A Study on the Readability of Scientific Publications

Thanasis Vergoulis[1(✉)], Ilias Kanellos[1,2], Anargiros Tzerefos[3], Serafeim Chatzopoulos[1,3], Theodore Dalamagas[1], and Spiros Skiadopoulos[3]

[1] IMSI, Athena Research and Innovation Center, 15125 Athens, Greece
{vergoulis,ilias.kanellos,schatz,dalamag}@athenarc.gr
[2] School of Electrical and Computer Engineering, NTUA, 15780 Athens, Greece
[3] Department of Informatics and Telecommunications, University of the Peloponnese, 22100 Tripoli, Greece
{dit13139,spiros}@uop.gr

Abstract. Several works have used traditional readability measures to investigate the readability of scientific texts and its association with scientific impact. However, these works are limited in terms of dataset size, range of domains, and examined readability and impact measures. Our study addresses these limitations, investigating the readability of paper abstracts on a very large multidisciplinary corpus, the association of expert judgments on abstract readability with traditional readability measures, and the association of abstract readability with the scientific impact of the corresponding publication.

Keywords: Readability · Scientific impact · Text analysis

1 Introduction

Reporting scientific issues with clarity in publications is a fundamental part of the scientific process, since it aids the comprehension of research findings and establishes the foundation for future research work. In addition, well-written scientific texts help the comprehension of research findings and scientific knowledge by journalists, educators, science enthusiasts, and the public, in general, preventing the dissemination of inaccuracies and misconceptions.

For these reasons, measuring and studying the readability in academic writing is of great importance and many studies have been conducted to investigate relevant issues. Most of the studies rely on traditional readability measures originally introduced to help in selecting appropriate teaching materials [21], or quantify the minimum required educational level for a text to be understood.

In this work, we focus our investigation on the readability of scientific paper abstracts. In particular, we focus on the following research questions:

– RQ1: How does the readability of publication abstracts, as calculated by traditional readability measures, evolve over time?

A. Doucet et al. (Eds.): TPDL 2019, LNCS 11799, pp. 136–144, 2019.
https://doi.org/10.1007/978-3-030-30760-8_12

- RQ2: To what extent are these measures associated with what is considered by domain experts as a well-written scientific text?
- RQ3: To what extent is the readability of a publication abstract associated with the scientific impact of the corresponding publication?

Existing literature investigates some of the previous research questions to a limited extent only, e.g., most works only focus on particular scientific domains, use small datasets, or examine only few readability and impact measures (details in Sect. 2). Our contributions are the following:

- We investigate readability over time on a multidisciplinary corpus an order of magnitude larger than those used by previous studies (∼12M abstracts).
- To the best of our knowledge, this is the first work to examine the agreement of readability as it is perceived by domain experts, compared to that calculated by traditional readability measures. Additionally, we make our dataset of the expert judgments publicly available at Zenodo (see Sect. 3.1)
- We examine the association of readability, as measured both by traditional measures and expert judgements, to impact. We employ three different impact measures, capturing slightly different notions of scientific impact.

2 Related Work

Several studies investigated the readability of scientific texts (abstracts and/or full texts) over time and its association to paper impact. However, most studies investigate small datasets, restricted to a particular domain (e.g., management and marketing [1,4,16,18], psychology [9,10], chemistry [3], information science [11]). Only few studies investigated multiple disciplines [6,15].

Longitudinal studies examining readability of scientific texts report varying results. In [6] FRE was measured for 260,000 paper abstracts revealing no significant changes in readability over time. In [20] the 100 most highly cited neuroimaging papers were examined in terms of readability, using an average of five grade level readability formulas, showing no relationship between readability and the papers' publication years. In [11] FRE and SMOG were used on papers of the Information Science Journals, published in the span of a decade, reporting only a trivial decrease of abstract readability and a respective increase in full text readability. Another recent research, however, examined more than 700,000 abstracts from PubMed using the FRE and Dale-Chall measures, reporting a statistically significant decrease in readability over time [15]. The association of paper impact and readability has also been examined, with most studies reporting no significant association between readability and citation counts [6,11,20]. However, in [10], although no correlation between citation counts and FRE was found, the authors additionally consider existing curated selections of prestigious publications finding, in this case, that readability and impact did correlate.

