QSS

an open access journal

Check for updates

Citation: Vincent-Lamarre, P., & Larivière, V. (2021). Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. *Quantitative Science Studies*, *2*(2), 662–677. https://doi.org/10.1162/qss_a_00125

DOI:
https://doi.org/10.1162/qss_a_00125

Peer Review:
https://publons.com/publon/10.1162/qss_a_00125

Supporting Information:
https://doi.org/10.1162/qss_a_00125

Received: 18 March 2020
Accepted: 17 December 2020

Corresponding Author:
Vincent Larivière
vincent.lariviere@umontreal.ca

Handling Editor:
Ludo Waltman

Copyright: © 2021 Philippe Vincent-Lamarre and Vincent Larivière. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

RESEARCH ARTICLE

Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome

Philippe Vincent-Lamarre[1,2] and **Vincent Larivière**[1]

[1]École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Canada
[2]École de psychologie, Université d'Ottawa, Ottawa, Canada

Keywords: artificial intelligence, peer review, readability

ABSTRACT

We analyzed a data set of scientific manuscripts that were submitted to various conferences in artificial intelligence. We performed a combination of semantic, lexical, and psycholinguistic analyses of the full text of the manuscripts and compared them with the outcome of the peer review process. We found that accepted manuscripts scored lower than rejected manuscripts on two indicators of readability, and that they also used more scientific and artificial intelligence jargon. We also found that accepted manuscripts were written with words that are less frequent, that are acquired at an older age, and that are more abstract than rejected manuscripts. The analysis of references included in the manuscripts revealed that the subset of accepted submissions were more likely to cite the same publications. This finding was echoed by pairwise comparisons of the word content of the manuscripts (i.e., an indicator of semantic similarity), which were more similar in the subset of accepted manuscripts. Finally, we predicted the peer review outcome of manuscripts with their word content, with words related to machine learning and neural networks positively related to acceptance, whereas words related to logic, symbolic processing, and knowledge-based systems negatively related to acceptance.

1. INTRODUCTION

Peer review is a fundamental component of the scientific enterprise and acts as one of the main sources of quality control of the scientific literature (Ziman, 2002). The primary form of peer review occurs before publication (Wakeling, Willett et al., 2019) and it is often considered as a stamp of approval from the scientific community (Mayden, 2012; Mulligan, 2005). Peer-reviewed publications have a considerable weight in the attribution of research and academic resources (McKiernan, Schimanski et al., 2019; Moher, Naudet et al., 2018; Tregellas, Smucny et al., 2018).

One of the main concerns about peer review is its lack of reliability (Bailar, 1991; Cicchetti, 1991; Lee, 2012). Most studies on the topic find that agreement between reviewers is barely greater than chance (Bornmann, Mutz, & Daniel, 2010; Forscher, Brauer et al., 2019; Price, 2014), which highlights the considerable amount of subjectivity involved in the process. This leaves room for a lot of potential sources of bias, which have been reported in several studies (De Silva & Vance, 2017; Lee, Sugimoto et al., 2013; Murray, Siler et al., 2018). A potential silver lining is that it appears that the process has some validity. For instance, articles accepted at a general medicinal journal (Jackson, Srinivasan et al., 2011) and journals in the domain of ecology (Paine & Fox, 2018) were more cited than the rejected articles published elsewhere, and

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

the process appears to improve the quality of manuscripts, although marginally (Calcagno, Demoinet et al., 2012; Goodman, Berlin et al., 1994; Pierie, Walvoort, & Overbeke, 1996). It is therefore surprising that a process that has little empirical support for its effectiveness, but a lot of evidence for its downsides (Smith, 2010) has so much importance.

The vast majority of studies on peer review have focused on the relationship between the socio-demographical attributes of the actors involved in the process and its outcome (Sabaj Meruane, González Vergara, & Pina-Stranger, 2016). Comparatively little research has focused on the association between the content of the manuscripts and the peer review process. This isn't surprising, given that there are few publicly available data sets of manuscripts annotated as rejected or accepted, and whenever they are made available to researchers it is usually through smaller samples designed to answer specific questions. Another factor contributing to this gap in the literature is that it is more time consuming to analyze textual data (either the referee's report or the reviewed manuscript) than papers' metadata. However, the increasing popularity of open access (Piwowar, Priem et al., 2018; Sutton & Gong, 2017) allows for a greater access to the full text of scientific manuscripts.

By scraping the content of arXiv, Kang, Ammar et al. (2018) developed a new method to identify manuscripts that were accepted at conferences after the peer review process based on submissions around the time of major NLP, machine learning (ML), and artificial intelligence (AI) conferences. These preprints were then matched with manuscripts that were published at the target venues as a way to determine whether they were accepted or "probably rejected." In addition, the manuscripts and peer-review outcomes were collected from conferences that agreed to share their data. Kang et al. (2018) were able to achieve decent accuracy at predicting the acceptance of the manuscripts in their data set. Other groups were able to obtain good performance at predicting paper acceptance with different ML models based on the text of the manuscripts (Jen, Zhang, & Chen, 2018), sentiment analysis of referee's reports (Ghosal, Verma et al., 2019), or the evaluation score given by the reviewers (Qiao, Xu, & Han, 2018).

For this study, we take advantage of the full text (title, abstract, and introduction) access to those manuscripts and explore linguistic and semantic features that correlate with the peer review outcome. We first used two readability metrics (the Flesch Reading Ease [FRE] and the New Dale-Chall Readability [NDC] Formula), as well as some indicators of scientific jargon content, and found that manuscripts that were less readable and used more jargon were more likely to get accepted. Accepted and rejected manuscripts were compared on their psycholinguistic and lexical attributes and we found that accepted manuscripts used words that were more abstract, less frequent, and acquired at a later age compared to rejected manuscripts. We then compared manuscripts on their word content and their referencing patterns through bibliographic coupling (BC), and found that the subset of accepted manuscripts were semantically closer than rejected manuscripts. Finally, we used the word content of the manuscripts to predict their acceptance, and found that specific topics were associated with greater odds of acceptance.

2. METHODS

2.1. Manuscript Data

We used the publicly available PeerRead data set (Kang et al., 2018) to analyze the semantic and lexical differences between accepted and rejected submissions to some natural language processing, artificial intelligence, and ML conferences (Table 1). We therefore used content from six platforms archived in the PeerRead data set: three arXiv subrepositories tagged by subject including submissions from 2007 to 2017 (AI: artificial intelligence, CL: computation and language, LG: machine learning), as well as submissions to three other venues: (ACL 2017: Association for

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

Table 1. Number of papers per platform

Platform	# Papers	# Accepted
ICLR 2017	427	172
ACL 2017	137	–
CoNLL 2016	22	–
arXiv:ai	4,092	418
arXiv:cl	2,638	646
arXiv:lg	5,048	1,827

Computational Linguistics; CoNLL 2016: Conference on Computational Natural Language Learning; ICLR 2017: International Conference on Learning Representations). This resulted in a data set with 12,364 submissions. Although the submissions to ACL 2017 and CoNLL 2016 had an acceptance rate in Kang et al. (2018), the information for each submission was not available in the data set at the time of the analysis.

