic Journals, in: *System*, 40 (4/2012), 544-553, doi:10.1016/j.system.2012.10.002; Brian Paltridge, *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*, London 2017, 49-50.

14 The database was supplied to us electronically for offsite use and storage.

15 Catriona J. MacCallum, ONE for All: The Next Step for PLoS, in: *PLOS Biology*, 4 (11/2006), e401, doi:10.1371/journal.pbio.0040401.

16 PLOS, Journal Information, in: *PLOS ONE*, 2016, URL: http://www.plosone.org/static/information [last accessed: Mar. 31, 2021]. Note though that even this definition is contentious. See Richard Poynder, *PLoS ONE, Open Access, and the Future of Scholarly Publishing*, 2011, URL: https://richardpoynder.co.uk/PLoS_ONE.pdf [last accessed: Mar. 31, 2021].

this existing system." In this new model, it was claimed, "acceptance to publication [would] be a matter of days."[17]

We were provided by PLOS with a database consisting of 229,296 usable peer-review reports written between 2014 and 2016 from *PLOS ONE*. There were other reports in this database, but the identifiers assigned to them made it impossible to group these reports by review round and so these data were discarded. We wanted to know: how have the radical propositions that led to the creation of *PLOS ONE* affected actual practices on the ground in the title? Do PLOS reviewers behave as one might expect given the radicalism on which *PLOS ONE* was premised? And what can we learn about organisational change and its drivers? These broader questions are addressed in the book that came out of the project.[18]

In order to understand the composition of the archive and to communicate these findings in a way that does not cite any material directly, for reasons of data protection, we undertook a qualitative coding exercise (specifically domain and taxonomic descriptive coding) in which three research assistants collaboratively built a taxonomy of statements derived from the longer reviews.[19] In order to achieve intersubjective and, as far as possible, some intercultural linguistic assessment of the database, we had a diverse team of coders. Two of the research assistants were native English speakers based in London in the United Kingdom, although we note that the policed boundary of "native" and "non-native" speakers comes with both challenges for the specific study of peer review, but also with postcolonial overtones.[20] The third research assistant was an L2 English speaker (English as a second language) based in Lethbridge in Canada with significant social scientific background experience, including with this kind of coding work.

The goal of our coding exercise was to delve into the linguistic structures and semantic comment types that are used by reviewers, following previous work by Fortanet.[21] In order to militate against identity subjectivities in the coding process, each report was coded in triplicate – in which each research assistant worked at first individually but then regrouped to build collaborative consensus among the group on both sentiment and thematic classification – thereby constructing an

17 MacCallum, ONE for All: The Next Step for PLoS.

18 Eve et al.,*Peer Review and Institutional Change in Academia* .

19 David W. McCurdy/James P. Spradley/Dianna J. Shandy, *The Cultural Experience: Ethnography in Complex Society*, Long Grove, IL, 2005, 44-45.

20 250Karent Englander, Revision of Scientific Manuscripts by Non-Native English-Speaking Scientists in Response to Journal Editors' Language Critiques, in: *Journal of Applied Linguistics and Professional Practice*, 3 (2/2006), 129-161, doi:10.1558/japl.v3i2.129.

21 Inmaculada Fortanet, Evaluative Language in Peer Review Referee Reports, in: *Journal of English for Academic Purposes*, 7 (1/2008), 27-37, doi:10.1016/j.jeap.2008.02.004.

intersubjective agreement on the labels assigned for each term.[22] The downside of this approach is that, clearly, we traded accuracy for volume. This resulted in 78 triplicate tagged reports, consisting of 2,049 statements. Given the constraints on our resources, we hoped nonetheless that we could use the coded statements as a training resource for a neural network text classifier, to extrapolate up our claims about the archive.

Our coding exercise eventually built the following taxonomy of peer-review statements:

Table 1: The taxonomy of statements built for the Reading Peer Review project from the PLOS ONE database.

High-Level Category	Fine-Grained Category	Explication
Data	Data	A reference to results and/or data.
Data	Data commentary	A description of or commentary upon data. For instance, a reference to a chart's legend.
Data	Interpretation	Extrapolation from data. This category can overlap with data analysis/treatment.
Data	Analysis/treatment	How data are treated after collection. This includes data analysis and statistical analysis. It can also refer to secondary data (sets).
Data	Presentation	Includes reference to data display. Also includes comments on formatting, size of tables, redundancy of images, visibility of images, and size of the images.
Field of Knowledge	(Knowledge) Statement	A statement that the reviewer makes (about fact or community agreed notions). Does not apply to the reviewer paraphrasing the original article. Relates to knowledge claims by the reviewer and/or authors.

22 Such a triplicate coding approach had been used previously by Lutz Bornmann/Christophe Weymuth/Hans-Dieter Daniel, A Content Analysis of Referees' Comments: How Do Comments on Manuscripts Rejected by a High-Impact Journal and Later Published in Either a Low- or High-Impact Journal Differ?, in: *Scientometrics*, 83 (2/2010), 493-506, doi:10.1007/s11192-009-0011-4.

