Journal of High School Science

A multi-dimensional analysis of linguistic differences in scientific communication across audiences and disciplines

Ranjan R*, Sinha V*

Submitted: July 3, 2025, Revised: version 1, September 3, 2025, version 2, September 20, 2025
Accepted: September 30, 2025

Abstract

Quantifying differences across disciplines and audiences is crucial to expanding the accessibility of scholarly material. In this study, we investigated the linguistic features of lay and technical research paper summaries from the biological, physical, and social sciences based on word-level co-occurrence graph networks and the 14 conventional indices of syntactic complexity. Our findings showed that while technical summaries exhibited similar characteristics across disciplines, lay summaries displayed notable variation: physical science summaries favored compact, nominal-heavy structures, whereas social science summaries relied more on clausal elaboration and subordination. To assess graph-theoretic textual features, we used Pointwise Mutual Information (PMI) to construct Word Adjacency Graphs for both technical and lay summaries, and computed six normalized graph-theoretic indices: lexical diversity (connected nodes per word), semantic connectivity (average nodal degree), cohesion (density), modularity (average clustering), and conceptual integration (largest connected component size / nodes). As observed with the conventional measures of syntactic complexity, while technical summaries were generally homogenous across disciplines, we found that the lay summaries differed significantly regarding graph-theoretic textual metrics. These findings highlight that not only did scientific domains differ in their linguistic structures, but these differences were audience-dependent. Our approach offers a qualitative framework for evaluating semantic complexity in science writing and has implications for both automated readability assessments and cross-disciplinary science education.

Keywords

Disciplinary conventions, Text readability, Syntactic complexity, Audience analysis, Semantic complexity, Graph-Theoretical Textual analysis, Scientific communication, Technical summaries, Lay summaries, Scholarly material accessibility

Ritvik Ranjan, Wheeler High School, 375 Holt Rd NE, Marietta, GA 30068, USA. ritvikranjan@gmail.com
Vidya Sinha, Wheeler High School, 375 Holt Rd NE, Marietta, GA 30068, USA. vidya.sinha@wheelermagnet.com
*These authors contributed equally to this work.

https://doi.org/10.64336/001c.145850

Introduction

Efforts to expand public access to scientific knowledge have increased in recent years due to open-access publishing policies (1, 2), science journalism (3, 4), and journals explicitly including lay summaries (5, 6). This expansion helps democratize academic developments and foster public interest in science. The Proceedings of the National Academy of Sciences (PNAS) is a notable journal that participates in this trend; since August 2013, PNAS has required authors to submit a "Significance Statement" alongside their abstract. While the abstract targets specialists, the significance statement is for non-experts, such as policymakers, scientists in other fields, and educated lay readers. This dual-summary format provides a controlled setting for comparative linguistic analysis, providing more robust findings rather than comparing unrelated texts.

Linguistic complexity directly affects a reader's ability to understand scientific texts. Syntactic complexity, in particular, is linked to increased cognitive processing load and comprehension difficulty (7, 8). Traditional readability studies often use simple measures like sentence length and word frequency, showing that many scientific texts are too complex for the general public (9). However, these measures overlook deeper grammatical and semantic nuances, which has led to the development of more precise syntactic complexity metrics, such as those based on T-units (10).

Despite the advances, gaps remain in analyzing linguistic complexity in scientific communication. First, while syntactic complexity has been studied in second-language writing (11, 12), its examination in a controlled setting that directly compares lay and technical summaries of the same texts is limited. Second, although graph-theoretic approaches to linguistic complexity have been applied in various areas (13, 14), a systematic application of this method to quantify differences across both audiences and scientific disciplines simultaneously has not been performed. Third, there is limited understanding of how complexity varies across disciplines. Disciplinary conventions influence lexical and syntactic structures (15, 16), but comparative linguistic analyses across scientific fields, especially those considering multiple audience types, are scarce, with most studies focusing on a single genre or audience (17). These unaddressed areas highlight a need for research that investigates the interplay of audience, discipline, and multi-dimensional linguistic complexity.

This study addresses these gaps. It analyzes 300 research articles published in PNAS after 2013, extracting their significance statements (lay summaries) and abstracts (technical summaries). The study uses two analytical frameworks: 14 syntactic metrics from Lu (10)'s L2 Syntactic Complexity Analyzer and 6 graph-theoretic metrics that capture word co-occurrence and association structures. This dual-method approach aims to answer the following research questions. 1] How do syntactic and structural complexities differ between lay and technical summaries of scientific articles? 2] To what extent do these differences vary across scientific disciplines within both lay and technical summaries?

By combining traditional syntactic analysis with network-based modeling, this study introduces a new multi-dimensional framework for evaluating linguistic complexity in scientific communication. This framework contributes to theoretical models of language complexity and practical discussions about accessible science communication.

Literature review

Linguistic complexity encompasses lexical, syntactic, and discourse-level features. In applied linguistics and natural language processing (NLP), syntactic complexity indicates language development, text difficulty, and genre variation (12, 18). Recently, researchers have also proposed structural representations of language using graph-theoretic methods, motivated by the limits of traditional, linear metrics in capturing semantic organization and cohesion (14, 19). This section reviews work in syntactic complexity, graph-theoretic text modeling, and their applications in science communication and disciplinary writing.

Syntactic complexity

Syntactic complexity refers to the sophistication of grammatical structures in a text (20). Unlike readability, which typically focuses on surface-level features like sentence length or syllable count (21), syntactic complexity examines the underlying grammatical structure. Early research in second language acquisition (SLA) identified syntactic complexity as a core dimension distinguishing writing proficiency (12, 22). Measures like the number of dependent clauses per T-unit or the ratio of complex nominals to total noun phrases were shown to increase with proficiency and task complexity.

To measure these concepts, Lu (10) developed the L2 Syntactic Complexity Analyzer (L2SCA), a tool that computes 14 indices based on established categories. These indices include sentence-level metrics (Mean Length of Sentence (MLS), Clauses per Sentence (C/S)); T-unit metrics (Mean Length of T-unit (MLT), Dependent Clauses per T-unit (DC/T)); and clause-level metrics (Mean Length of Clause (MLC), Coordinate Phrases per Clause (CP/C)). Lu's tool was validated on large corpora of native and non-native writing and has since been widely adopted in SLA, writing assessment, and corpus-based text analysis (11, 23, 24, 25).

Syntactic complexity measures are sensitive to genre and audience. Biber (26) showed that academic prose differs from spoken or narrative language in its use of subordination, nominalization, and passive constructions. Crossley and McNamara (27) found a correlation between increased syntactic sophistication in writing and higher essay quality ratings, especially in academic contexts. Additional studies on genre variation found that written genres like scientific articles show compressed, noun-heavy syntax, while spoken and informal genres use more coordination and clause chaining (28, 29). However, most of this work has focused on either academic writing or student essays. Few studies have applied these syntactic metrics to scientific communication tailored for non-expert audiences. Consequently, there is a need for more precise syntactic models, like those tested in Lu's (10)

L2SCA, to be used in science communication research.

Disciplinary variation in scientific writing

Scientific writing varies across disciplines. Research in applied linguistics consistently shows that different scientific fields adopt distinct rhetorical structures, terminology, and syntactic conventions (30, 31). For example, Hyland (32) found that papers in different scientific fields show distinct patterns in reporting verbs and hedging, reflecting discipline-specific norms for presenting findings and expressing certainty. Similarly disciplinary cultures within STEM fields influence the use of rhetorical moves and linguistic features in research articles, affecting their perceived objectivity and persuasiveness. Flowerdew and Wang (33) further highlight how specific disciplines use distinct syntactic patterns, such as the prevalence of nominalization in some fields versus more verbal constructions in others.