We limited our analysis to the title, abstract, and introduction of the manuscripts, because the methods and results contained formulas, mathematical equations, and variables, which made them unsuitable for textual analysis.

2.2. Semantic Similarity

The textual data of each article, including the title, abstract, and introduction, were cleaned by making all words lowercase, and eliminating punctuation, single-character words, and common stopwords. For all analyses except for readability, scientific jargon, and psycholinguistic matching, the stem of the word was extracted using the Porter algorithm (Porter, 1980). We used the Term Frequency Inverse Document Frequency (tf-idf) algorithm to create vectorial representations based on the field of interest (title, abstract, or introduction). We then used those vectors to compute the cosine similarity between the pairs of documents.

2.3. Reference Matching and Bibliographic Coupling

To obtain the manuscripts' BC, we developed a reference matching algorithm because their format was not standardized across manuscripts. We used four conditions to group references together:

1. They were published in the same year.
2. They had the same number of authors.
3. The authors had a similarity score above 0.7 (empirically determined after manual inspection of matching results) with a fuzzy matching procedure (*Token Set Ratio* function from the FuzzyWuzzy python library, https://github.com/seatgeek/fuzzywuzzy) on the author's names.
4. The article titles had a similarity score above 0.7.

We also used two measures to determine the BC of the manuscripts. We used the number of common references to get an intersection-based BC. We also used the Jaccard Index, $\frac{R_i \cap R_j}{R_i \cup R_j}$, where R_i and R_j are the references of articles i and j.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

2.4. Psycholinguistic and Readability Variables

For word frequency estimation, we used the SUBTLEXUS corpus (Brysbaert & New, 2009) from which we used the logarithm of the estimated word frequency + 1. The word frequencies were obtained from a large corpus of television and film subtitles, and reflects the frequency count of each word in the English language. For concreteness, we used the Brysbaert, Warriner, and Kuperman (2014) data set providing a concreteness rating for 40,000 commonly known English words. The concreteness of each word included in this database was rated by every participant on a five-point scale, from abstract (1) to concrete (5). For age of acquisition, we used the Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) age of acquisition ratings for 30,000 English words. This variable provides an estimation of the average age at which each of the words in the data set was acquired by the sampled individuals.

We used the readability functions as implemented in Plavén-Sigray, Matheson et al. (2017). We used the FRE (Flesch, 1948; Kincaid, Fishburne et al., 1975) and the New Dale-Chall Readability Formula (NDC; Chall & Dale, 1995). The FRE is calculated based on the number of syllables per word and the number of words per sentence. The NDC is based on the number of words per sentence and the proportion of difficult words that are not part of a list of "common words." We also included two sources of jargon developed by Plavén-Sigray et al. (2017). The first is science-specific common words, which are words used by scientists that are not in the NDC's list of common words. The other is general science jargon, which are words frequently used in science, but aren't specific to science (see Plavén-Sigray et al. (2017) for methods). Finally, we complied a list of AI jargon from three online glossaries (https://developers.google.com/machine-learning/glossary/, http://www.wildml.com/deep-learning-glossary/, and https://en.wikipedia.org/wiki/Glossary_of_artificial_intelligence).

2.5. Logistic Regression Analysis and Keywords Importance

We used a logistic regression to predict the acceptance of the manuscripts based on their word content. The results are reported as the average precision, recall, and F1-scores, which are based on different ratios of True and False Positives and Negatives (TP, FP, and FN). We used the following definitions: $precision = \frac{TP}{TP + FP}$, $recall = \frac{TP}{TP + FN}$, $F1 = 2 \times \frac{PrecisionRecall}{Precision + Recall}$. The average is obtained with a 10-fold cross-validation, and the different scores are macro-averaged, which averages the score of each class (rejections and acceptances), and then reports the average of both. For instance, the overall recall would be computed as such: $recall_{macro} = \frac{recall_{accepted} + recall_{rejected}}{2}$. Therefore a random performance would give a score of 0.5.

We also computed a tf-idf score for each stem, and then averaged the score of each words for rejected and accepted manuscripts separately. We then computed the difference between those two scores as a measure of the importance of each stem to predict the status of the manuscripts.

2.6. Data Analysis

Because of the exploratory nature of the study and the large size of the data sets, null hypothesis significance testing has many shortcomings (Szucs & Ioannidis, 2017). In some cases, we performed statistical analysis of the results and reported the *p*-value, but those results should be interpreted carefully. Our analyses rely on the effect size, as well as the cross-validated effects on the independent subsets of the PeerRead data set (manuscripts from different venues and online repositories). All error bars represent the standard error of the mean.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

2.7. Identification of Geographic Location

We searched through the email addresses of the authors to identify research within and outside the United States. We considered a manuscript as U.S.-based if at least one author had an email address that ended with ".edu."

3. RESULTS

3.1. Readability

The readability of scientific articles has been steadily declining in the last century (Plavén-Sigray et al., 2017). One possible explanation for this is that writing more complex sentences and using more scientific jargon increase the likelihood that a manuscript will get accepted at peer review. To investigate this hypothesis, we used two measures of readability on our data: the Flesch Reading Ease (FRE) score and the New Dale-Chall Readability Formula (NDC). FRE scores decrease as a function of a ratio of the number of syllables per word and the number of words per sentence. NDC scores increase as a function of the number of words in each sentence and as the proportion of difficult words increases (words that are not present in the NDC list of common words). We also included the proportion of words from a science-specific common words and general science jargon list (constructed by Plavén-Sigray et al. (2017)). To control for potential demographic confounders, we divided our data set into two categories: manuscripts from within and outside the United States.

We found that both indicators of readability were correlated with the peer review outcome. FRE (higher score = more readable) was lower for accepted manuscripts, while NDC (higher score = less readable) was higher for accepted manuscripts (Figure 1). This was the case for both U.S. and non-U.S. manuscripts (with no sizeable differences), for every section of the manuscript (title, abstract, and introduction) and the effect was replicated within most platforms, except for the introduction with the FRE indicator.

As Plavén-Sigray et al. (2017) reported that the proportion of scientific jargon has increased over the last century, we also wondered if the peer review process would reflect this effect. Using a list of general and specific scientific jargon, we found that the manuscripts containing higher ratios of jargon were associated with higher acceptance rates. However, our list of science jargon was biased towards content from the life sciences. To confirm the relevance of these results to our data set, we generated an AI jargon list (see Methods). Using this new list, we found a robust effect across platforms and document section, where a larger proportion of AI jargon predicted greater odds of acceptance for the manuscripts.