Field of Knowledge	Information for author(s)	Statements that indicate a reviewer's subjective opinion. E.g., "I consider it appropriate to…"
Field of Knowledge	Positioning	Reference to ways in which/ to what extent the authors position concepts/ideas in relation to others. Can also imply/require the Literature tag (see below).
Field of Knowledge	Literature	Explicit reference to secondary literature. Negative sentiment score in this category refers to misinterpretation or misrepresentation of literature, or lack of relevance of references employed.
Field of Knowledge	Revision	A comment on whether revisions have been made. A positive sentiment score in this category indicates revisions met while a negative means the opposite. This category also includes corrections and reference to subsequent/previous revisions.
Field of Knowledge	Holistic revision	Reviewer signals a range of issues to be fixed through revisions (referring to multiple categories).
Field of Knowledge	Fallibility	Instances where the reviewer admits they may not be correct in their opinion/criticism or admits inadequacy of and uncertainty around judgement.
Field of Knowledge	Tone	Tone of reviewer exhibits bias against non-western submission/language (patronising). Tone of reviewer exhibits *ad hominem* attack on author or team of researchers. Also used to denote overly familiar personal register/tone. Awarded appropriate sentiment score if tone implies praise or critique of manuscript.
Field of Knowledge	Potential/significance	A remark upon the significance of findings/data/results/work. This also includes the potential of contribution to knowledge or research; references to reproduction of experiments. Also used to flag poor scholarship and auto-plagiarism via a lack of novelty. Note that this category of "significance" should *not* be a criterion used for judgement of admission within the *PLOS ONE* ecosystem.

Expression	(English) Language	Reference to use of English, languages other than English, native/non-native speakers.
Expression	Typographical errors	Reference to surface level errors, including grammatical errors. Lack of consistency denotes strongly negative sentiment. Trivial typos are low sentiment score. Comments on punctuation are attributed using this tag.
Expression	Expression	Communicative quality - coherence of style and academic/scientific register. Choice of language. Rewording. Definitions of terms/acronyms.
Expression	Terminology	Use/deployment of subject-specific terminology. Can refer to accessibility of terms.
Expression	Cohesion	Comments on linkage between sections of paper in terms of correlation, structure and organisation.
Expression	Style	Comments on adhesion to house style.
Expression	Citation	Referencing and citation practice; includes lack of appropriate citation.
Expression	Summary	When a reviewer summarises or signals a section of paper. Also used as a form of transition before critique. Includes quotations from original text, including title.
Expression	Transition	A transitive statement which makes no reference to the manuscript. Includes notes to editors.
Methodology	Methodology	Broader approach to methods adopted. Also refers to rationale, justification or basis for research. Ethical issues/concerns.
Methodology	Statistics	In general and/or explicit reference to statistics including statistical tests. Explicit reference to or use of statistical tests such as Analysis of Variance (ANOVA), Student's T-Test, Pearson, correlation coefficients, Mann Whitney (package).
Methodology	Experimental design	Reference to a series of experiments, hypotheses, sample size, control groups, parameters, data collection tools, inferential/descriptive statistics, correlation, data modeling.

Methodology	Method	Refers to the description of method, including procedures, techniques, and discussion of advantageous alternatives.
Methodology	Limitations	Discussion of limitations.
Omission	Implied omission	Implies that something is missing without explicitly stating it.
Omission	Omission	Explicitly states that something is missing.
Omission	Accuracy	Comments on the accuracy of (data) description (& definitions). Can refer to factual or descriptive inaccuracy. Can also refer to (lack of) precision.
Omission	Elaboration	Request for more detail, information, clarification or precision. Different to omission in the sense that omission is about something that isn't there at all whereas this tag calls for supplementation.
Omission	Argument/analysis	Discussion of data/results. When there is "omission," it is unlikely that this tag will be used also.
Omission	Ambiguity	Reference to clarity, vagueness. Can connote positive (as clear, well worded etc.) as well as negative sentiment. Instances where something not clear to reviewer.
Omission	Argument	Pertains to clarity of argument - exposing point of view. Distinct from "argument/analysis" in that it deals with literature. Can also refer to the phrasing of an argument. Negative sentiment can refer to redundant or unconvincing argument. Explicit reference to logic or logical can imply this category. Also refers to the coherence of an argument. Claims implying criticism/agreement.
Omission	Implied criticism	Used for tagging questions from reviewers. Negative meaning/critique implicit. For instance, "Would this manuscript benefit from X?"
Section	Outcome	Publishability and suitability of results/data/findings. Relates to publishability of specific paper. Usually with reference to the admissibility criteria of *PLOS ONE*.

Section	Overarching comment	Used for tagging comments that broadly apply to the whole manuscript.
Section	Conclusion	Reference to the results of interpretation and/or analysis. Can also refer to results/findings. Reference to implications of results. Also refers to limitations of study.
Section	Abstract	Reference to the work's abstract.
Section	Appendix	Reference to an appendix in a work.

In order to conduct our computational reading test, we built a multi-class and multi-label text classifier based on the TenCent NeuralClassifier toolkit.[23] Although multi-class and multi-label text classification is a difficult task and even though we were only possessed of a relatively minimal, albeit robust, training set, the neural network was good at classifying certain types of input text. In particular, the network performed well at recognising requests for revision and/or outcome statements. For example, the generic statement "I do not recommend publication" was tagged by the network as pertaining to "revision" and "outcome." Some other types of broad statements were also accurately classified: "In particular I am left confused as to how the results fit in here" was marked as "ambiguity" and "cohesion" by the software.