This disciplinary specificity arises because scientific language is socially constructed and context-dependent, shaped by disciplinary norms and the epistemological status of knowledge within each field. Bazerman (34) examined historical shifts in the rhetorical forms of scientific articles, noting, for instance, that physicists moved toward more impersonal, hypothesis-driven exposition, while biologists tended to retain more descriptive conventions. These differences appear not only in vocabulary and citation patterns but also in fine-grained syntactic features like clause length, coordination, and the use of complex noun phrases (35).

Despite extensive research on disciplinary communication, few studies have systematically compared syntactic complexity across scientific fields using standardized metrics like those from L2SCA. Staples et al. (36) applied syntactic indices to disciplinary writing samples and found statistically significant differences in mean clause length and subordination between biology, engineering, and philosophy texts. However, their analysis focused only on technical academic writing and did not consider any lay-oriented genres or public-facing scientific content. Furthermore, their measurement of syntactic complexity was limited; while they reported some clause-based measures, much of the analysis relied on broader proxies, specifically mean word length, which did not capture the full range of grammatical structuring in scientific texts. This did not account for important aspects of syntactic variation such as phrase-level embedding, coordination, and dependent clause usage, limiting the interpretability of their disciplinary comparisons. The influence of disciplinary norms on linguistic features, combined with the scarcity of systematic, detailed syntactic comparisons across disciplines (especially for lay audiences), points to a significant knowledge gap that this study addresses.

Graph-Theoretic approaches to text analysis

In addition to traditional methods for analyzing syntactic structure, graph-theoretic methods have become common in evaluating linguistic complexity and cohesion. By viewing language as a network of semantic or lexical relationships, these methods offer insights into a text's linguistic properties by extracting

information from its structural features. Seminal work by Ferrer i Cancho and Solé (37) showed that syntactic networks exhibit small-world and scale-free properties, indicating structural organization inherent in natural language. Steyvers and Tenenbaum (19) extended this framework to semantic networks, using word co-occurrence relations to model the mental lexicon and uncovering similar topological properties. These fundamental attributes of language make graph-theoretic measures well-suited to provide useful information about a text's linguistic characteristics.

Applications of these insights led to tools like TextRank (38) and LexRank (39), which use graph centrality to prioritize textual elements for summarization and keyword extraction. These studies reinforced the utility of graph-theoretic metrics, including average degree, density, and clustering coefficient, as numerical indicators of lexical cohesion and information salience. More recently, graph-theoretic measures have been extended to genre analysis and complexity assessment. Amancio et al. (14) found that the graphical properties of clustering and path length reflected differences in text coherence and readability between human and machine-generated summaries. Similarly, Antoniak et al. (13) used network statistics to analyze syntactic differences across corpora, while Mota et al. (40) used graph fragmentation as a proxy for thought disorder in psychiatric patients. These investigations aimed to capture aspects of linguistic complexity using graph characteristics, an approach heavily utilized in the present study. Network-based approaches represent a conceptual advancement in complexity analysis, addressing the limits of traditional linguistic models in capturing the intricate semantic organization and cohesion of text. This allows for a more complete understanding of textual structure.

Complexity in scientific communication

Science communication aims to make scientific knowledge accessible to non-specialist audiences without compromising accuracy. Studies in this field typically emphasize lexical simplification, metaphor use, and discourse framing, rather than syntactic or structural complexity (41). While journalistic adaptations of research articles often involve lexical substitution and content omission, the linguistic mechanisms driving accessibility remain under examined.

Prior analyses of public science writing offer some insights. Lay summaries have been shown to use shorter sentences, more personal pronouns, and fewer complex clauses. More recently, Rakedzon et al. (42) applied NLP tools to analyze scientific abstracts and their corresponding press releases, finding that lexical and syntactic features strongly predict public interest, as measured by media coverage. The emergence of dual-summary formats, such as those implemented by PNAS, provides an ideal setting for investigating how linguistic complexity is modulated across audience types. This dual-summary format was utilized by Kang et al. to evaluate differences between lay and medical summaries of medical texts from The New England Journal of Medicine. However, no prior study has integrated both traditional syntactic and

network-based complexity metrics to comprehensively assess this variation. The current study contributes to this underexplored area by integrating established syntactic measures (10) with novel graph-theoretic indices (14) in a comparative analysis of lay and technical summaries from a single, controlled source.

Methods

Corpus and data collection

This study analyzed a corpus of 300 research articles published in the Proceedings of the National Academy of Sciences (PNAS). Articles were selected only if published after August 2013, when PNAS began requiring authors to submit a "Significance Statement" alongside their abstract. This ensured consistency in the dual-summary format across the corpus. The articles were selected through sorting of the most relevant on PNAS' database. Articles were deemed relevant based on the amount of views they received in the past 30 days. This selection process prioritized articles with the highest recent readership, ensuring the corpus reflected research with the broadest public interest and relevance. Our selection criteria led to a minor imbalance in the number of papers per subject, with biological sciences being the most represented, followed by physical and then social sciences. We determined that this variation did not compromise our results because the disciplinary sample sizes were comparable, and our analysis was based on normalized metrics rather than absolute article counts. For each selected article, the significance statement served as the lay summary, and the abstract represented the technical summary.

Syntactic complexity analysis

To quantify syntactic complexity, Lu's (10) L2 Syntactic Complexity Analyzer (L2SCA) computed 14 established syntactic metrics for each text capturing various dimensions of grammatical structure across different linguistic units. A comprehensive list and definitions of all 14 syntactic metrics are in Appendix A1. These measures were computed directly from the raw text without stopword removal to preserve syntactic structures.

Graph-Theoretic analysis

To assess structural and semantic complexities, word co-occurrence graphs were constructed for each summary. Unique words within a summary were represented as nodes, and edges between nodes were weighted by their Pointwise Mutual Information (PMI). PMI values were calculated using a sentence-based co-occurrence window, meaning words within the same sentence were considered to co-occur. Only positive PMI values were retained to capture meaningful associations.

From these Word Adjacency Graphs, six normalized graph-theoretic metrics were computed for each summary. 1] ***Connected nodes per word:*** An indicator of lexical diversity, computed as the number of nodes connected to at least one other node divided by the total number of words in the text. 2] ***Average nodal degree*****:** A measure of semantic connectivity, reflecting the average number of connections (co-occurrences) a word has within the graph. 3] ***Graph density:*** An indicator of

overall textual cohesion, representing the proportion of actual connections relative to all possible connections. 4] ***Average clustering coefficient:*** A measure of modularity, indicating the tendency of words to form tight, interconnected groups (clusters). 5] ***Largest connected component size / nodes:*** A metric for conceptual integration, reflecting the size of the largest group of interconnected words relative to the total number of words. 6] ***Average edge weight associated with a node:*** A measure of lexical cohesion, the average strength of association between co-occurring words

For preprocessing, texts were lowercased and stripped of punctuation and non-alphabetic tokens. Stopwords were retained for syntactic analysis but included in the graphs only if connected via PMI. All graphs were undirected and constructed using NetworkX 3.1. Metrics were normalized as appropriate to account for textual length (e.g., degree per node, largest connected component size / nodes). For the "Connected nodes / words" metric, all words were included, which encompassed duplicates and words not included in the PMI graphs.

Statistical analysis

Distributions were summarized using medians and interquartile ranges (IQRs). For each complexity measure and subject, the proportion of outliers relative to the total number of data points did not exceed 25%, indicating sufficient regularity for the IQR to provide a reliable measure of variability. To examine differences between disciplines, we used paired *t*-tests, as our research questions emphasized specific pairwise contrasts rather than an overall omnibus effect. A repeated-measures ANOVA could have been applied, but this approach primarily tests for overall group effects, which were not central to our aims. Moreover, preliminary Levene's tests revealed significant violations of the homogeneity of variances assumption for several key metrics (all $p < 0.05$), further limiting the suitability of ANOVA. Because paired *t*-tests operate on within-item difference scores rather than raw group variances, they are robust to such heteroskedasticity and were therefore chosen as the primary statistical method.