The replication of our results independently for U.S. and non-U.S.-based manuscripts suggests that the effect is not driven by geographic locations. Statistical analyses are summarized in Tables S1 and S2.

3.2. Lexical Correlates of Peer Review Outcome

We then investigated the differences between accepted and rejected submissions based on lexical and psycholinguistic attributes. Given that we have not found consistent differences between U.S. and non-U.S.-based submissions, we pooled all manuscripts together for the rest of the analysis. We used the number of tokens (total number of words in a document) as well as two measures of lexical diversity: the number of types (unique words in a document) and the ratio between the types and token (Type-Token Ratio, TTR). We also used three psycholinguistic variables: the age of acquisition (AOA), concreteness, and frequency (on a logarithmic scale). We computed the average values of those psycholinguistic variables on all types and all tokens.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

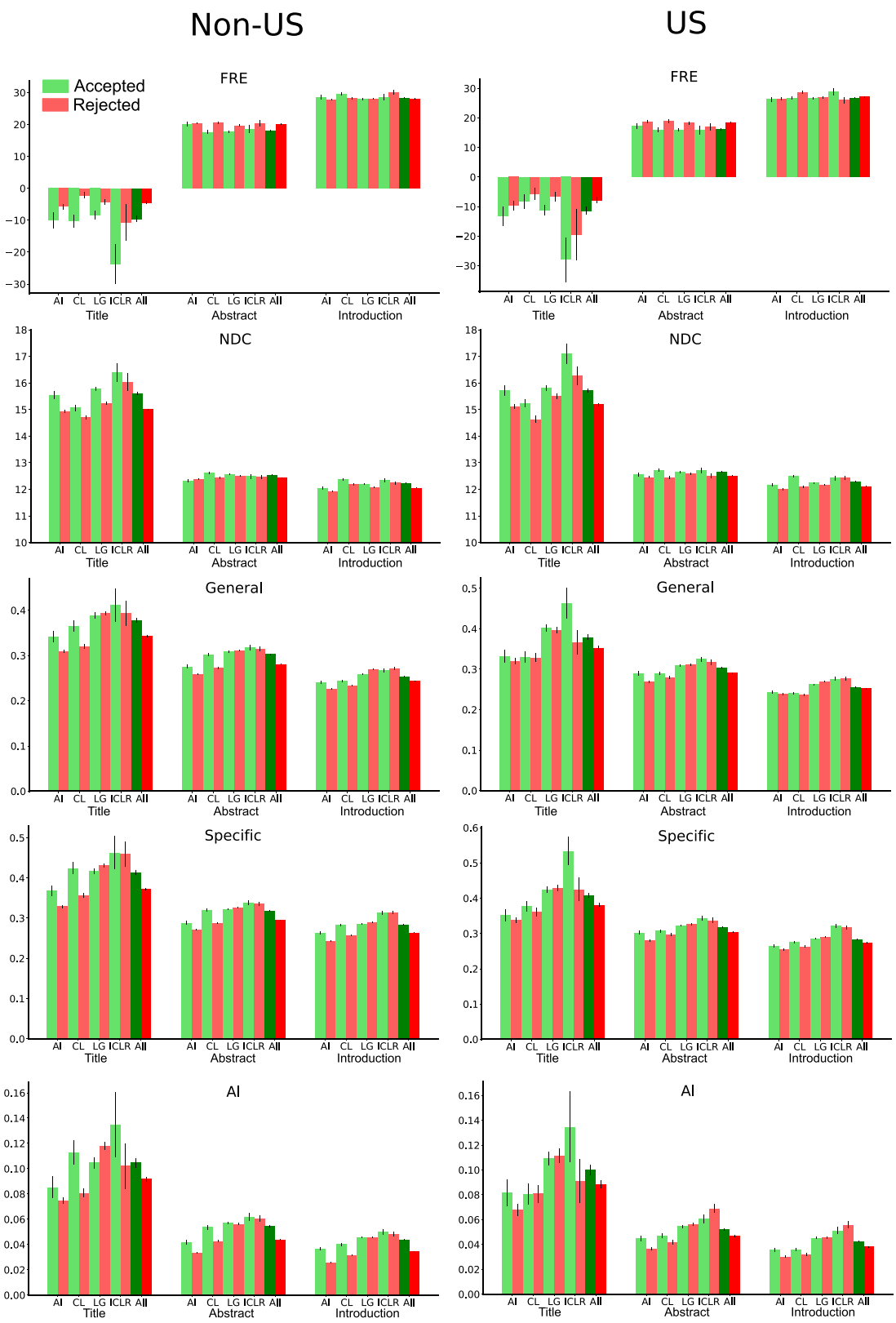

Figure 1. Average readability and jargon proportion of U.S. and non-U.S. manuscripts.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

We found consistent effects between the psycholinguistic variables and across platforms and sections, with few exceptions. Words used in accepted manuscripts were less frequent, acquired later in life and more abstract than in rejected manuscripts on average (Figure 2). The effects were consistent across all platforms except ICLR (which is smaller than the other ones).

Interestingly, we found that shorter titles, abstracts, and introductions were all associated with higher acceptance rates. Unsurprisingly, this translated into higher acceptance rates for manuscripts with lower total types (for every section). However, when taking the ratio between the two (TTR, an indicator of lexical richness), we found that this variable was positively associated with manuscript acceptance. The results from the statistical analysis are summarized in Table S3.

3.3. High-Level Semantic Correlates of Peer Review Outcome

3.3.1. Bibliographic coupling and semantic similarity

We then looked at how similar the accepted manuscripts were compared to the rejected ones based on their semantic content. First we looked at the similarity of their title, abstract, or introduction based on a tf-idf representation of their word content. Secondly, we looked at their degree of BC. We only compared pairs of manuscripts that shared at least one common reference for the next analysis (see Table 2).