However, the specific challenges of implementing an accurate classification system were many. First, the tagged data proved insufficient for these purposes. The labour-intensive processes of triplicate tagging gave us the confidence that we needed in the material that had been tagged, but this came at the expense of volume. Further, since each tagged statement was relatively short it was difficult to train natural-language processing toolkits to identify salient features; there is not

23 *An Open-Source Neural Hierarchical Multi-Label Text Classification Toolkit: Tencent/NeuralNLP-NeuralClassifier*, 2019, URL: https://github.com/Tencent/NeuralNLP-NeuralClassifier [last accessed: Mar. 31, 2021]. For more on paradigms of so-called "distant reading," see Franco Moretti, *Distant Reading*, London 2013; Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History*, London 2007; Ted Underwood, A Genealogy of Distant Reading, in: *Digital Humanities Quarterly*, 11 (2/2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html [last accessed: Mar. 31, 2021]; Andrew Piper, *Enumerations: Data and Literary Study*, Chicago, IL, 2018; Nan Z. Da, The Computational Case against Computational Literary Studies, in: *Critical Inquiry*, 45 (3/2019), 601-639, doi:10.1086/702594; Martin Paul Eve, *Close Reading With Computers: Textual Scholarship, Computational Formalism, and David Mitchell's* Cloud Atlas, Stanford, CA, 2019; Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change*, Chicago, IL, 2019.

a huge volume in each case for the network to identify. As above, there are also instances where we did not find and tag particular types of statement, such as those pertaining to ethics. Finally, since each statement was written by different authors (reviewers), with different primary languages, the strength of these linguistic differentiations – as opposed to the words used within different types of classificatory statements – appear to be pulled to the fore.[24] As such, this study is limited to a relatively small sample size with a relatively good accuracy level within that sample.

Hence, while the network appears to work well at classifying statements that have appeared in almost all reviews – for instance, the outcome example above – it performed poorly at identifying less frequent types, such as "fallibility." The network was unable, for example, to ascribe a label to the statement "I must confess that I am not an expert with respect to these methods," a clear assertion of fallibility. Further, various statements around originality were not tagged with any accuracy. For instance, "There is nothing technically wrong with the paper, but it is not that original" was marked as an "overarching comment," which is a fair assessment. However, no label noting that this was a statement about originality or novelty was ascribed, regardless of the training parameters that we fed to the network.

A further method for "distant reading" the corpus is, of course, to conduct a simple text search through the reviews. This is how we identified the "missing" statements on ethics to which we earlier referred. This can be useful to find examples of specific kinds of practice. For instance, to identify overly aggressive reports we used a simple tool, "grep" (globally search a regular expression and print), to look for instances of the word "useless" in the top-800 longest reports. This yielded harsh reports that included phrases such as "Fig 12 is almost useless"; "the null model seems somewhat useless"; "remove the repeated useless sentences"; "I found the [secondary subject matter] results to be EXTREMELY distracting, and essentially useless"; "this work appears to be all but useless" and so on. What such searching cannot tell us, though, is the prevalence of such practices. For instance, the above examples were found using an extremely simple keyword search pulled from the top of our heads. There will be many instances of *ad hominem* or vicious attack that use different terms and the only reliable way, at present, to identify these is to read and to tag the reports themselves.

In addition to this, capital letters (as in the above "EXTREMELY" example) are relatively easy to detect and sometimes indicate strong sentiment of one kind or another. However, detection of these is not as simple as a regular expression ("\b[A-Z][A-Z]+\b") as this will also pull out the many acronyms used in scientific practice

24 For more on authorship signals among others, see Sarah Allison et al., *Quantitative Formalism: An Experiment*, Stanford 2011, URL: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [last accessed: Mar. 31, 2021]; Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Urbana, IL, 2013.

("BDNF," "ROC" etc.). It is also not clear that capital letters denote strong sentiment in one direction or another; "WOW" can indicate "WOW this is a brilliant paper," but it can equally likely specify "WOW, this paper was terrible." Furthermore, on occasion capital letters are used to denote section headings and/or specific portions of a paper ("in the METHOD section"). In this way, the extraction of capital letters – without a pre-built blacklist of words to exclude – is likely to result in many false positives.

A further way of exploring the corpus at scale is to use the techniques of "topic modeling," a technique that finds co-occurring words and bundles them into so-called "topics." Hence, a "topic" in a recipe book might be: "sugar," "icing," "sweet," "jelly." Topic modeling generally uses a process called "Latent Dirichlet Allocation" (LDA) in order to cluster together terms that probabilistically co-occur in similar contexts. This is a useful way to explore a dataset and to infer the groups of terms that most frequently crop up together; that is, which "topics" are explored within a corpus (a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics).[25] As Ben Schmidt notes, this approach "does a good job giving an overview of the contents of large textual collections; it can provide some intriguing new artifacts to study; and it even holds [...] some promise for structuring non-lexical data like geographic points."[26]

However, LDA is also a dangerous method. This is because there is no way to infer *why* topics have been grouped together. In particular, surprising groupings that appear to exhibit coherence may not be as well bound as we would like to think. As Schmidt continues,

> still, excitement about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions LDA creates will be helpful, and how often merely surprising. A poorly supervised machine learning algorithm is like a bad research assistant. It might produce some unexpected constellations that show flickers of deeper truths; but it will also produce tedious, inexplicable, or misleading results.[27]

That said, as an exploratory exercise that others may wish to take further, we produced a twenty-topic model using the MALLET tool based on the same corpus of 800 reports using the default hyperparameters and with stop words excluded. The results for this are shown in Table 2.