For comparisons between lay and technical summaries of the same text, paired *t*-tests were used; for comparisons between summaries of distinct subjects, independent *t*-tests were applied. In lay/technical comparisons, each technical summary was directly paired with the corresponding lay summary of the same paper. Normality assumptions were evaluated both visually and statistically prior to analysis. All statistical procedures were performed in Python 3.10 using the SciPy 1.11.1 package.

Data reporting

For the tables comparing lay and technical summaries overall, we removed the rows with p-values > 0.05 to maintain the focus on significant metrics. For tables depicting comparisons between the three disciplinary umbrellas, we removed the rows that had no p-values < 0.05 and preserved the rows that had at least one p-value < 0.05. In the discussion section, non-significant pairwise differences are discussed for each included metric.

Results

Research question 1: How do syntactic and structural complexities differ between lay and technical summaries of scientific articles?

The first portion of our analysis focused on pairwise comparisons of the 14 traditional indices of syntactic complexity between technical and lay summaries. T-tests indicated that eight of the fourteen indices indicated statistically significant differences. Of these, the MLS, MLT, MLC, and VP/T demonstrated highly significant differences ($p < 0.001$), the DC/C and DC/T exhibited moderately significant differences, and the CT/T and C/T displayed marginally significant differences.

The MLS, MLT, and MLC indicated that technical summaries tended to utilize longer production units and these units generally included longer clauses. Additionally, lay summaries exhibited a significantly higher incidence of verb phrases (VP/T) and total clauses (C/T) as well as a higher proportion of dependent clauses (DC/C, DC/T). Curiously, lay summaries also included a higher proportion of complex t-units (CT/T), although this effect size was comparatively small. This suggests that, despite their aim for clarity, lay summaries may introduce complexity by layering clauses and embedding ideas, rather than by using dense vocabulary or long noun phrases.

Table 1. Comparisons between Lay and Technical summaries for traditional indices of syntactic complexity

Metric	Lay Summary	Technical Summary	P-value (t-value)
MLS	23.9 (20.79, 27.0)	25.66 (23.0, 28.11)	**<0.001** (-4.270)
MLT	22.8 (19.82, 25.4)	23.76 (21.12, 26.85)	**<0.001** (-3.061)
MLC	15.13 (12.58,17.57)	16.46 (14.28, 18.69)	**<0.001** (-3.284)
VP/T	2.2 (1.88, 2.67)	2.09 (1.78, 2.44)	**<0.001** (-4.573)
C/T	1.5 (1.29, 1.8)	1.44 (1.25, 1.67)	0.043 (2.029)
DC/C	0.37 (0.25, 0.46)	0.33 (0.23, 0.41)	0.008 (2.650)
DC/T	0.5 (0.33, 0.8)	0.45 (0.29, 0.68)	0.001 (3.247)
CT/T	0.49 (0.29, 0.62)	0.4 (0.27, 0.57)	0.048 (1.985)

Values in columns 2-3 are presented as Median (First Quartile, Third Quartile). Columns 4 presents P-value (t-value). P values < 0.05 p appear in bold.

The second portion of our analysis centered on pairwise comparisons of the six graph-theoretic measures of lexical complexity between technical and lay summaries. Our results indicate that five of the six metrics displayed highly significant results (connected nodes / words, average nodal degree, density, average clustering, largest component size, and average edge weight associated with a node.)

Lay summaries indicated significantly greater overall lexical cohesion (connected nodes /

words, density, average clustering). However, despite displaying a higher degree of cohesion within lexically linked text-portions, lay summaries were associated with a weaker average association between words and fewer overall connections per word globally (average edge weight associated with a node, average nodal degree). This suggests that while words in lay summaries tend to co-occur more often and form tighter local clusters, they are less interconnected across the broader text, potentially reflecting a more diverse vocabulary with lower semantic redundancy and specificity.

Table 2. Comparisons between Lay and Technical summaries for Graph-Theoretic metrics

Metric	Lay Summary	Technical Summary	P-value (t-value)
Connected nodes / words	0.67 (0.64, 0.71)	0.58 (0.54, 0.62)	**< 0.001** (23.9072)
Average nodal degree (full graph)	29.50 (26.62, 33.41)	37.24 (33.35, 41.24)	**< 0.001** (-19.4025)
Density	0.44 (0.39, 0.48)	0.35 (0.31, 0.39)	**< 0.001** (18.1891)
Average clustering	0.80 (0.78, 0.83)	0.75 (0.73, 0.77)	**< 0.001** (23.4018)
Average edge weight (PMI) associated with a node	315.40 (279.06, 355.88)	409.81 (354.99, 455.95)	**< 0.001** (-18.7654)

Values in columns 2-3 are presented as Median (First Quartile, Third Quartile). Columns 4 presents P-value (t-value). P values < 0.05 p appear in bold.

Research question 2: To what extent does linguistic complexity vary across scientific disciplines within both lay summaries and technical summaries?

To address Research Question 2, we first computed differences in the 14 syntactic complexity indices between subject-specific lay summaries. The results indicate that the physical and social sciences exhibited the greatest number of statistically significant differences (MLC, C/S, C/T, DC/C, DC/T, CT/T, CN/C), followed by the physical and biological sciences (MLS, MLT, MLC, CN/T, CN/C). The biological and social sciences were the most similar, differing significantly on only five metrics (C/S, C/T, DC/C, DC/T, CT/T).

The physical sciences employed longer clauses (MLC), but included less clauses per production unit (C/S, C/T) and a lower proportion of complex t-units (CT/T) than the social sciences. They also utilized a lower proportion of dependent clauses (DC/C, DC/T) but a higher number of complex nominals (CN/C). This implies that the physical sciences tend to use more information-dense noun phrases within longer, more syntactically compressed structures than the social sciences, favoring nominal complexity over clausal subordination. The physical sciences implemented longer production units (MLS, MLT) and clauses (MLC) as well as a higher proportion of complex nominals (CN/C) than the biological sciences, reflecting the previously established tendency of the physical

sciences favoring longer and more nominally complex structures.The biological sciences featured less clauses than the social sciences (C/S, C/T) as well as less dependent clauses per production unit (DC/C, DC/T) and a lower ratio of complex t-units (CT/T), suggesting that the biological sciences rely on simpler syntactic structures than the social sciences with fewer layers of subordination and complexity, favoring clarity and conciseness over elaboration.

Table 3. Disciplinary comparisons within Lay summaries for traditional indices of syntactic complexity