As the two approaches quantify the content similarity of the documents, we wanted to verify whether those two metrics measured different aspects of the document content. It was previously reported that there is a moderate correlation between the two measures in the field of economics (Sainte-Marie, Mongeon, & Larivière, 2018). We correlated the semantic distance with the BC of the document submitted to each platform. We used a semantic distance metric based on the cosine similarity between the tf-idf representation of each document, as well as both the reference

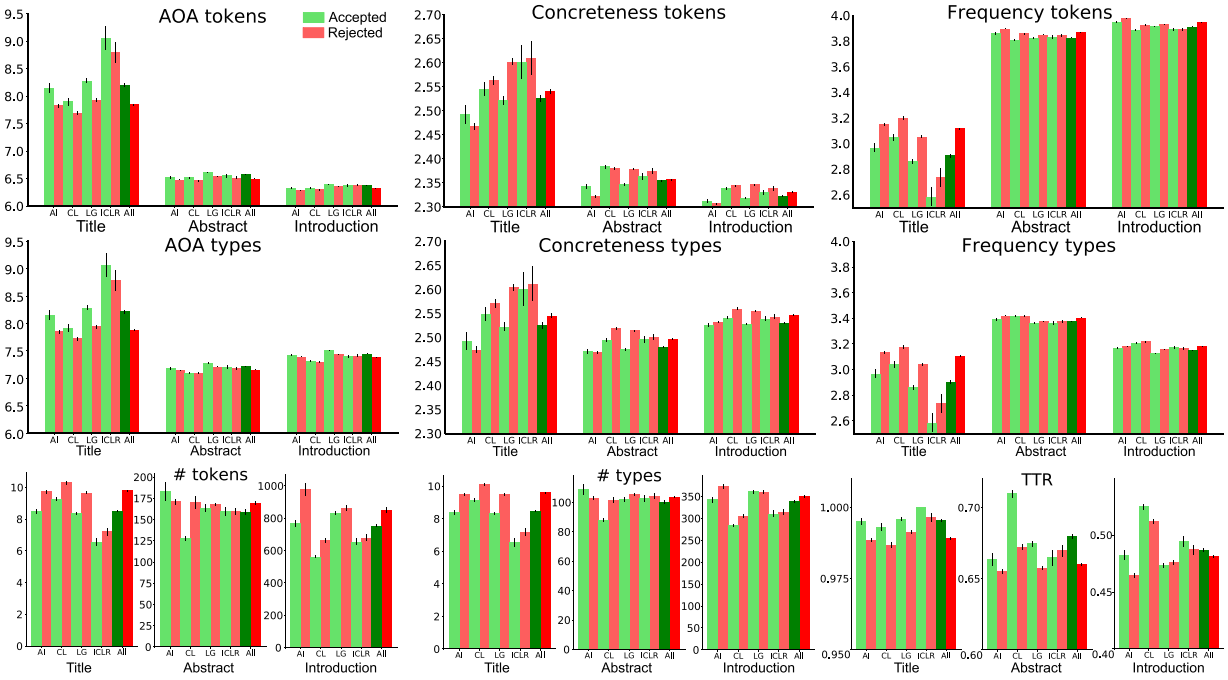

Figure 2. Psycholinguistic and lexical scores of manuscripts.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

Table 2. Reference intersection for all papers

# of common references	Frequency
1	1,619,173
2	379,161
3	124,574
4	47,254
5–9	36,753
10–19	2,527
20–29	230
30–39	68
40–49	31
50–59	11
60–80	4

intersection (# common references) and the Jaccard similarity coefficient (#references in common/# references in total) as a measure of BC. We found a moderate correlation (Pearson $r > 0.20$ and < 0.40) between both measures of BC and semantic distance when pooling all platforms together depending on what section of the manuscript was compared (Figure 3). This suggests that those two measures are not redundant features of semantic content, and that they might capture different aspects of it. This also validates our algorithm for citation disambiguation, as comparable correlations between the BC and textual similarity were reported in Sainte-Marie et al. (2018).

3.3.2. *Bibliographic coupling and peer review outcome*

We then looked at how accepted and rejected manuscripts differed based on the characteristics of their references (BC). We compared all pairs of manuscripts on the two indicators of BC (intersection and Jaccard index). Each pair of manuscripts was categorized as one of the following: "accepted": the two submissions were accepted; "rejected": the two submissions were rejected; and "mixed": one document was rejected and the other was accepted.

We found that accepted manuscripts had more references in common (Figure 4) than the two other categories of manuscripts. The effect was slightly weaker for the Jaccard similarity (intersection over union of references) and less consistent across platforms than the intersection. However, both metrics account for about 0.2% of the variance (all platforms, Jaccard: 0.228% and intersection: 0.21%).

3.3.3. *Semantic similarity and peer review outcome*

Having established that semantic similarity and BC capture different aspects of the relationship between documents, we also analyzed the semantic similarity of the documents from the four platforms. Thus, for each platform we computed the td-idf distance between all pairs of document based on their word stem.

Overall, we found that accepted manuscripts were more similar to each other than rejected manuscripts based on their abstracts and introduction (Figure 5). We found a stronger effect for