Our work extends previous studies threefold: first, we use four measures to examine abstract readability over time on a larger corpus and time span, compared to previous work. Second, we investigate the association of readability

measures to expert readability judgements on scientific abstracts. Finally, we study the association of readability and impact using three impact measures capturing different impact aspects.

3 Methods

3.1 Datasets

Publication Abstracts and Impact (D1). To study the readability of scientific publications over time (RQ1) and its correlation to scientific impact (RQ3), we used a large multidisciplinary collection of scientific texts. We gathered all publications (distinct DOIs) included in the OpenCitations COCI dataset[1]. We collected their abstracts and titles from the Open Academic Graph[2] [17,19] and the Crossref REST API[3], keeping only publications for which the abstract was available. Then, we performed basic cleaning by removing publications containing XML tags in the abstract and ignoring publications with abstracts containing less than three sentences[4]. This resulted in a dataset containing abstracts and citations for 12,534,077 publications. Finally, we used this dataset to calculate citation counts and additionally gathered extra impact scores (i.e., PageRank and RAM) about all the collected publications using BIP! Finder's API[5].

Domain Expert Readability (D2). To investigate RQ2 and RQ3, we gathered judgments for the readability of publication abstracts from 10 data and knowledge management experts (PhD students or post-docs) through a Web-based survey. The abstracts were a subset of AMiner's DBLP citation dataset[6]. To guarantee that most of the abstracts would be relevant to the area of expertise of our experts, we only used abstracts containing the terms illustrated in Table 1. Each expert provided judgments for a small subset of these abstracts (34−202). Upon reviewing a particular publication, an expert had to read its abstract and then, answer three questions relevant to different aspects of abstract readability. These questions were worded as shown in Table 2 and the allowed answers were based on a 5 point scale[7]. Each time an expert requested to review a new abstract, the system provided either an abstract already rated by other experts, or one unrated. To guarantee a substantive overlap between the sets of abstracts rated by each expert, we used the following procedure: an unrated abstract was provided to the expert only after rating 10 abstracts previously rated by others. Dataset D2 is openly available at Zenodo[8].

[1] http://opencitations.net/download (November 2018 Dump).
[2] https://www.openacademic.ai/oag/.
[3] https://www.crossref.org/services/metadata-delivery/rest-api/.
[4] This is a restriction imposed by the `textstat` library (see Sect. 3.2).
[5] http://bip.imsi.athenarc.gr:4000/documentation.
[6] https://aminer.org/citation.
[7] For each question, the interpretation of the extreme scale values (i.e., 1 and 5) were provided (actual wording is described in the dataset description page in Zenodo).
[8] https://doi.org/10.5281/zenodo.2651009.

Table 1. List of terms used to construct D2

"database"	"machine learning"	"information retrieval"	"data management"
"cloud computing"	"data mining"	"algorithms"	"classification"
"query processing"	"networks"	"indexing"	"distributed systems"

Table 2. The questions of the Web-based survey

Q1	"Please rate how well-written the abstract is"
Q2	"Does the abstract contain linguistic errors?"
Q3	"Please rate how clear the contribution of the paper is (based on the abstract)"

3.2 Examined Readability and Impact Measures

In our experiments we examine abstract readability based on four measures: FRE [5], SMOG, [13], Dale-Chall (DC) [21], and Gunning Fog (GF) [8]. The former two use statistics such as sentence length and average number of syllables per word, while the latter two also take into account "difficult" words (e.g., based on syllable length, or dictionaries). For FRE a higher score indicates a more readable text, while the opposite holds for the other measures. All readability scores were calculated using the `textstat`[9] (release 0.5.6) Python library.

Additionally we calculate three scientific impact measures: citation counts, PageRank [14], and RAM [7]. Citation counts are the de facto measure used in evaluations of academic performance. PageRank differentiates citations, based on the paper making them, following the principle that "good papers cite other good papers". Finally, RAM considers recent citations as more important, aiming to overcome the citation bias against recently published papers.