Metric	Biological Sciences	Physical Sciences	Social Sciences	Physical vs. Biological Sciences	Physical vs. Social Sciences	Biological vs. Social Sciences
MLS	23.29 (20.5, 25.2)	24.6 (21.75, 28.5)	24.2 (20.42, 28.25)	**0.0031** (2.984)	0.1622 (1.405)	0.3055 (-1.027)
MLT	21.63 (19.67, 24.8)	24.25 (20.5, 27.27)	23.25 (19.0, 26.38)	**0.0028** (3.025)	0.1245 (1.545)	0.3913 (-0.859)
MLC	14.81 (12.68, 16.86)	16.86 (13.72, 20.30)	14.38 (12.1, 15.93)	**0.0002** (3.787)	**$9.594x10^{-5}$** (4.014)	0.0778 (1.772)
C/S	1.6 (1.33, 1.8)	1.5 (1.23, 1.78)	1.67 (1.43, 2.0)	0.3669 (-0.904)	**0.0075** (-2.712)	**0.0077** (-2.690)
C/T	1.5 (1.29, 1.75)	1.4 (1.2, 1.75)	1.67 (1.43, 1.85)	0.3944 (-0.853)	**0.0063** (-2.771)	**0.0200** (-2.342)
DC/C	0.38 (0.25, 0.45)	0.33 (0.18, 0.44)	0.4 (0.29, 0.5)	0.1344 (-1.502)	**0.0021** (-3.131)	**0.0186** (-2.372)
DC/T	0.5 (0.33, 0.8)	0.5 (0.2, 0.75)	0.67 (0.43, 1.0)	0.3120 (-1.013)	**0.0072** (-2.728)	**0.0186** (-2.370)
T/S	1.0 (1.0, 1.17)	1.0 (1.0, 1.0)	1.0 (1.0, 1.17)	0.9898 (-0.013)	0.4809 (-0..707)	0.4519 (-0.754)
CT/T	0.4 (0.29, 0.6)	0.4 (0.2, 0.67)	0.5 (0.33, 0.71)	0.3013 (-1.036)	**0.0146** (-2.472)	**0.0388** (-2.078)
CN/T	3.37 (2.8, 4.2)	3.6 (3.08, 4.25)	3.67 (2.57, 4.75)	**0.0287** (2.201)	0.6155 (0.503)	0.1849 (-1.330)
CN/C	2.27 (1.8, 2.75)	2.43 (2.0, 3.2)	2.27 (1.65, 2.94)	**0.0034** (2.962)	**0.0084** (2.674)	0.4641 (0.733)

Values in columns 2-4 (Biological Science, Physical Sciences, and Social Sciences) are presented as Median (First Quartile, Third Quartile). Columns 5-7 present P-value (t-value). P values < 0.05 p appear in bold.

Next, we computed differences between subject-specific technical summaries in the 14-indices. The number of statistically significant differences shrunk from 18 to 9, implying that many of the inter-disciplinary differences that were present in the lay summaries were absent

in technical summaries.The physical sciences continued to feature longer production units (MLS, MLT) than the biological sciences but also exhibited a higher frequency of verb phrases, (VP/T) a difference that was not present in the lay summaries. The rest of the previously observed differences between the physical and biological sciences within the lay summaries were statistically insignificant for the technical summaries. All of the differences between the physical and social sciences, which had previously exhibited the highest number of statistically significant differences within the lay summaries, were rendered insignificant for technical summaries. The biological sciences featured longer clauses (MLC) than the social sciences but continued to harbor lower values of other measures of textual complexity (C/S, VP/T, C/T, CN/T) and feature less subordination (DC/T). Thus, the physical and social sciences differed substantially less in technical contexts compared to lay writing, the physical and biological social sciences differed moderately less in technical contexts, and the biological and social sciences continued to demonstrate several significant differences for both technical and lay audiences. This implies that technical writing generally tends to be more uniform than writing directed towards a lay audience.

Table 4. Disciplinary comparisons within Technical summaries for traditional indices of syntactic complexity

Metric	Biological Sciences	Physical Sciences	Social Sciences	Physical vs. Biological Sciences	Physical vs. Social Sciences	Biological vs. Social Sciences
MLS	25.42 (22.64, 27.90)	26.57 (23.55, 28.35)	25.4 (23.14, 28.33)	**0.0226** (2.295)	0.1190 (1.568)	0.9933 (-0.008)
MLT	23.36 (20.86, 26.3)	25.1 (22.09, 27.27)	23.67 (21.05, 26.71)	**0.0138** (2.483)	0.0951 (1.680)	0.9339 (-0.083)
MLC	16.64 (14.64, 19.06)	16.38 (14.31, 19.88)	15.82 (13.26, 17.96)	0.0868 (1.720)	0.0743 (1.798)	**0.0168** (2.410)
C/S	1.5 (1.3, 1.72)	1.56 (1.32, 1.82)	1.6 (1.44, 1.94)	0.2894 (1.062)	0.3218 (-0.994)	**0.0184** (-2.375)
VP/T	2.0 (1.7, 2.34)	2.11 (1.86, 2.47)	2.22 (1.88, 2.62)	**0.0254** (2.250)	0.7127 (-0.369)	**0.0077** (-2.690)
C/T	1.38 (1.22. 1.67)	1.5 (1.29, 1.71)	1.57 (1.29, 1.75)	0.1309 (1.516)	0.4172 (-0.814)	**0.0123** (-2.524)
DC/T	0.44 (0.27, 0.67)	0.45 (0.28, 0.67)	0.55 (0.33, 0.8)	0.4776 (0.711)	0.2888 (-1.065)	**0.0310** (-2.171)
CN/T	3.41 (3.0, 4.14)	3.7 (3.18, 4.35)	3.83 (2.90, 4.58)	0.0521 (1.952)	0.9683 (-0.040)	**0.0488** (-1.981)

Values in columns 2-4 (Biological Science, Physical Sciences, and Social Sciences) are presented as Median (First Quartile, Third Quartile). Columns 5-7 present P-value (t-value). P values < 0.05 p appear in bold.

Subsequently, we used the graph-theoretic indices to investigate differences between the scientific disciplines. The differences were far less pronounced than with the traditional indices. For lay summaries, the sole statistically significant differences were in the average nodal degree, connected nodes/words, and average edge weight between the physical and biological sciences. The physical science summaries exhibited higher average nodal degrees, connected nodes/words, and higher average nodal edge weights, suggesting that the physical sciences were more conceptually dense and interconnected, potentially reflecting a tighter integration of technical terminology and more cohesive conceptual frameworks within their discourse. A representative lay summary and technical summary, along with corresponding network visualizations and a table detailing the nodal connections, are presented below for illustrative purposes.

Representative Lay summary

The northern white rhinoceros (NWR; Ceratotherium simum cottoni) is functionally extinct, with only two nonreproductive females remaining alive. Extraordinary measures are underway to rescue this species, including using a collection of NWR induced pluripotent stem cells (iPSCs) to generate gametes for assisted reproduction technologies. Because of the critical importance of genomic integrity in germ cells used for reproduction, these approaches require extensive genomic analyses to exclude aberrations that are acquired during culture of iPSCs. In order to support those efforts, we have generated a chromosome-level genome assembly of northern white rhinoceros and used this reference genome to evaluate the genomic integrity of iPSCs cultured for the generation of artificial gametes.

Representative Technical summary

The northern white rhinoceros (NWR; Ceratotherium simum cottoni) is functionally extinct, with only two nonreproductive females alive. Efforts to rescue the NWR from its inevitable demise have inspired the exploration of unconventional conservation methods, including the development of induced pluripotent stem cells (iPSCs) for the in vitro generation of artificial gametes. The integrity of iPSC genomes is critical for in vitro gametogenesis to be used for assisted reproductive technologies using NWR iPSCs. We generated a chromosome-level NWR reference genome that meets or exceeds the metrics proposed by the Vertebrate Genome Project, using complementary sequencing and mapping methods. The genome represents 40 autosomes, an X and a partially resolved Y chromosome, and the mitochondrial genome. Using comparative FISH mapping, we confirmed a general gene order conservation between the NWR and horse genomes. We aligned the NWR genome with that of the southern white rhinoceros (SWR; Ceratotherium simum simum), a population that has been physically separated from the NWR for tens of thousands of years, and we found that the two subspecies are very similar on the chromosome level. Comparing long-read data from NWR iPSC lines and the fibroblast cultures used for reprogramming, we identified copy number variations that were likely to have been introduced during in vitro iPSC expansion. The NWR reference genome allows for efficient, rapid, and accurate

assessment of the genomic integrity of iPSC lines to direct their differentiation. This will assist in strategies to rescue the NWR through extraordinary measures like cloning and the generation of embryos from iPSC-derived gametes.