---

25 David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, 3 (2003), 993-1022.

26 Benjamin Schmidt, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities*, 2 (1/2012), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 1, 2021].

27 Schmidt, Words Alone: Dismantling Topic Models in the Humanities.

Table 2: A topic model of the 800 longest reports in our database of reviews at PLOS ONE.

Topic Number	Topic Terms
1	authors manuscript paper study comments data review results current previous discussion work addressed major provide reviewer research information important studies
2	interests competing samples genetic dna populations population gene pcr table strains sequences figure structure loci sequencing individuals analysis chromosomes number
3	paper data case make point time clear important find fact understand general evidence e.g i.e model number high approach literature
4	line lines page sentence paragraph change suggest section figure results text table discussion reference manuscript remove delete add replace information
5	genes gene expression analysis number sequences sequence genome rna authors expressed species biological fig methods transcripts data results transcriptome proteins
6	model method models paper approach parameters system distribution set network number dataset parameter distributions author proposed performance simulations equation networks
7	species study habitat lines area model spatial population areas line fish variables sites models distance size individuals data prey year
8	study treatment patients group participants trial intervention studies outcome pain analysis groups clinical outcomes research control care reported patient measures
9	patients study blood clinical studies disease plasma levels authors activity acute group tissue serum negative ace sensitivity mir cortisol healthy
10	social behavior females males male authors individuals scer_scrt lcl study female group behaviors human sex calls behaviour pointing sexual attention
11	fire area page e.g subject trees specific stand bone signals motion intensity study science frequency subjects biochemistry atmospheric stands fires

12	food hsv animals infection mice diet authors bees response dose resistance treatment weight pigs bacteria immune intake group larvae virus
13	participants task authors condition experiment effect stimuli results memory performance responses effects experiments stimulation response conditions experimental visual trials learning
14	cells authors cell figure fig expression data shown protein experiments levels control show mrna mice state manuscript results effect antibody
15	species phylogenetic taxa tree xxx based diversity sequences analysis clade character genus specimens support trees present phylogeny group taxonomic found
16	age health risk table page women population study prevalence model factors children results years variables cases analysis paragraph hiv year
17	null partly disease n/a cancer vaccine hpv page cervical women vaccination safety doi group gardasil adverse rate don't map human
19	data results authors analysis table methods study differences significant discussion statistical figure time section values effect test information sample size
19	species water soil temperature change growth plant plants concentrations climate samples concentration biomass sites fish site study conditions carbon community
20	fig protein binding manuscript light proteins figure structure shown images domain mutant site sequence image residues cry region structures pax

Some of these topics appear easy to interpret. Group one, comprising "authors," "manuscript," "paper" and so on cluster meta-statements about the paper, its submission, and the review process. It is curious, though, that "important" should find its way into the work here (although "not important" would also trigger this, so no sentiment value should be inferred). Certainly, there are multiple contexts within which the word "important" can appear. For instance, "it is important that the author address these points" is as likely a statement as "this paper is extremely important." Nonetheless, given that *PLOS ONE* specifically disavows importance from its criteria, it is significant that the term should appear so prominently among statements that are otherwise common in opening gambits.

Topic four, by contrast, clearly pertains to the mechanics of a paper and suggested corrections. Its functional emphasis on the "line," "sentence," "page," "figure," "table" and so forth – coupled with "suggest," "add," "replace," and "delete" is the archetypical set of terms that we find in revision requests. In our experience of tagging, such language is prevalent during line-by-line commentaries that usually take the form of "line 123: suggest adding X."

Several of the topics relate to subject matter that is clearly of disciplinary interest to and prominent within *PLOS ONE*. Topics two and five, for instance, are

concerned with genetics. Topic seven appears to be biology; topics eight and nine circle around medicine and clinical trials; topic ten relates to reproduction, mating, and sexuality; topic twelve seems to indicate dietary behaviours; topic fifteen is about biological taxonomies; topic sixteen is on ageing; and so on.

Of course, anyone who knows anything about *PLOS ONE* might have guessed that such terms would cluster together and be found as separate strata. For us, the more useful indicators are not the subject groupings, which one would expect, but the functional parameters. We can anticipate scenarios under which knowledge of the distinct linguistic layer of line-by-line corrections, for instance, could be extracted and formed into editorial "to-do" lists. We could also imagine automatic detection of appraisal of novelty and importance, and a flagging system that could warn the editor of such an approach (and that it should not be used in the judgement of articles). The challenge, as ever with topic modeling, though, is that the topics that seem clearly thematically clustered are obvious, while the ones that exhibit less coherence (say, topic 20) are baffling.

3. Conclusions

The resources required to train a neural network for accurate multi-class and multi-label identification over the whole corpus were greater than those available to us. Indeed, the quality of the classification engine is directly proportional to the volume and accuracy of the training data. While our exercise yielded insights – particularly through LDA and plain-text search methods – to classify accurately the whole corpus and then to make deductive statements with any certainty requires a great deal more work at the corpus preparation stage. In short, while our experiment in using machine learning to examine the entire corpus of reviews might have worked well for certain types of statement, such as those pertaining to outcome, the uncertainty around, and low levels of, accuracy mean that any quantitative analysis based on the broader corpus, read at distance, would be unacceptably imprecise. Nonetheless, the moments of success in the network seem to indicate that those with broader resources for tagging and access to a large corpus of review reports might, in future, see some benefit in using this approach. For instance, we can envisage situations where such a network could detect hostile tone and warn the reviewer that s/he is being overly harsh or *ad hominem*. We could also imagine situations in which such a classifier could distinguish reviews that were structured in an unusual/idiosyncratic manner. While this would not rule out the review from being useful, it could give an indication that the reviewer is inexperienced or working away from norms of the form. That said, if the network were used by publishers to insist on normative practices in review, then this could stifle new ways of writing and operating.