4 Results and Discussion

4.1 Longitudinal Study of Readability

In this section we focus on research question RQ1. To examine temporal changes in readability, we calculated the FRE, SMOG, GF, and DC scores on dataset D1 and measured the yearly average scores (Fig. 1). We observe that, generally, abstract readability seems to be decreasing over time, based on all measures[10]. These findings are in agreement with the results of [11] which showed an insignificant downtrend in FRE on Information Science Journals, however they do not demonstrate as dramatic a drop in readability, as shown in [15] for PubMed papers. On the other hand, our findings contrast previous domain specific works that report relatively constant readability with time [6]. The trend of decreasing readability could be attributed, as previous works have stated, on factors such as the increased use of scientific jargon [15].

[9] https://github.com/shivam5992/textstat.

[10] Recall that FRE scores increase with readability, contrary to the other measures.

4.2 Readability Measures Vs Expert Judgments

Since traditional readability measures were initially introduced for testing the readability level of school textbooks [21] their suitability for use in the context of scientific articles (as conducted in previous studies) could be debatable. In this section, we investigate this matter using dataset `D2`. For each abstract in `D2` we calculated (a) its score based on each of the four readability measures used in our study and (b) the average score it gathered for each question posed to the experts. In our experiments, to avoid biases, we kept only abstracts judged by at least four experts resulting in a set of 172 publication abstracts.

Table 3 illustrates the correlation (Spearman's ρ and Kendall's τ) of the four readability measures to the average score for each question. Interestingly, only extremely weak correlations were found. Although the dataset is relatively small, following the reasoning in [2], if a true significantly stronger correlation (e.g., $\tau > 0.3$) existed, we would expect to have measured greater values of

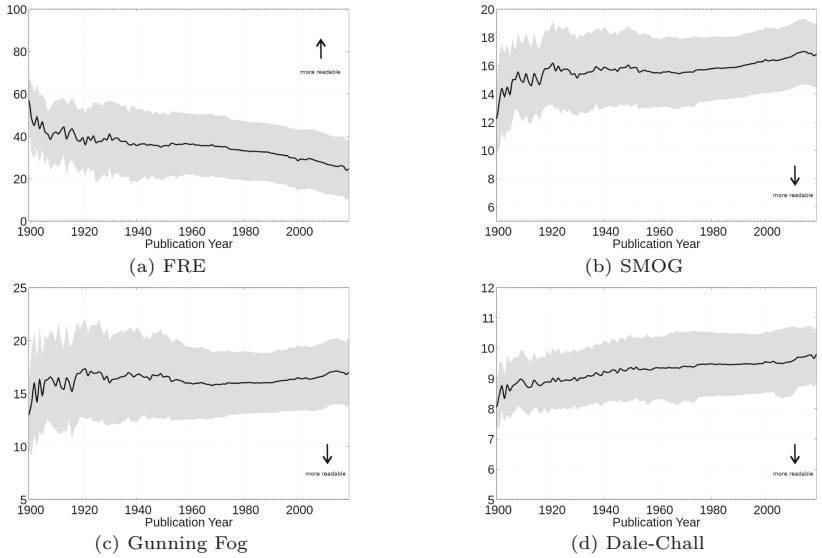

Fig. 1. Average scores per year (with st. deviation)

Table 3. Correlations of expert judgments to readability measures. FRE scores were reversed for reasons of uniformity, i.e. readability decreases with score, for all measures.

	Spearman's ρ				Kendall's τ			
	FRE	SMOG	DC	GF	FRE	SMOG	DC	GF
Q1	−0.0776	0.0371	−0.0372	−0.0552	−0.0509	0.0256	−0.0275	−0.0395
Q2	−0.0346	−0.0100	0.0946	0.0794	−0.0247	−0.0135	0.0657	0.0542
Q3	−0.0884	0.0216	−0.1033	−0.0712	−0.0584	0.0114	−0.0741	−0.0494

Table 4. Pairwise correlations (τ) of expert judgments on question Q1. *Corr. coefficients significant at $p < 10^{-3}$. **Corr. coefficients significant at $p < 10^{-5}$.