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

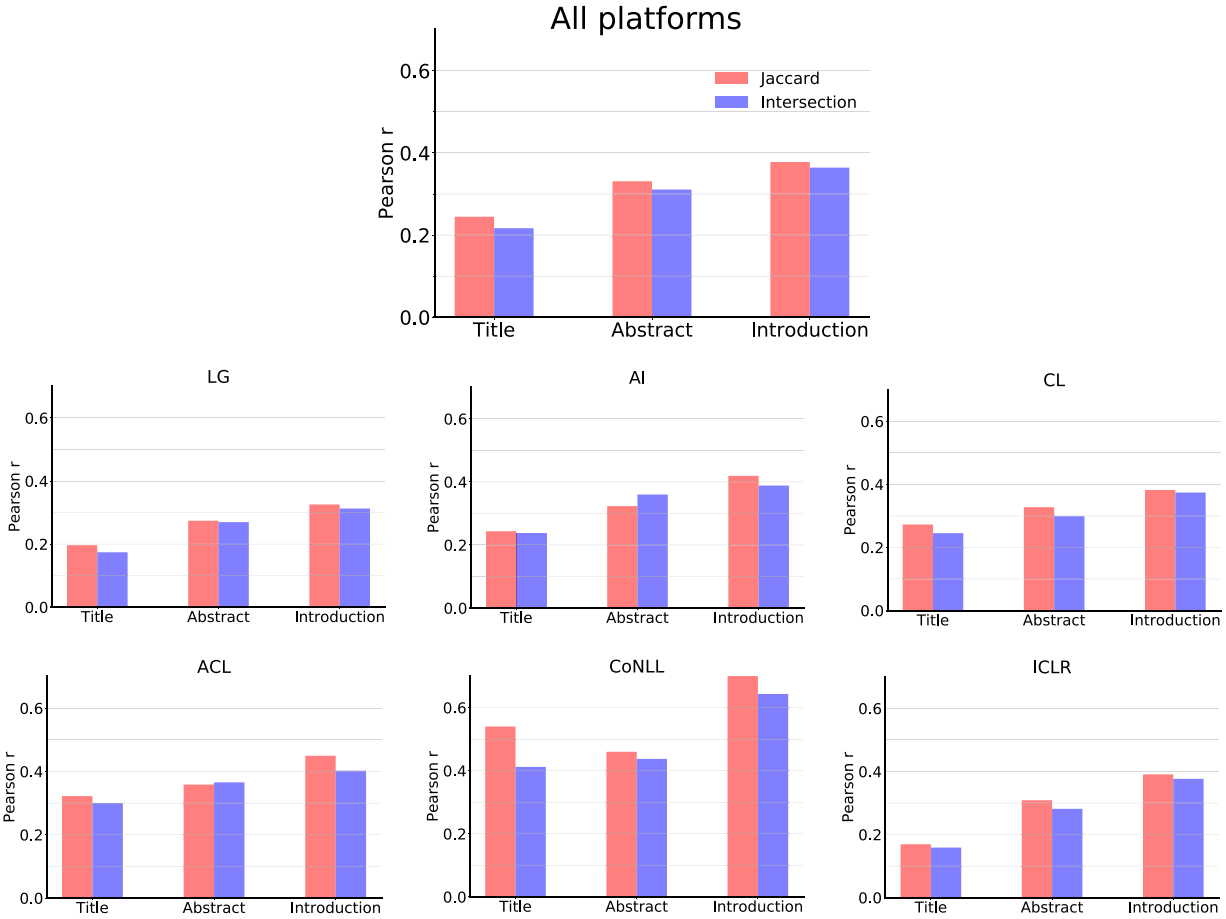

Figure 3. Correlation between semantic similarity and bibliographic coupling.

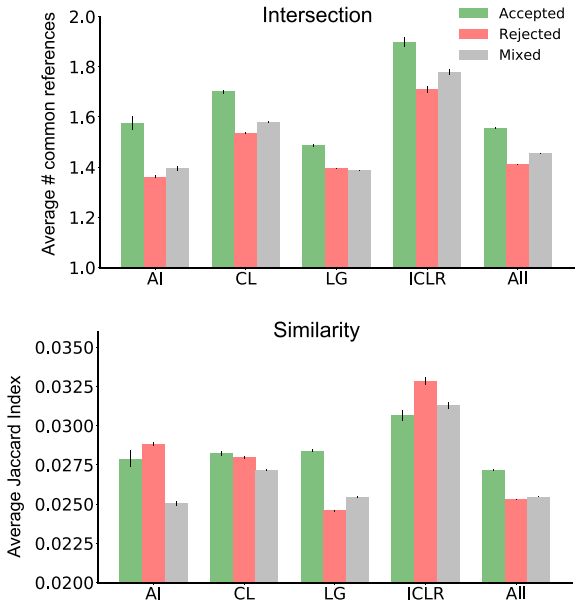

Figure 4. Bibliographic coupling bet ween accepted and rejected manuscripts.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

the similarity between the introduction (R^2 = 0.01) than for the abstract (R^2 = 0.006) when all platforms were pooled together. In other words, accepted pairs of manuscripts were more similar to each other compared to the other two pair types.

This analysis of the semantic similarity of documents (for both citations and text) showed some high-level trends based on whether or not the manuscripts were accepted after peer review. We therefore next examined the text content of the manuscripts with a more detailed approach to gain more insights into the patterns uncovered by the analysis on BC and textual similarity.

3.4. Words as a Predictor of Acceptance

Finally, after having established high-level associations between the rejected and accepted manuscripts, we attempted to predict the peer review outcome with a logistic regression using a bag-of-words approach. Overall, the model was fairly successful at predicting the peer review outcome on a 10-fold cross-validated data set (Tables S4, S5, and S6, with F-score ranging from 0.58 to 0.68 across all platforms depending of the part of text used for the prediction (performance with random assignation of outcomes is ∼0.5 for all three metrics). The model was the

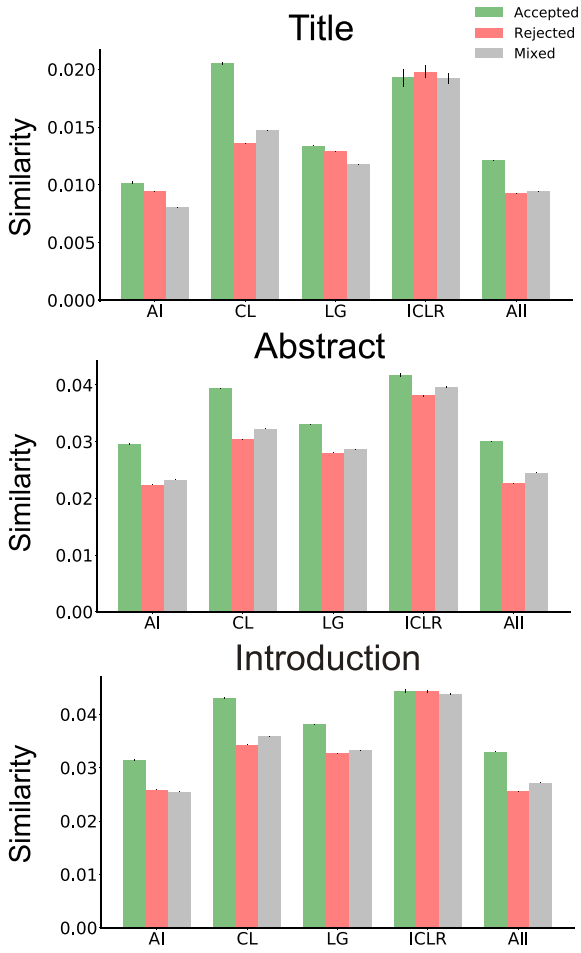

Figure 5. Semantic similarity between accepted and rejected manuscripts.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

most successful when the text of the introduction was used, followed by the text of the abstract and of the title.

After having established that we could predict to some extent the outcome of the peer review process with the word content of the manuscripts, we performed a more detailed analysis to try to get some insight into the key predictors of the outcome. We therefore computed the average tf-idf score of each stem for accepted and rejected manuscripts, and obtained measures of "importance" based on the difference between the two averages. This approach allowed us to identify the most important keywords predicting the acceptance of a manuscript (Figure 6).

Although some differences were noticeable across platforms regarding the predictors of acceptance (Tables S7, S8, and S9) and rejection (Tables S10, S11, and S12), some robust patterns emerged. Word stems related to subfields of neural networks and ML (e.g., learn, neural,