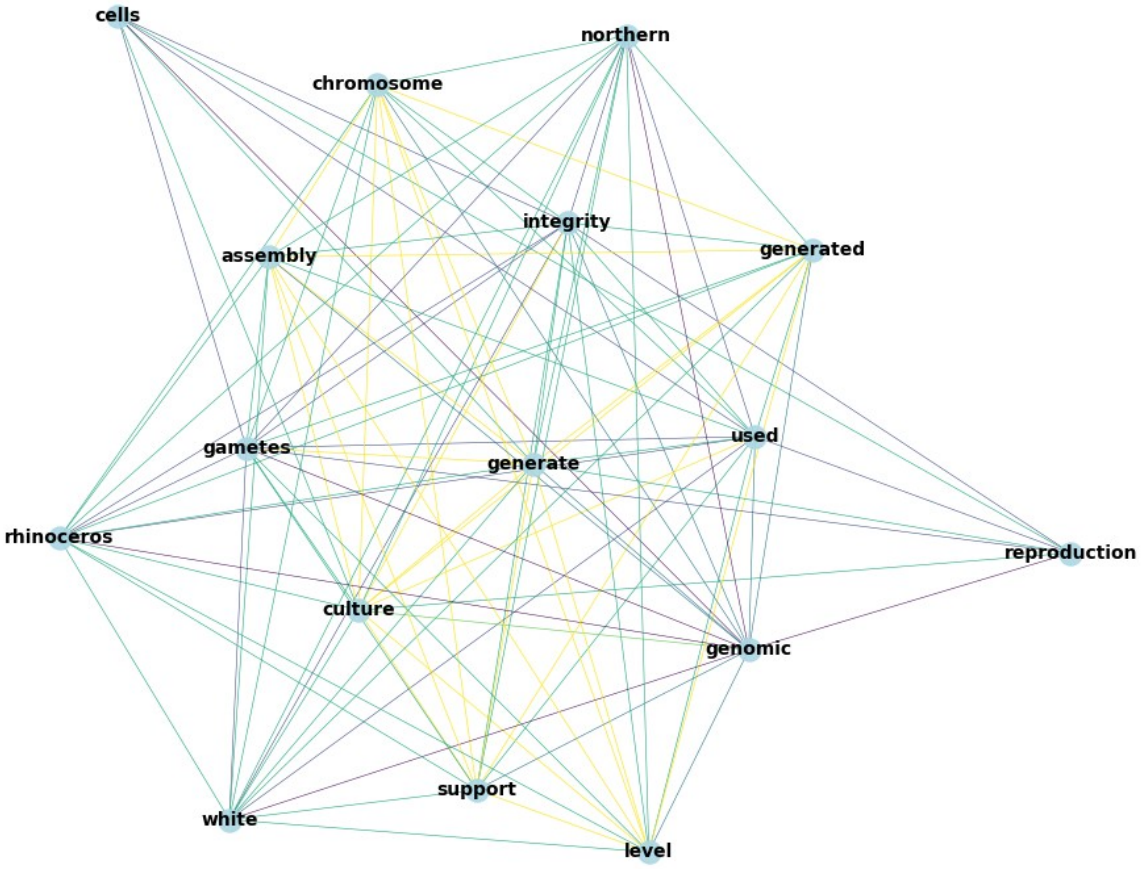

Figure 1. Word Adjacency Graph for a Lay Summary. This graph represents the lay summary of a scientific article. Lighter edges indicate stronger word associations, while darker edges show weaker ones. For visual clarity, this subgraph contains only the 20% of non-stop words with the highest degrees.

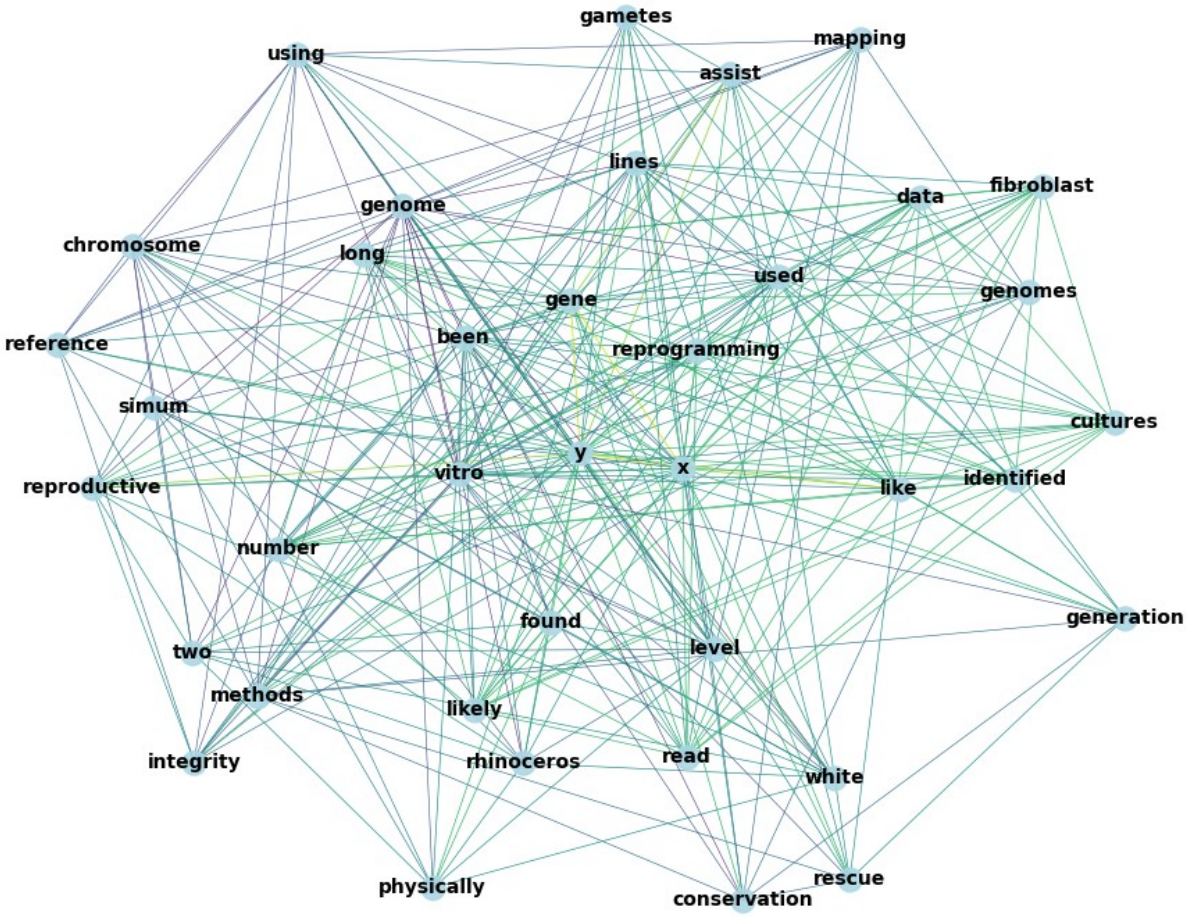

Figure 2. Word Adjacency Graph for a Technical Summary. This graph is a representative Word Adjacency Graph for the technical summary of the same scientific article as Figure 1. It shares the same properties, with lighter edges indicating stronger word associations and darker edges showing weaker ones. For visual clarity, this subgraph also contains only the 20% of non-stop words with the highest degrees.