Q1	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
E1	1.0									
E2	0.27	1.0								
E3	0.05	0.51**	1.0							
E4	0.20	0.07	0.18	1.0						
E5	0.50*	0.23	0.26	0.38	1.0					
E6	0.31	0.37*	0.42*	0.09	0.35	1.0				
E7	0.20	0.27	0.46*	0.23	0.35	0.48**	1.0			
E8	0.22	0.28	0.21	0.22	0.38	0.31*	0.26	1.0		
E9	0.17	0.32*	0.51**	0.24	0.17	0.34*	0.35*	0.43**	1.0	
E10	0.28	0.60*	0.68**	0.45	0.47	0.50	0.68*	0.33	0.40	1.0

correlation. This result may hint that mechanically applying classic readability measures in the context of scientific texts, a common practice in the literature, may not be entirely appropriate. While this is not to say that readability measures are entirely useless, it does point out the need for additional methods particularly tailored to measure readability in this context.

Another interesting subject for investigation is whether the notion of being "readable" is compliant between different experts and between the different questions of Table 2. Table 4 shows the correlation[11] between the average scores given by the experts to question Q1 for the abstracts in `D2`[12]. We observe that the answers of reviewers agree substantially only in few cases (e.g., $\tau = 0.68$ for researchers E3-E10) and overall expert responses do not seem to correlate at all (similar results were found for Q2 and Q3). These results indicate that each individual's idea of what defines a "well written" text may differ. The above may be to some degree reflected in the correlation of averages given to questions Q1-Q3. We found less than perfect correlation of these results to each other ($0.48 < \rho < 0.77$, and $0.34 < \tau < 0.59$ between averages for all pairs of Q1-Q3) which additionally hints that these questions indeed capture different semantics.

4.3 Abstract Readability vs Paper Impact

In this section we focus on research question RQ3, examining the association of publication readability and impact on dataset `D1`. First, we measure Spearman's ρ[13] between readability rankings (FRE, SMOG, GF, DC) and impact rankings (Citation Counts, PageRank, and RAM). Overall we report very weak

[11] Due to lack of space we omit ρ values, however the results were similar.
[12] For this measurement, we used all overlapping `D2` abstracts for each expert pair.
[13] We omit τ since it runs very slow on this dataset ($\sim$ 12M papers).

correlations between readability and impact measures (Table 5). This is in agreement with previous research which focused on particular domains [6,11,20]. An interesting observation is that, among the other impact measures, RAM achieves a significantly higher (but not moderate) correlation to the readability measures in comparison to Citation Counts and PageRank. This finding could be explained as follows: due to its de-bias mechanism, a large proportion of the top-ranked publications based on RAM are recently published articles. In addition, based on Figs. 1a–d, recent publications tend to have less readable abstracts. Therefore, since both RAM and readability scores favor recent publications, it is not surprising that we observe a higher correlation in this case.

Table 5. Correlations (ρ) of readability measures to impact measures (FRE scores reversed for uniformity, star notation same as in Table 4).

	FRE	SMOG	DC	GF
Citation count	−0.0525**	0.0656**	−0.0013**	0.03800**
PageRank	0.0001	0.0076**	−0.01635**	0.0011*
RAM	0.1169**	0.1257**	0.0397**	0.0837**

Since we generally found disagreements between traditional readability measures and expert judgments (Sect. 4.2), we also measure readability based on the averages of expert responses compared to impact measures. We note similar relative values for Spearman's ρ and Kendall's τ, that correspond to very weak and statistically insignificant correlations (Table 6). One conclusion based on the above is that readability does not seem to play a key role in whether a paper will be cited. Our results show that this holds regardless of whether we consider readability measures, or expert judgments. Along with discussion in [6] this counters claims that simple abstracts correlate with citation counts [12].

Table 6. Correlations of expert judgments to impact measures.