Our experiments in using machine learning to read at scale taught us that, in essence, the results are only ever as good as the data that are fed in. Garbage in? Garbage out. We did not have garbage; our training data were robust, but they were not voluminous. However, to build an ultra-robust and massive corpus that would have made the AI methods work at scale would have required more labour effort than we could afford. This is perhaps the lesson that our AI approach taught us: it is resourcing and people that are the scarcities, not technology.

Bibliography

Aktas, Rahime Nur/Cortes, Viviana, Shell Nouns as Cohesive Devices in Published and ESL Student Writing, in: *Journal of English for Academic Purposes*, 7 (1/2008), 3-14, doi:10.1016/j.jeap.2008.02.002.

Alberts, Bruce/Hanson, Brooks/Kelner, Katrina L., Reviewing Peer Review, in: *Science*, 321 (5885/2008), 15, doi:10.1126/science.1162115.

Allison, Sarah, et al., *Quantitative Formalism: An Experiment*, Stanford 2011, URL: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf [last accessed: Mar. 31, 2021].

An Open-Source Neural Hierarchical Multi-Label Text Classification Toolkit: Tencent/NeuralNLP-NeuralClassifier, 2019, URL: https://github.com/Tencent/NeuralNLP-NeuralClassifier [last accessed: Mar. 31, 2021].

Bakanic, Von/McPhail, Clark/Simon, Rita J., Mixed Messages: Referees' Comments on the Manuscripts They Review, in: *The Sociological Quarterly*, 30 (4/1989), 639-654, URL: http://www.jstor.org/stable/4121469.

Baldwin, Melinda, Scientific Autonomy, Public Accountability, and the Rise of "Peer Review" in the Cold War United States, in: *Isis*, 109 (3/2018), 538-558, doi:10.1086/700070.

Baldwin, Melinda, In Referees We Trust?, in: *Physics Today*, 70 (2/2017), 44-49, doi:10.1063/PT.3.3463.

Bardy, A. H., Bias in Reporting Clinical Trials, in: *British Journal of Clinical Pharmacology*, 46 (2/1998), 147-150, doi:10.1046/j.1365-2125.1998.00759.x.

Batagelj, Vladimir/Ferligoj, Anuška/Squazzoni, Flaminio, The Emergence of a Field: A Network Analysis of Research on Peer Review, in: *Scientometrics*, 113 (1/2017), 503-532, doi:10.1007/s11192-017-2522-8.

Besselaar, Peter van den, et al., Explaining Gender Bias in ERC Grant Selection – Life Sciences Case, in: *STI 2018 Conference Proceedings*, Leiden University, 2018, 346-352.

Beukeboom, Camiel J./Burgers, Christian, Linguistic Bias, in: *Oxford Research Encyclopedia of Communication*, 2017, doi:10.1093/acrefore/9780190228613.013.439.

Blei, David M./Ng, Andrew Y./Jordan, Michael I., Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, 3 (2003), 993-1022.

Bonjean, Charles M/Hullum, Jan, Reasons for Journal Rejection: An Analysis of 600 Manuscripts, in: *PS*, 11 (1978), 480-483.

Bornmann, Lutz, Peer Review and Bibliometrics: Potentials and Problems, in: Jung Cheol Shin/Robert K. Toutkoushian/Ulrich Teichler (eds.), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*, Dordrecht, 2011, 145-164, doi:10.1007/978-94-007-1116-7.

Bornmann, Lutz, Scientific Peer Review, in: *Annual Review of Information Science and Technology*, 45 (1/2011), 197-245, doi:10.1002/aris.2011.1440450112.

Bornmann, Lutz/Daniel, Hans-Dieter, The Effectiveness of the Peer Review Process: Inter-Referee Agreement and Predictive Validity of Manuscript Refereeing at *Angewandte Chemie*, in: *Angewandte Chemie International Edition*, 47 (38/2008), 7173-7178, doi:10.1002/anie.200800513.

Bornmann, Lutz/Weymuth, Christophe/Daniel, Hans-Dieter, A Content Analysis of Referees' Comments: How Do Comments on Manuscripts Rejected by a High-Impact Journal and Later Published in Either a Low- or High-Impact Journal Differ?, in: *Scientometrics*, 83 (2/2010), 493-506, doi:10.1007/s11192-009-0011-4.

Brainard, Jeffrey/You, Jia, What a Massive Database of Retracted Papers Reveals about Science Publishing's "Death Penalty," in: *Science | AAAS*, 25.10.2018, URL: https://www.sciencemag.org/news/2018/10/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty [last accessed: Mar. 31, 2021].

Brembs, Björn/Button, Katherine/Munafò, Marcus, Deep Impact: Unintended Consequences of Journal Rank, in: *Frontiers in Human Neuroscience*, 7 (2013), 1-12, doi:10.3389/fnhum.2013.00291.

Budden, Amber E., et al., Double-Blind Review Favours Increased Representation of Female Authors, in: *Trends in Ecology & Evolution*, 23 (1/2008), 4-6, doi:10.1016/j.tree.2007.07.008.