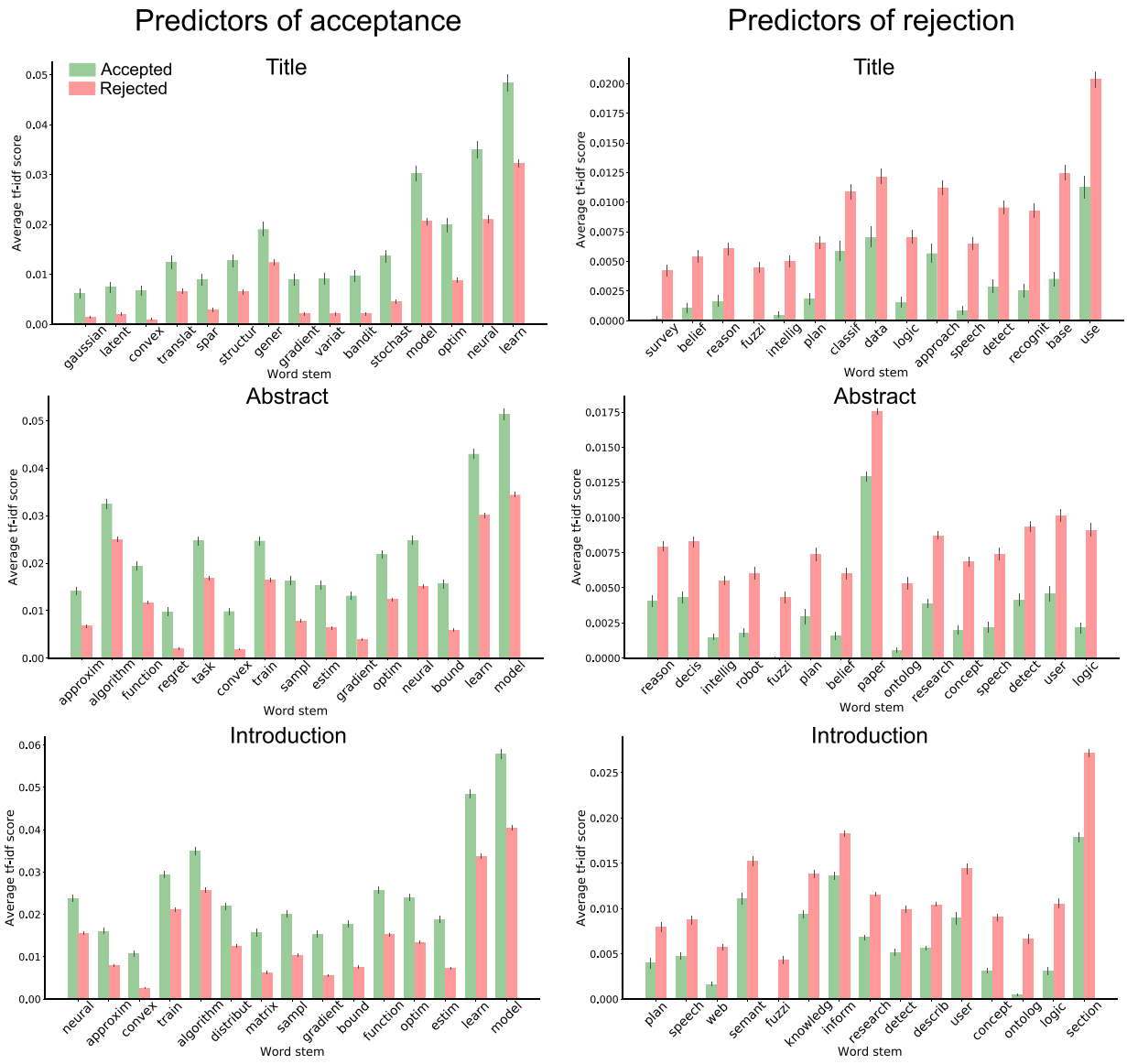

Figure 6. Most important word stems for predicting peer review outcome.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

gradient, train) were increasing the odds of the manuscript being accepted. However word stems related to the subfields of logic, symbolic processing, and knowledge representation (e.g, use, base, system, logic, fuzzi, knowledg, rule) were decreasing the odds that a manuscript would get accepted.

4. DISCUSSION AND CONCLUSION

4.1. Summary of Results

From a linguistic point of view, our results suggest that accepted manuscripts could be written in more complex, less accessible English. Using two indices of readability, one of which is agnostic to the word content of the manuscript (FRE), we found that the accepted manuscripts obtained lower readability scores. Strikingly, we found the same effect for almost all our independent data sets. The same pattern was also observed for the title, the abstract, and the introduction. Using a different type of readability indicator—the proportion of general, specific scientific, and AI jargon words—we found that manuscripts that contained a greater proportion of jargon words were more likely to be accepted. This finding may partly explain that the readability of manuscripts has steadily declined during the last century (Plavén-Sigray et al., 2017). In other words, it is possible that part of this decline could be attributed to a selection process taking place during peer review.

When considering the word content of the accepted manuscripts, we found that they had words that were acquired at a later age, that were more abstract, and that were less common than the words from the rejected manuscripts. Additionally, these manuscripts were shorter and had increased lexical diversity. Once again, the effect sizes were small given the highly multivariate determination of the peer review outcome. The effects were replicated across multiple independent data sets from different fields in AI, which strengthens the conclusions of our analysis.

From a content point of view, we compared manuscripts based on their referencing patterns and word content. We compared the coupling based on both the number of common references (intersection) and the fraction of overlap between the manuscripts' references (Jaccard similarity). We found that accepted pairs had a larger intersection than other pairs, and found a similar but less reliable effect for similarity. We used a tf-idf vectorial representation of the text from all manuscripts in the database, comparing all possible pairs of manuscripts, and we found that pairs of accepted manuscripts had considerably more overlap between their word content. This high-level analysis of the manuscript's content revealed that some topics might be associated with different odds of acceptance. We performed a correlation between the BC and the semantic similarity to get an idea of the how independent was the information provided by these two semantic indicators. As reported previously (Sainte-Marie et al., 2018), we found a weak to moderate correlation between the two, which suggests that they provide distinct sources of information about topic similarity in accepted manuscripts.

Finally, we built a logistic regression to predict the peer review outcome, which revealed that using the title, abstract, or introduction words leads to robust predictions. Our results are compatible with the presence of content bias, where trending topics in AI, such as ML and neural networks, were linked with greater acceptance rate, whereas words related to symbolic processing and knowledge-based reasoning led to lower acceptance rates.

4.2. Implications

A possible interpretation of our findings can be linked to two types of biases that have been hypothesized to take place during the peer review process: language and content bias. In language bias, authors who aren't native English speakers could receive more negative evaluations due to

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

the linguistic level of their manuscripts (Herrera, 1999; Ross, Gross et al., 2006; Tregenza, 2002). Our understanding of the extent to which such a bias could play a role in research evaluation is still limited, which is worrying given the increasingly globalized scientific system that relies on one language: English (Larivière & Warren, 2019). In terms of content bias, innovative and unorthodox methods are less likely to be judged favorably (Lee et al., 2013). This type of bias is also quite likely to play a role in fields that are dominated by a few mainstream approaches such as AI (Hao, 2019). Conservatism in this field could impede the emergence of breakthrough or novel techniques that don't fit with the current trends.

Taken together, our analysis of the linguistic aspects of manuscripts is coherent with linguistic biases during peer review. It has been reported that writers using English as their main language (L1) use words that are more abstract and less frequent than writers with English as their second language (L2) (Crossley & McNamara, 2009). Additionally, this effect is exacerbated by L2 proficiency (where larger differences are observed for beginners than advanced L2 speakers) (Crossley, Salsbury et al., 2011). The complexity of L2 writing was also shown to correlate with proficiency (Kim, 2014; Lahuerta Martínez, 2018; Radhiah & Abidin, 2018). Our results are therefore compatible with the hypothesis that L2 writers are less likely to get their manuscript accepted at peer review.