Table 5. Comparison of Graph-Theoretic metrics in a Lay–Technical summary pair

Metric	Lay Summary	Technical Summary
Connected nodes / words	0.6577	0.4331
Average nodal degree (full graph)	30.92	19.05
Density	0.4589	0.1907
Average clustering	0.7639	0.7296
Largest component size / number of total nodes (full graph) (S/N)	0.9359	0.9167
Average edge weight (PMI) associated with a node	351.7	248.5

Table 6. Disciplinary comparisons within Lay summaries for Graph-Theoretic metrics

Metric	Biological Sciences	Physical Sciences	Social Sciences	Physical vs. Biological Sciences	Physical vs. Social Sciences
Connected nodes / words	0.67 (0.63, 0.70)	0.68 (0.66, 0.72)	0.66 (0.63, 0.71)	**0.0411** (2.054)	0.1202 (1.563)
Average nodal degree (full graph)	29.05 (26.01, 32.94)	30.48 (27.97, 33.20)	29.78 (26.71, 33.67)	**0.0165** (2.414)	0.6140 (0.505)
Average edge weight (PMI) associated with a node	321.08 (282.27, 362.41)	340.95 (303.77, 378.27)	326.03 (290.00, 369.16)	**0.0118** (2.539)	0.5424 (0.611)

Values in columns 2-4 (Biological Science, Physical Sciences, and Social Sciences) are presented as Median (First Quartile, Third Quartile). Columns 5-7 present P-value (t-value). P values < 0.05 p appear in bold.

As observed with the 14-indices, differences in the graph-theoretic indices between subjects were even less pronounced for technical summaries than for lay summaries. The only statistically significant differences were observed between the physical and biological sciences for the average degree of a node and between the biological and social sciences for graph density. Physical science summaries displayed higher average nodal degrees, indicating greater lexical cohesion than biological science summaries as observed in lay summaries. Social science summaries were associated with higher graph density than biological science summaries, possibly indicating that they tend to include a tighter conceptual framework than the biological sciences. However, due to the lack of significant differences in other metrics, the evidence is insufficient to conclude that there is

a meaningful difference in lexical cohesion between any two subjects. Overall, the data corroborated the previously demonstrated trend that technical writing tends to be more unvarying than writing directed towards a lay audience.

Table 7. Disciplinary comparisons within Technical summaries for Graph-Theoretic metrics

Metric	Biological Sciences	Physical Sciences	Social Sciences	Physical vs. Biological Sciences	Physical vs. Social Sciences
Average nodal degree (full graph)	35.92 (32.44, 40.54)	38.0 (33.71, 41.57)	38.17 (34.10, 42.11)	**0.0439** (2.026)	0.8559 (0.182)
Density	0.33 (0.30, 0.38)	0.35 (0.31, 0.38)	0.37 (0.33, 0.43)	0.0627 (1.871)	0.1811 (-1.344)

Values in columns 2-4 (Biological Science, Physical Sciences, and Social Sciences) are presented as Median (First Quartile, Third Quartile). Columns 5-7 present P-value (t-value). P values < 0.05 p appear in bold.

Discussion

Audience-based and disciplinary similarities in textual complexity

Our analysis revealed stable foundations of linguistic structures across both audiences and disciplines. Despite their divergent approaches to complexity, lay and technical summaries did not differ in measures of syntactic density, such as clauses per sentence or the use of complex nominals. Likewise, the overall conceptual integration, measured by the relative size of the largest semantic network, remained consistent between both summary types. This suggests that the shared constraints of summarizing identical content within a strict word limit compel authors to preserve a common syntactic and semantic scaffold, even while adapting surface-level features for different audiences.

While these findings highlight the shared linguistic foundations of scientific writing, they simultaneously accentuate the more pronounced differences that emerge. The following sections examine these audience- and discipline-specific divergences, demonstrating how writers strategically adapt linguistic complexity to address distinct communicative goals.

To further clarify why certain metrics converge across audiences and disciplines while others diverge, we provide interpretive tables that highlight the sources of overlap. These summaries illustrate how conventions of scientific summarization constrain linguistic variation across contexts.

Table 8. Explanations for similarities in linguistic metrics across disciplines (Lay summaries)

Metric	**Physical vs. Biological Sciences**	**Physical vs. Social Sciences**	**Biological vs. Social Sciences**
Connected nodes / words	N/A	Physical and social sciences draw on narrower, subject-specific lexicons, producing similar lexical connectivity.	Although social sciences rely on somewhat more specific terminology than biological sciences, the vocabulary remains moderately diffuse, resulting in substantial overlap and non-significance.
Average nodal degree (full graph)	N/A	Both physical and social sciences feature higher nodal degrees due to more condensed vocabularies.	Social sciences exhibit moderately higher interconnections than biology, but variability reduces statistical significance.
Average edge weight (PMI) associated with a node	N/A	Physical and social sciences display stronger word associations due to constrained conceptual frameworks.	Social sciences' conceptual frameworks are intermediate between physical and biological sciences, producing overlap that prevents significance.
MLS, MLT	N/A	Physical and social sciences tend to employ longer sentences and T-units due to greater syntactic layering.	Social sciences exhibit only moderate layering, resulting in overlap with biological sciences and non-significant differences.
MLC	N/A	N/A	Social and biological sciences both employ shorter clauses than physical sciences, relying more on coordination and subordination.
C/S, C/T	Biological and physical sciences exhibit fewer clauses per sentence/T-unit than social sciences, but distributions overlap.	N/A	N/A
DC/C, DC/T	Biological and physical sciences show lower rates of dependent clauses than social sciences, reflecting less subordination.	N/A	N/A

T/S	The number of T-units per sentence is consistently close to 1 across disciplines, reflecting avoidance of multiple independent clauses and preference for syntactic simplicity.		
CT/T	Biological and physical sciences employ fewer complex T-units with similar interquartile ranges, indicating less reliance on subordination/coordination than social sciences.	N/A	N/A
CN/T	N/A	Physical and social sciences have higher CN/T than biological sciences, but IQR overlap reduces statistical separation.	Social sciences show slightly higher median than biology, but greater variability and overlapping IQRs reduce effect size and significance.
CN/C	N/A	N/A	Social and biological sciences exhibit fewer complex nominals with substantial IQR overlap, reflecting lower rates of discipline-specific terminology.

Audience-based differences in linguistic complexity

The findings show that technical summaries are not simply inherently more complex than lay summaries. Instead, both summary types show distinct forms of complexity, each aligned with their communication goals. Technical summaries rely more on grammatical compression, while lay summaries often use more elaboration and surface-level cohesion.

Syntactically, technical summaries consistently had significantly higher values for Mean Length of Sentence (MLS), Mean Length of T-unit (MLT), and Mean Length of Clause (MLC). These results indicate that technical summaries use longer production units, and these units generally contain longer clauses. This reflects the dense constructions typical of academic writing, where conciseness and high information density are prioritized for an expert audience. Conversely, lay summaries showed a significantly higher incidence of verb phrases (VP/T) and total clauses (C/T), as well as a higher proportion of dependent clauses (DC/C, DC/T). While lay summaries also showed a higher proportion of complex T-units (CT/T), the effect size for this metric was comparatively small. This pattern suggests a deliberate shift from information-dense nominalizations, common in technical writing, to more reader-oriented syntactic choices and clausal elaboration. Technical writing prioritizes information density through compression, assuming expert knowledge, while lay writing prioritizes ease of processing through explicit, clause-based structures, reducing cognitive load for a broader audience.

Graph-theoretic metrics reveal textual characteristics not easily captured by syntactic analysis alone (14, 19). Lay summaries showed significantly higher values in graph density, clustering, and the ratio of connected nodes to total words. These features suggest that lay summaries are more locally cohesive; terms are more likely to co-occur in small, tightly interconnected groups, which likely reflects a strategic approach to reinforce key concepts and maintain clarity for a non-specialist audience (14). However, despite this heightened local cohesion, lay summaries were also associated with a weaker average association between words and fewer overall connections per word globally (lower average edge weight and average nodal degree). A higher average edge weight, as seen in technical summaries, implies a stronger, more specific semantic relationship between co-occurring words, indicating a more tightly integrated and specialized conceptual structure (44, 45). This apparent contradiction suggests a sophisticated adaptation strategy of achieving clarity by creating strong, localized semantic anchors, but avoiding overly specialized global semantic networks that might overwhelm a non-expert. This contrasts with technical summaries, which build highly interconnected, specific semantic fields for experts.