Burgers, Christian/Beukeboom, Camiel J., Stereotype Transmission and Maintenance Through Interpersonal Communication: The Irony Bias, in: *Communication Research*, 43 (3/2016), 414-441, doi:10.1177/0093650214534975.

Ceci, Stephen J./Peters, Douglas P., Peer Review: A Study of Reliability, in: *Change*, 14 (6/1982), 44-48.

Chawla, Dalmeet Singh, Thousands of Grant Peer Reviewers Share Concerns in Global Survey, in: *Nature*, 15.10.2019, doi:10.1038/d41586-019-03105-2.

Chubin, Daryl E./Hackett, Edward J., *Peerless Science: Peer Review and U.S. Science Policy*, Albany, NY, 1990.

Coniam, David, Exploring Reviewer Reactions to Manuscripts Submitted to Academic Journals, in: *System*, 40 (4/2012), 544-553, doi:10.1016/j.system.2012.10.002.

Connor, Ull/Mauranen, Anna, Linguistic Analysis of Grant Proposals: European Union Research Grants, in: *English for Specific Purposes*, 18 (1/1999), 47-62, doi:10.1016/S0889-4906(97)00026-4.

Cronin, Blaise, Vernacular and Vehicular Language, in: *Journal of the American Society for Information Science and Technology*, 60 (3/2009), 433-433, doi:10.1002/asi.21010.

Da, Nan Z., The Computational Case against Computational Literary Studies, in: *Critical Inquiry*, 45 (3/2019), 601-639, doi:10.1086/702594.

Dall'Aglio, Paolo, Peer Review and Journal Models, in: *ArXiv*, 2006, URL: http://arxiv.org/abs/physics/0608307 [last accessed: Mar. 31, 2021].

Daniel, Hans-Dieter, *Guardians of Science: Fairness and Reliability of Peer Review*, Weinheim, 1993.

Dickersin, Kay, et al., Publication Bias and Clinical Trials, in: *Controlled Clinical Trials*, 8 (4/1987), 343-353, doi:10.1016/0197-2456(87)90155-3.

Dickersin, Kay/Min, Yuan-I./Meinert, Curtis L., Factors Influencing Publication of Research Results: Follow-up of Applications Submitted to Two Institutional Review Boards, in: *JAMA*, 267 (3/1992), 374-378, doi:10.1001/jama.1992.03480030052036.

Englander, Karent, Revision of Scientific Manuscripts by Non-Native English-Speaking Scientists in Response to Journal Editors' Language Critiques, in: *Journal of Applied Linguistics and Professional Practice*, 3 (2/2006), 129-161, doi:10.1558/japl.v3i2.129.

Ernst, E./Kienbacher, T., Chauvinism, in: *Nature*, 15.08.1991, 560, doi:10.1038/352560b0.

Eve, Martin Paul, *Close Reading With Computers: Textual Scholarship, Computational Formalism, and David Mitchell's* Cloud Atlas, Stanford, CA, 2019.

Eve, Martin Paul, et al., *Peer Review and Institutional Change in Academia*, Cambridge 2021.

Fanelli, Daniele, Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data, in: *PLoS ONE*, 5 (4/2010), doi:10.1371/journal.pone.0010271.

Fang, Ferric C./Bowen, Anthony/Casadevall, Arturo, NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity, in: *ELife*, 5 (2016), e13323, doi:10.7554/eLife.13323.

Fish, Stanley, No Bias, No Merit: The Case against Blind Submission, in: *PMLA*, 103 (5/1988), 739-748.

Fogg, Louis/Fiske, Donald W., Foretelling the Judgments of Reviewers and Editors, in: *American Psychologist*, 48 (3/1993), 293-294, doi:10.1037/0003-066X.48.3.293.

Fortanet, Inmaculada, Evaluative Language in Peer Review Referee Reports, in: *Journal of English for Academic Purposes*, 7 (1/2008), 27-37, doi:10.1016/j.jeap.2008.02.004.

Fyfe, Aileen, et al., Managing the Growth of Peer Review at the Royal Society Journals, 1865-1965, in: *Science, Technology, & Human Values* 45 (3/2019), 405-429, doi:10.1177/0162243919862868.

Giannoni, Davide Simone, Medical Writing at the Periphery: The Case of Italian Journal Editorials, in: *Journal of English for Academic Purposes*, 7 (2/2008), 97-107, doi:10.1016/j.jeap.2008.03.003.

Gillespie, Gilbert W./Chubin, Daryl E./Kurzon, George M., Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change, in: *Science, Technology, & Human Values*, 10 (3/1985), 44-54, doi:10.1177/016224398501000306.

Godlee, Fiona/Smith, Jane/Marcovitch, Harvey, Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent, in: *BMJ*, 342 (2011), c7452, doi:10.1136/bmj.c7452.

Goodman, Steven N., Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine, in: *Annals of Internal Medicine*, 121 (1/1994), 11, doi:10.7326/0003-4819-121-1-199407010-00003.

Grimaldo, Francisco/Marušić, Ana/Squazzoni, Flaminio, Fragments of Peer Review: A Quantitative Analysis of the Literature (1969-2015), in: *PLOS ONE*, 13 (2/2018), e0193148, doi:10.1371/journal.pone.0193148.