Our results are also compatible with a content bias where manuscripts on the topics of ML techniques and neural networks have greater odds to be accepted at peer review. Leading figures in the AI community have raised their voice against the overwhelming dominance of neural networks and deep learning in the domain of AI (Jordan, 2018; Knight, 2018; Marcus, 2018). The recent successes of deep learning and neural networks might explain their dominance in the field, but a bias against other techniques might impede developments similar to those that led to the breakthroughs underlying the deep learning revolution (Krizhevsky, Sutskever, & Hinton, 2012). Following this idea, several researchers have indicated that symbolic processing could hold the answer to shortcomings of deep learning (Garnelo & Shanahan, 2019; Geffner, 2018; Marcus, 2018).

However, the analyses that we performed are all correlational in nature, and given the complexity of the peer review process, other factors than the biases we highlighted could explain our results. For instance, it is possible that the features we identified that were correlated with the peer review outcome are characteristics of true scientific quality. However, a considerable number of studies have previously reported the presence of those biases in peer review (Lee et al., 2013). The data we analyzed have been historically limited and restricted and despite the limitation of this observational design, our analysis fills an important gap in the literature and is a natural step towards establishing whether or not content and linguistic biases are factors contributing to the outcome of the peer review process.

Other than biases during peer review, our results have implications for the quality of scientific communications. A recent report on reproducibility in ML found a positive correlation between the readability of a manuscript and successful replication of the claims that it makes (Raff, 2019). Selecting for less readable manuscripts during peer review may therefore increase the proportion of nonreproducible research findings.

4.3. Limitations

Our analyses are correlational and there are possible confounders that could explain our results. For instance, while our findings that some linguistic aspects of the manuscripts—the readability and psycholinguistic attributes—were correlated with the peer review outcome, we cannot infer that those variables directly lowered the odds of acceptance. It is also possible that less readable

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

manuscripts would be of higher quality due to factors unaccounted for in our study. However, this hypothesis is hard to test as we do not have a way to independently assess the true unbiased "quality" of a scientific manuscripts, as the gold standard to establish the quality of scientific manuscripts is the process under study itself.

Similarly, alternative explanations to conservatism bias are plausible. For instance, some reviewers in the field of AI might give more weight to the benchmark performance of an algorithm. This could lead to higher acceptance of manuscripts using state-of-the-art techniques. The resulting "bias" against nonmainstream approaches would be the results of the normal reviewing process, and not a bias against novelty.

Another limitation to our findings is the methodology of the peer read data set (Kang et al., 2018). For most manuscripts included in the data set, their status is inferred and the true outcome of the peer review process is unknown. Although Kang et al. (2018) validated their method on a subset of their data, the accuracy is not perfect. However, we believe that the large size of the data set is enough to counteract this source of noise. Only the minority of manuscripts included in their data set had a true peer review outcome provided by the publishing venue. This highlights the need for publishers and conferences to open their peer review process to further advance our understanding of the strengths and limitations of the peer review process.

AUTHOR CONTRIBUTIONS

Philippe Vincent-Lamarre: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing—Original draft, Writing—Review & editing. Vincent Larivière: Conceptualization, Writing—Original draft, Writing—Review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research was funded by the Canada Research Chair program and the Social Sciences and Humanities Research Council of Canada.

DATA AVAILABILITY

The custom Python scripts and the data used to generate the results of this manuscript can be found at https://github.com/lamvin/PeerReviewAI.git.

REFERENCES

Bailar, J. C. (1991). Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, *14*(1), 137–138. **DOI:** https://doi.org/10.1017/S0140525X00065705

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS ONE*, *5*(12), e14331. **DOI:** https://doi.org/10.1371/journal.pone.0014331, **PMID:** 21179459, **PMCID:** PMC3001856

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. **DOI:** https://doi.org/10.3758/BRM.41.4.977, **PMID:** 19897807

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. **DOI:** https://doi.org/10.3758/s13428-013-0403-5, **PMID:** 24142837

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & Mazancourt, C. D. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, *338*(6110), 1065–1069. **DOI:** https://doi.org/10.1126/science.1227833, **PMID:** 23065906

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline, MA: Brookline Books.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral*

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

and Brain Sciences, *14*(1), 119–135. **DOI:** https://doi.org/10.1017/S0140525X00065675

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119–135. **DOI:** https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. **DOI:** https://doi.org/10.1177/0265532210378031

De Silva, P. U. K., & Vance, C. K. (2017). Preserving the quality of scientific research: Peer review of research articles. In P. U. K. De Silva & C. K. Vance (Eds.), *Scientific scholarly communication: The changing landscape*, Fascinating Life Sciences (pp. 73–99). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-319-50627-2_6

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. **DOI:** https://doi.org/10.1037/h0057532, **PMID:** 18867058

Forscher, P. S., Brauer, M., Cox, W. T., & Devine, P. G. (2019). How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals? **DOI:** https://doi.org/10.31234/osf.io/483zj

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations. *Current Opinion in Behavioral Sciences*, *29*, 17–23. **DOI:** https://doi.org/10.1016/j.cobeha.2018.12.010

Geffner, H. (2018). Model-free, model-based, and general intelligence. *arXiv:1806.02308 [cs]*. arXiv: 1806.02308. **DOI:** https://doi.org/10.24963/ijcai.2018/2

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2019). A sentiment augmented deep architecture to predict peer review outcomes. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 414–415). **DOI:** https://doi.org/10.1109/JCDL.2019.00096

Goodman, S., Berlin, J., Fletcher, S., & Fletcher, R. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine*, *121*(1), 11–21. **DOI:** https://doi.org/10.7326/0003-4819-121-1-199407010-00003, **PMID:** 8198342

Hao, K. (2019). We analyzed 16,625 papers to figure out where AI is headed next. *MIT Technology Review*, January 25. https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/

Herrera, A. J. (1999). Language bias discredits the peer-review system. *Nature*, *397*(6719), 467. **DOI:** https://doi.org/10.1038/17194, **PMID:** 10028961

Jackson, J. L., Srinivasan, M., Rea, J., Fletcher, K. E., & Kravitz, R. L. (2011). The validity of peer review in a general medicine journal. *PLOS ONE*, *6*(7), e22475. **DOI:** https://doi.org/10.1371/journal.pone.0022475, **PMID:** 21799867, **PMCID:** PMC3143147

Jen, W., Zhang, S., & Chen, M. (2018). Predicting conference paper acceptance. page 7. http://cs229.stanford.edu/proj2018/report/117.pdf