	Spearman's ρ			Kendall's τ		
	CC	PR	RAM	CC	PR	RAM
Q1	0.1925	0.1896	0.2242	0.1358	0.1286	0.1539
Q2	0.1827	0.1433	0.1963	0.1273	0.0946	0.1366
Q3	0.162	0.1285	0.2192	0.1139	0.0878	0.1526

5 Conclusion

In this work we investigated several issues regarding the readability of publications. First, we conducted a longitudinal study using ∼12M publication abstracts

from many scientific disciplines. To the best of our knowledge, this is the largest collection of scientific texts analyzed in terms of readability so far. Our findings support the results of some earlier studies (e.g., [11,15]), that the overall readability of scientific publications tends to decrease. Second, we examine if the experts' opinion about the readability of scientific texts is compliant with the notion of readability captured by traditional measures. Our findings suggest that these measures are not in absolute agreement. This indicates that there is a need for new, specialized readability measures tailored for scientific texts. Finally, we examined how readability of publications (both as perceived by domain experts and as captured by traditional measures) associates with different aspects of scientific impact. Our results have shown no significant correlation of readability and impact.

Acknowledgments. We acknowledge support of this work by the project "Moving from Big Data Management to Data Science" (MIS 5002437/3) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. Bauerly, R.J., Johnson, D.T., Singh, M.: Readability and writing well. Mark. Manag. J. **16**(1), 216–227 (2006)
2. Bonett, D.G., Wright, T.A.: Sample size requirements for estimating pearson, kendall and spearman correlations. Psychometrika **65**(1), 23–28 (2000)
3. Bottle, R.T., Rennie, J.S., Russ, S., Sardar, Z.: Changes in the communication of chemical information I: some effects of growth. J. Inf. Sci. **6**(4), 103–108 (1983)
4. Crosier, K.: How effectively do marketing journals transfer useful learning from scholars to practitioners? Mark. Intell. Plan. **22**(5), 540–556 (2004)
5. Flesch, R.: A new readability yardstick. J. Appl. Psychol. **32**(3), 221 (1948)
6. Gazni, A.: Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. J. Inf. Sci. **37**(3), 273–281 (2011)
7. Ghosh, R., Kuo, T.T., Hsu, C.N., Lin, S.D., Lerman, K.: Time-aware ranking in dynamic citation networks. In: 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 373–380. IEEE (2011)
8. Gunning, R.: The Technique of Clear Writing. McGraw-Hill, New York (1952)
9. Hartley, J., Pennebaker, J.W., Fox, C.: Abstracts, introductions and discussions: how far do they differ in style? Scientometrics **57**(3), 389–398 (2003)
10. Hartley, J., Sotto, E., Pennebaker, J.: Style and substance in psychology: are influential articles more readable than less influential ones? Soc. Stud. Sci. **32**(2), 321–334 (2002)
11. Lei, L., Yan, S.: Readability and citations in information science: evidence from abstracts and articles of four journals (2003–2012). Scientometrics **108**(3), 1155–1169 (2016)
12. Letchford, A., Preis, T., Moat, H.S.: The advantage of simple paper abstracts. J. Informetr. **10**(1), 1–8 (2016)

13. Mc Laughlin, G.H.: Smog grading-a new readability formula. J. Read. **12**(8), 639–646 (1969)
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab (1999)
15. Plavén-Sigray, P., Matheson, G.J., Schiffler, B.C., Thompson, W.H.: The readability of scientific texts is decreasing over time. Elife **6**, e27725 (2017)
16. Sawyer, A.G., Laran, J., Xu, J.: The readability of marketing journals: are award-winning articles better written? J. Mark. **72**(1), 108–117 (2008)
17. Sinha, A., et al.: An overview of Microsoft academic service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web, pp. 243–246. ACM (2015)
18. Stremersch, S., Verniers, I., Verhoef, P.C.: The quest for citations: drivers of article impact. J. Mark. **71**(3), 171–193 (2007)
19. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD, pp. 990–998. ACM (2008)
20. Yeung, A.W.K., Goto, T.K., Leung, W.K.: Readability of the 100 most-cited neuroimaging papers assessed by common readability formulae. Front. Hum. Neurosci. **12**, 308 (2018)
21. Zamanian, M., Heydari, P.: Readability of texts: state of the art. Theory Pract. Lang. Stud. **2**(1), 43–53 (2012)