Grimaldo, Francisco/Paolucci, Mario/Sabater-Mir, Jordi, Reputation or Peer Review? The Role of Outliers, in: *Scientometrics*, 116 (3/2018), 1421-1438, doi:10.1007/s11192-018-2826-3.

Hames, Irene, *Peer Review and Manuscript Management of Scientific Journals Guidelines for Good Practice*, Malden, MA, 2007.

Harwood, Nigel, "I Hoped to Counteract the Memory Problem, but I Made No Impact Whatsoever": Discussing Methods in Computing Science Using I, in: *English for Specific Purposes*, 24 (3/2005), 243-267, doi:10.1016/j.esp.2004.10.002.

Harwood, Nigel, "Nowhere Has Anyone Attempted … In This Article I Aim to Do Just That": A Corpus-Based Study of Self-Promotional I and We in Academic Writing across Four Disciplines, in: *Journal of Pragmatics*, Focus-on Issue: Marking Discourse, 37 (8/2005), 1207-1231, doi:10.1016/j.pragma.2005.01.012.

Herron, Daniel M., Is Expert Peer Review Obsolete? A Model Suggests That Post-Publication Reader Review May Exceed the Accuracy of Traditional Peer Review, in: *Surgical Endoscopy*, 26 (8/2012), 2275-2280, doi:10.1007/s00464-012-2171-1.

Ingelfinger, Franz J., Peer Review in Biomedical Publication, in: *The American Journal of Medicine*, 56 (1974), 686-692.

Ioannidis, John P. A., Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials, in: *JAMA*, 279 (4/1998), 281-286, doi:10.1001/jama.279.4.281.

Jockers, Matthew L., *Macroanalysis: Digital Methods and Literary History*, Urbana, IL, 2013.

Kravitz, Richard L., et al., Editorial Peer Reviewers' Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care?, in: *PLOS ONE*, 5 (4/2010), e10072, doi:10.1371/journal.pone.0010072.

LaFollette, Marcel C., *Stealing into Print: Fraud, Plagiarism, and Misconduct in Scientific Publishing*, Berkeley, CA, 1992.

Lamont, Michèle, *How Professors Think: Inside the Curious World of Academic Judgment*, Cambridge, MA, 2009.

Lillis, Theresa/Curry, Mary Jane, The Politics of English, Language and Uptake: The Case of International Academic Journal Article Reviews, in: *AILA Review*, 28 (2015), 127-150, doi:10.1075/aila.28.06lil.

Link, Ann M., US and Non-US Submissions: An Analysis of Reviewer Bias, in: *JAMA*, 280 (3/1998), 246-247, doi:10.1001/jama.280.3.246.

Lloyd, Margaret E., Gender Factors in Reviewer Recommendations for Manuscript Publication, in: *Journal of Applied Behavior Analysis*, 23 (4/1990), 539-543, doi:10.1901/jaba.1990.23-539.

Lock, Stephen, *A Difficult Balance: Editorial Peer Review in Medicine*, Philadelphia, PA, 1986.

MacCallum, Catriona J., ONE for All: The Next Step for PLoS, in: *PLOS Biology*, 4 (11/2006), e401, doi:10.1371/journal.pbio.0040401.

Mahoney, Michael J., Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System, in: *Cognitive Therapy and Research*, 1 (2/1977), 161-175, doi:10.1007/BF01173636.

McCurdy, David W./Spradley, James P./Shandy, Dianna J., *The Cultural Experience: Ethnography in Complex Society*, Long Grove, IL, 2005.

Meruane, Omar Sabaj/Vergara, Carlos González/Pina-Stranger, Álvaro, What We Still Don't Know About Peer Review, in: *Journal of Scholarly Publishing*, 47 (2/2016), 180-212, doi:10.3138/jsp.47.2.180.

Moore, Samuel, et al., Excellence R Us: University Research and the Fetishisation of Excellence, in: *Palgrave Communications*, 3 (2017), doi:10.1057/palcomms.2016.105.

Moretti, Franco, *Distant Reading*, London 2013.

Moretti, Franco, *Graphs, Maps, Trees: Abstract Models for Literary History*, London 2007.

Mulligan, Adrian/Hall, Louise/Raphael, Ellen, Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers, in: *Journal of the American Society for Information Science and Technology*, 64 (1/2013), 132-161, doi:10.1002/asi.22798.

Murray, Dakota, et al., Author-Reviewer Homophily in Peer Review, in: *BioRxiv* 400515 (2019), doi:10.1101/400515.

Mustaine, Elizabeth Ehrhardt/Tewksbury, Richard, Reviewers' Views on Reviewing: An Examination of the Peer Review Process in Criminal Justice, in: *Journal of Criminal Justice Education*, 19 (3/2008), 351-365, doi:10.1080/10511250802476178.

Okike, Kanu, et al., Single-Blind vs Double-Blind Peer Review in the Setting of Author Prestige, in: *JAMA*, 31 (12/2016), 1315-1316, doi:10.1001/jama.2016.11014.

Paltridge, Brian, *The Discourse of Peer Review: Reviewing Submissions to Academic Journals*, London 2017.

Peters, Douglas P./Ceci, Stephen J., Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again, in: *Behavioral and Brain Sciences*, 5 (2/1982), 187-195, doi:10.1017/S0140525X00011183.

Petty, Richard E./Fleming, Monique A./Fabrigar, Leandre R., The Review Process at PSPB: Correlates of Interreviewer Agreement and Manuscript Acceptance, in: *Personality and Social Psychology Bulletin*, 25 (2/1999), 188-203, doi:10.1177/0146167299025002005.