Disciplinary variation in linguistic complexity

The study's findings on disciplinary variation show a contrast between lay and technical summaries, suggesting different factors influence linguistic style.

Lay summaries: disciplinary voice re-emerges

When the strict constraints of technical language are relaxed, a distinct disciplinary voice becomes more pronounced in lay summaries, as scientists appear to use the rhetorical norms of their fields (15, 46). The analysis of lay summaries showed significant syntactic differences across disciplines. Our results suggest that the physical sciences favor information-dense noun phrases within syntactically compressed structures, prioritizing nominal complexity over clausal subordination to convey meaning, even when writing for general audiences. This may reflect underlying assumptions about the technical accuracy needed in fields like physics and engineering, where precision and conciseness are paramount (34). In contrast, lay summaries in the social sciences made heavier use of clause-based elaboration, including more dependent clauses and more clause-per-T-unit structures. This stylistic choice likely reflects a disciplinary tradition of emphasizing argumentation, explanation, and social context, even when addressing non-specialists (15). Biological sciences lay summaries showed intermediate values, differing in some ways from both the physical and social sciences. This variation may stem from the field’s dual identity as both experimentally grounded and socially relevant, requiring both technical precision and public engagement. The distinct disciplinary patterns observed in lay summaries suggest that "simplification" for a general audience is not a uniform process, but rather a context dependent adaptation, indicating a deeper, discipline specific rhetorical practice.

Technical summaries: homogeneity across disciplines

In contrast to the lay summaries, the analysis of subject-specific technical summaries showed a decrease of statistically significant inter-disciplinary differences across the 14 syntactic indices. The only notable exceptions were persisting differences between the biological and social sciences. This homogeneity suggests that technical writing styles, perhaps driven by academic conventions, journal requirements, and the shared imperative of precise scientific language, tend to be relatively uniform across scientific fields (30, 31). The strong constraints of the technical writing appear to homogenize scientific communication, overriding disciplinary stylistic preferences that re-emerge when writing for a broader audience.

To illustrate the basis for these convergences, we provide a companion table describing why specific differences between disciplines in technical summaries are largely absent.

Table 9. Explanations for similarities in linguistic metrics across disciplines (Technical summaries)

Metric	**Physical vs. Biological Sciences**	**Physical vs. Social Sciences**	**Biological vs. Social Sciences**
Average nodal degree (full graph)	N/A	Both physical and social sciences feature higher nodal degrees due to more condensed vocabularies.	Although social sciences show moderately higher connectivity than biological sciences, overlapping ranges reduce statistical separation.
Density	N/A	Both physical and social sciences feature higher densities due to more interrelated concepts	Social sciences display higher median densities than biological sciences, but due to substantial overlap in interquartile ranges, this effect size is reduced.
MLS, MLT	N/A	Physical sciences trend longer than social sciences, but technical writing norms for clarity constrain extremes, producing overlap.	Sentence and t-unit lengths in biology and social sciences are nearly identical since both prioritize concise reporting of results.
MLC	Biological and physical sciences show similar clause lengths because both disciplines rely on moderately complex statements for precision.	Physical sciences display somewhat longer clauses than social sciences, but all disciplines use subordination sparingly in technical writing.	N/A

C/S	Sentence-level clause counts are nearly identical because both biology and physical sciences avoid stacking clauses to maintain clarity.	Physical and social sciences show close medians since both disciplines favor one to two clauses per sentence to convey results directly.	N/A
VP/T	N/A	Physical and social sciences overlap because both employ specialized noun phrases to describe processes, limiting VP/T divergence.	N/A
C/T	Physical sciences employ more clauses per T-unit than biology, but overlap occurs because all disciplines restrict excess coordination/subordination.	Physical and social sciences are broadly similar, as both prioritize compact, single-clause T-units for technical clarity.	N/A
DC/T	Physical and biological sciences exhibit nearly identical dependent-clause rates, as both disciplines favor parataxis over subordination.	Physical and social sciences show comparable use of dependent clauses, since technical prose avoids excessive subordination.	N/A
CN/T	Physical sciences exhibit moderately more complex nominals than biology, but overlap occurs because both disciplines rely heavily on formulaic noun phrases.	Physical and social sciences show nearly identical rates because both disciplines package concepts densely within noun phrases.	N/A

Implications and contributions

The findings of this study carry significant implications for several areas, particularly in enhancing science communication and education. For automated readability assessments, the results suggest that future tools need to account for multi-dimensional complexity and the audience-dependent linguistic strategies employed by writers, moving beyond surface-level indicators (9). Understanding these nuanced complexity profiles can lead to the development of more sophisticated algorithms that accurately gauge text accessibility for diverse audiences. Furthermore, these findings are highly relevant for cross-disciplinary science education, highlighting the need to teach not only scientific content but also the discipline-specific communication norms and audience adaptation strategies that are crucial for effective knowledge transfer (15). For instance, educators can use these insights to guide students in tailoring their scientific writing for different audiences, emphasizing the strategic choices involved in balancing precision and clarity. Ultimately, this study contributes to both theoretical models of language complexity and practical discussions around accessible science communication, clarifying the strategic balance between accuracy and accessibility that underlies effective scientific discourse.

Limitations

This study's findings should be considered within its specific scope. Our analysis was limited to articles published in a single journal, PNAS, which, while providing a controlled environment for comparative analysis, may not fully represent the diversity of scientific communication across all academic publishing venues. Future research could expand the corpus to include multiple journals and different types of scientific texts to enhance generalizability.

Perspectives

Building upon our findings that lay and technical scientific abstracts exhibit distinct linguistic and structural characteristics—particularly regarding syntactic complexity and word co-occurrence patterns (lay local-clusters vs. technical universal-connections)—a compelling avenue for future research lies in developing a quantitative measure of content distortion. This theoretical tool, which we term the **Out-of-Context Index** (OOCI) or **Exaggeration Index** (EI), would be designed to quantify the degree to which a lay summary oversimplifies, exaggerates, or takes a scientific finding out of its original, qualified, technical context.

The OOCI's theoretical framework requires a hybrid methodology that combines our automated quantitative metrics with a qualitative content analysis. Specifically, the index could be constructed by establishing a correlation between two main components: 1] **Linguistic Metric Ratio**: The ratio of lay 'local-clusters' (representing simplified or common-language motifs) to technical 'universal-connections' (i.e. those which would represent technically qualified and complex motifs). 2] **Manual Curation of Distortion**: A manually curated count of instances of exaggerated words or simplistic, unqualified descriptions within the lay summaries.

By correlating the ratio of these structural differences to the manually curated content distortions, the OOCI could serve as an indicator for quantifying the potential for misinterpretation inherent in translational communication.

This index would hold significant value for diverse audiences. It could help prevent embellishment and misinformation of socially important research. For investors, the OOCI could provide an objective metric to evaluate the verisimilitude of corporate summaries—such as the "management summary" or sections of "10-K" reports—against the actual, qualified findings reported in the underlying technical publications.

As an example that illustrates this content distortion, consider a pharmaceutical trial concerning an Alzheimer's drug. The technical summary may state: "These findings are suggestive of an increased expression of the Alzheimer Neuro-fibriallary tangle reducing protein X upon the administration of drug Y, in-part due to the release of transcription factor Z from mutated complexasome A... Therefore, the effect of the drug Y may be conditional upon a mutation in the binding pocket of complexasome A, limiting its potential to patients who exhibit this mutation." This text is

highly qualified, establishing a narrow scope for efficacy.