Jordan, M. (2018). Artificial intelligence—The revolution hasn't happened yet. *Harvard Data Science Review*, *1*(1). **DOI:** https://doi.org/10.1162/99608f92.f06c6e61

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., … Schwartz, R. (2018). A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. *arXiv preprint* arXiv:1804.09635. **DOI:** https://doi.org/10.18653/v1/N18-1149

Kim, J.-Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27–51. **DOI:** https://doi.org/10.15858/engtea.69.4.201412.27

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida. **DOI:** https://doi.org/10.21236/ADA006655

Knight, W. (2018). One of the fathers of AI is worried about its future. *MIT Technology Review*, November 17. https://www.technologyreview.com/2018/11/17/66372/one-of-the-fathers-of-ai-is-worried-about-its-future/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. **DOI:** https://doi.org/10.3758/s13428-012-0210-4, **PMID:** 22581493

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, *35*, 1–11. **DOI:** https://doi.org/10.1016/j.asw.2017.11.002

Larivière, V., & Warren, J.-P. (2019). Introduction: The dissemination of national knowledge in an internationalized scientific community. *Canadian Journal of Sociology*, *44*(1), 1–8. **DOI:** https://doi.org/10.29173/cjs29548

Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. *Philosophy of Science*, *79*(5), 859–870. **DOI:** https://doi.org/10.1086/667841

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. **DOI:** https://doi.org/10.1002/asi.22784

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint* arXiv:1801.00631.

Mayden, K. D. (2012). Peer review: Publication's gold standard. *Journal of the Advanced Practitioner in Oncology*, *3*(2), 117–122. **DOI:** https://doi.org/10.6004/jadpro.2012.3.2.8, **PMID:** 25059293, **PMCID:** PMC4093306

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife*, *8*, e47338. **DOI:** https://doi.org/10.7554/eLife.47338, **PMID:** 31364991, **PMCID:** PMC6668985

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, *16*(3), e2004089. **DOI:** https://doi.org/10.1371/journal.pbio.2004089, **PMID:** 29596415, **PMCID:** PMC5892914

Mulligan, A. (2005). Is peer review in crisis? *Oral Oncology*, *41*(2), 135–141. **DOI:** https://doi.org/10.1016/j.oraloncology.2004.11.001, **PMID:** 15695114

Murray, D., Siler, K., Lariviére, V., Chan, W. M., Collings, A. M., … Sugimoto, C. R. (2018). Gender and international diversity improves equity in peer review. *bioRxiv*. **DOI:** https://doi.org/10.1101/400515

Nur Najah Radhiah, Z. A. (2018). A longitudinal corpus study of syntactic complexity development in L2 writing (Doctoral dissertation, University of Malaya).

Paine, C. E. T., & Fox, C. W. (2018). The effectiveness of journals as arbiters of scientific impact. *Ecology and Evolution*, *8*(19), 9566–9585. **DOI:** https://doi.org/10.1002/ece3.4467, **PMID:** 30386557, **PMCID:** PMC6202707

Pierie, J.-P. E., Walvoort, H. C., & Overbeke, A. J. P. (1996). Readers' evaluation of effect of peer review and editing on quality of articles in the Nederlands Tijdschrift voor Geneeskunde.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025

The Lancet, *348*(9040), 1480–1483. **DOI:** https://doi.org/10.1016/S0140-6736(96)05016-7

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., … Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. **DOI:** https://doi.org/10.7717/peerj.4375, **PMID:** 29456894, **PMCID:** PMC5815332

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (2017). The readability of scientific texts is decreasing over time. *eLife*, *6*, e27725. **DOI:** https://doi.org/10.7554/eLife.27725, **PMID:** 28873054, **PMCID:** PMC5584989

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. **DOI:** https://doi.org/10.1108/eb046814

Price, E. (2014). The NIPS experiment. *Moody Rd blog post*, December 15. http://blog.mrtz.org/2014/12/15/the-nips-experiment.html

Qiao, F., Xu, L., & Han, X. (2018). Modularized and attention-based recurrent convolutional neural network for automatic academic paper aspect scoring. In X. Meng, R. Li, K. Wang, B. Niu, X. Wang, & G. Zhao (Eds.), *Web Information Systems and Applications* (pp. 68–76). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-030-02934-0_7

Radhiah, N. N., & Abidin, Z. (2018). A longitudinal corpus study of syntactic complexity development in L2 writing.

Raff, E. (2019). A step toward quantifying independently reproducible machine learning research. *arXiv:1909.06674 [cs, stat]*. arXiv: 1909.06674.

Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., … Krumholz, H. M. (2006). Effect of blinded peer review on abstract acceptance. *JAMA*, *295*(14), 1675–1680. **DOI:** https://doi.org/10.1001/jama.295.14.1675, **PMID:** 16609089

Sabaj Meruane, O., González Vergara, C., & Pina-Stranger, Á. (2016). What we still don't know about peer review. *Journal of Scholarly Publishing*, *47*(2), 180–212. **DOI:** https://doi.org/10.3138/jsp.47.2.180

Sainte-Marie, M., Mongeon, P., & Larivière, V. (2018). Do you cite what I mean? Assessing the semantic scope of bibliographic coupling in economics. *23rd International Conference on Science and Technology Indicators* (pp. 649–657). Centre for Science and Technology Studies.

Smith, R. (2010). Classical peer review: An empty gun. *Breast Cancer Research*, *12*(4), S13. **DOI:** https://doi.org/10.1186/bcr2742, **PMID:** 21172075, **PMCID:** PMC3005733

Sutton, C., & Gong, L. (2017). Popularity of arXiv.org within computer science. *arXiv:1710.05225 [cs]*. arXiv: 1710.05225.

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*, 390. **DOI:** https://doi.org/10.3389/fnhum.2017.00390, **PMID:** 28824397, **PMCID:** PMC5540883

Tregellas, J. R., Smucny, J., Rojas, D. C., & Legget, K. T. (2018). Predicting academic career outcomes by predoctoral publication record. *PeerJ*, *6*, e5707. **DOI:** https://doi.org/10.7717/peerj.5707, **PMID:** 30310749, **PMCID:** PMC6174868

Tregenza, T. (2002). Gender bias in the refereeing process? *Trends in Ecology & Evolution*, *17*(8), 349–350. **DOI:** https://doi.org/10.1016/S0169-5347(02)02545-4

Wakeling, S., Willett, P., Creaser, C., Fry, J., Pinfield, S., … Medina Perea, I. (2019). 'No comment'? A study of commenting on PLOS articles. *Journal of Information Science*, *46*(1), 82–100. **DOI:** https://doi.org/10.1177/0165551518819965

Ziman, J. (2002). *Real science: What it is and what it means*. Cambridge: Cambridge University Press.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/2/662/1930783/qss_a_00125.pdf by guest on 18 January 2025