Pierie, Jean-Pierre E.N., Walvoort, Henk C./Overbeke, A. John P.M., Readers' Evaluation of Effect of Peer Review and Editing on Quality of Articles in the Nederlands Tijdschrift Voor Geneeskunde, in: *The Lancet*, 348 (9040/1996), 1480-1483, doi:10.1016/S0140-6736(96)05016-7.

Piper, Andrew, *Enumerations: Data and Literary Study*, Chicago, IL, 2018.

PLOS, Journal Information, in: *PLOS ONE*, 2016, URL: http://www.plosone.org/static/information [last accessed: Mar. 31, 2021].

Poynder, Richard, *PLoS ONE, Open Access, and the Future of Scholarly Publishing*, 2011, URL: https://richardpoynder.co.uk/PLoS_ONE.pdf [last accessed: Mar. 31, 2021].

Retraction Watch, URL: https://retractionwatch.com/ [last accessed: Mar. 31, 2021].

Ross, Joseph S., et al., Effect of Blinded Peer Review on Abstract Acceptance, in: *JAMA*, 295 (14/2006), 1675-1680, doi:10.1001/jama.295.14.1675.

Ross-Hellauer, Tony, What Is Open Peer Review? A Systematic Review, in: *F1000Research*, 6 (2017), 588, doi:10.12688/f1000research.11369.2.

Rothwell, Peter M./Martyn, Christopher N., Reproducibility of Peer Review in Clinical Neuroscience, in: *Brain*, 123 (9/2000), 1964-1969, doi:10.1093/brain/123.9.1964.

Schmidt, Benjamin, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities*, 2 (1/2012), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 1, 2021].

Shatz, David, *Peer Review: A Critical Inquiry*, Issues in Academic Ethics, Lanham, MD, 2004.

Shehzad, W., How to End an Introduction in a Computer Science Article: A Corpus-Based Approach, in E. Fitzpatrick (ed.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*, Amsterdam 2015, 227-241, URL: https://brill.com/view/title/30110 [last accessed: Mar. 31, 2021].

Silbiger, Nyssa J./Stubler, Amber D., Unprofessional Peer Reviews Disproportionately Harm Underrepresented Groups in STEM, in: *PeerJ*, 7 (2019), e8247, doi:10.7717/peerj.8247.

Smigel, Erwin O./Ross, H. Laurence, Factors in the Editorial Decision, in: *The American Sociologist*, 5 (1/1970), 19-21.

Smith, Richard, Classical Peer Review: An Empty Gun, in: *Breast Cancer Research*, 12 (4/2010), S13, doi:10.1186/bcr2742.

Smith, Richard, Peer Review: A Flawed Process at the Heart of Science and Journals, in: *Journal of the Royal Society of Medicine*, 99 (4/2006), 178-182.

Squazzoni, Flaminio, Peering Into Peer Review, in: *Sociologica*, 3 (2010), 1-27, doi:10.2383/33640.

Sternberg, Robert J., et al., Getting in: Criteria for Acceptance of Manuscripts in Psychological Bulletin, 1993-1996, in: *Psychological Bulletin*, 121 (2/1997), 321-323, doi:10.1037/0033-2909.121.2.321.

Sugimoto, Cassidy R./Cronin, Blaise, Citation Gamesmanship: Testing for Evidence of Ego Bias in Peer Review, in: *Scientometrics*, 95 (3/2013), 851-862, doi:10.1007/s11192-012-0845-z.

Swales, John, *Genre Analysis: English in Academic and Research Settings*, Cambridge 1990.

Tennant, Jonathan/Ross-Hellauer, Tony, The Limitations to Our Understanding of Peer Review, in: *SocArXiv*, 2019, doi:10.31235/osf.io/jq623.

Tewksbury, Richard/Ehrhardt Mustaine, Elizabeth, Cracking Open the Black Box of the Manuscript Review Process: A Look Inside, in: *Justice Quarterly*, *Journal of Criminal Justice Education*, 23 (4/2012), 399-422, doi:10.1080/10511253.2011.653650.

Travis, G. D. L./Collins, H. M., New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System, in: *Science, Technology, & Human Values*, 16 (3/1991), 322-341, doi:10.1177/016224399101600303.

Tregenza, Tom, Gender Bias in the Refereeing Process?, in: *Trends in Ecology & Evolution*, 17 (8/2002), 349-350, doi:10.1016/S0169-5347(02)02545-4.

Underwood, Ted, A Genealogy of Distant Reading, in: *Digital Humanities Quarterly*, 11 (2/2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html [last accessed: Mar. 31, 2021].

Underwood, Ted, *Distant Horizons: Digital Evidence and Literary Change*, Chicago, IL, 2019.

Wakefield, A. J.et al., RETRACTED: Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children, in: *The Lancet*, 351 (9103/1998), 637-641, doi:10.1016/S0140-6736(97)11096-0.

Weller, Ann C., Editorial Peer Review: Its Strengths and Weaknesses, in: *Journal of the Medical Library Association*, 90 (1/2002).

Woods, Peter, *Successful Writing for Qualitative Researchers*, London 2006.

Zuckerman, Harriet/Merton, Robert K., Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System, in: *Minerva*, 9 (1/1971), 66-100, doi:10.1007/BF01553188.