In contrast, the corresponding lay summary may simplify this by stating: "In a major achievement, drug Y was found to be effective against Alzheimer's disease in a phase III clinical trial." This simplified statement removes the critical contextual limitation (the required mutation), resulting in an exaggeration of the drug's true potential. The OOCI would, through its correlation with the cluster ratio, signal the extent to which the lay language oversimplifies or exaggerates the qualified technical finding.

Given that such an index does not yet exist, a dedicated, subsequent study focusing on the manual curation of a large corpus would represent a worthwhile and substantial contribution to the literature on translational science communication.

Conclusion

The results of this study indicate that lay and technical writing differ significantly in their complexity profiles, but not in simple or uniform ways. Technical summaries were characterized by syntactic compactness, showing longer sentences, T-units, and clauses, along with higher rates of nominal complexity. These features reflect the dense, information-rich constructions typical of academic writing, tailored for an expert audience (26, 29). In contrast, lay summaries consistently used greater clausal elaboration and showed higher local lexical cohesion, as evidenced by increased graph density and clustering. However, they also showed lower global semantic specificity, suggesting a deliberate strategy to achieve broader accessibility by reinforcing key concepts locally while avoiding an excessively dense global conceptual network.

Crucially, the study revealed a striking divergence in disciplinary linguistic patterns. While technical writing styles showed remarkable homogeneity across scientific fields (30, 31), lay summaries showed the re-emergence of distinct disciplinary voices. The study's dual-framework approach provides a more precise understanding of scientific language, showing the diverse strategies researchers use to balance accuracy and accessibility. It advances theoretical models of language complexity by demonstrating the complementary insights gained from integrating syntactic and network-based analyses, moving beyond single-dimensional metrics to capture the multifaceted nature of textual organization.

Future research should extend this inquiry by investigating the direct impact of these identified complexity patterns on actual reader comprehension and engagement. While the current metrics highlight how language is structured, they do not directly capture how it is received. Experimental work incorporating comprehension testing, eye-tracking, or think-aloud protocols could help connect these linguistic features to real-world outcomes, bridging the gap between linguistic analysis and communication effectiveness. Additionally, expanding this analysis to other journals, languages, or alternative modes of scientific communication (e.g., policy briefs,

science news articles, or social media posts) would provide a broader and more nuanced understanding of how scientific knowledge is transferred between expert and non-expert domains.

References

1. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

2. Sharp, P. A., Bonvillian, W. B., Brand, A., Goldston, D., Stebbins, M. (2024). The future of open research policy should be evidence based. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(32), e2412688121. https://doi.org/10.1073/pnas.2412688121

3. Schipani, V. (2024). Journalism and public trust in science. *PhilSci-Archive*. https://philsci-archive.pitt.edu/23677/

4. Metag, J., Wintterlin, F., Klinger, K. (2023). Editorial: Science communication in the digital age—New actors, environments, and practices. *Media and Communication*, *11*(1), 212–216. https://doi.org/10.17645/mac.v11i1.6905

5. Goldstein, C. M., Krukowski, R. A. (2023). The importance of lay summaries for improving science communication. *Annals of Behavioral Medicine*, *57*(7), 509–510. https://doi.org/10.1093/abm/kaad027

6. Kuehne, L. M., Olden, J. D. (2015). Opinion: Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, *112*(12), 3585–3586. https://doi.org/10.1073/pnas.1500882112

7. Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

8. Just, M. A., Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

9. 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. https://doi.org/10.7554/eLife.27725

10. Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

11. Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

12. Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

13. Antoniak, M., Mimno, D., Smith, N. A. (2015). A quantitative survey of lexical features in literature. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 437–446). Association for Computational Linguistics.

14. Amancio, D. R., Oliveira Jr, O. N., Costa, L. da F. (2012). Graph-based ranking algorithms for sentence extraction, applied to scientific articles. *Journal of Informetrics*, *6*(4), 427–434. https://doi.org/10.1016/j.joi.2012.01.002

15. Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, MI: University of Michigan Press.

16. Wang, Z., Jabar, M. A. A., Jalis, F. M. M. (2024). Cross-disciplinary analysis of the syntactic and lexical features of Chinese master thesis titles. *Theory and Practice in Language Studies*, *14*(9), 481–487.

17. Hartley, J. (2003). Improving the clarity of journal abstracts in psychology: The case for structure. *Science Communication*, *24*(3), 366–379. https://doi.org/10.1177/1075547002250301

18. Kyle, K., Crossley, S. A. (2015). Automatically assessing syntactic sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 761–786. https://doi.org/10.1002/tesq.194

19. Steyvers, M., Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3

20. Larsson, T., Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes*, *45*, 100850. https://doi.org/10.1016/j.jeap.2020.100850

21. DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.

22. Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii Press.

23. Casal, J. E., Lu, X., Qiu, X., Wang, Y., Zhang, G. (2021). Syntactic complexity across academic research article part-genres: A cross-disciplinary perspective. *Journal of English for Academic Purposes*, *52*, 100996. https://doi.org/10.1016/j.jeap.2021.100996

24. Nasrabady, P., Khoshsima, H., Mohammadian, A., Yarahmadzehi, N. (2025). A corpus-based evaluation of syntactic complexity measures as indices of advanced English text comprehension in EAP textbooks and academic research papers. *Research in English Language Pedagogy*, *13*(2), 1–17. https://doi.org/10.71673/relp.2025.1195182

25. Zhao, M., Ge, T. (2024). A comparative analysis of syntactic complexity in applied linguistics abstracts written by Chinese novice writers and native English advanced writers. *Open Journal of Applied Sciences*, *14*, 1–26. https://doi.org/10.4236/ojapps.2024.141001

26. Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

27. Crossley, S. A., McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

28. Rühlemann, C. (2007). *Conversation in context: A corpus-driven approach*. London, UK: Continuum.

29. Biber, D., Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge, UK: Cambridge University Press.

30. Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London, UK: Continuum.

31. Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

32. Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam, Netherlands: John Benjamins Publishing.

33. Flowerdew, J., Wang, H. (2015). Identity in academic discourse. *Annual Review of Applied Linguistics*, *35*, 81–99. https://doi.org/10.1017/S026719051400021X

34. Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison, WI: University of Wisconsin Press.

35. Pérez-Guerra, J., Smirnova, E. A. (2021). Disciplinary variation in syntactic complexity: A corpus analysis of professional academic writing. *CEUR Workshop Proceedings*, *Vol. 3090*. Retrieved from https://ceur-ws.org/Vol-3090/spaper29.pdf

36. Staples, S., Egbert, J., Biber, D., Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149–183. https://doi.org/10.1177/0741088316631527

37. Ferrer i Cancho, R., Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1482), 2261–2265. https://doi.org/10.1098/rspb.2001.1800

38. Mihalcea, R., Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP 2004* (pp. 404–411). Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/W04-3252/

39. Erkan, G., Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479. https://doi.org/10.1613/JAIR.1523

40. Mota, N. B., Copelli, M., Ribeiro, S. (2012). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis. *PLoS ONE*, *7*(5), e34438. https://doi.org/10.1371/journal.pone.0034438

41. Bubela, T., Nisbet, C., Borchelt, S., Brunger, F., Critchley, E., Einsiedel, E., Caulfield, T. (2009). Science communication reconsidered. *Nature Biotechnology*, *27*(6), 514–518. https://doi.org/10.1038/nbt0609-514

42. Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., Baram-Tsabari, A. (2017). Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLOS ONE*, *12*(8), e0181742. https://doi.org/10.1371/journal.pone.0181742

43. Kang, M., Jin, T., Lu, X., Zhang, H. (2024). Exploring the differences in syntactic complexity between lay summaries and abstracts: A case study of *The New England Journal of Medicine*. *Journal of English for Academic Purposes*, *72*, 101444. https://doi.org/10.1016/j.jeap.2024.101444

44. Masucci, A. P., Rodgers, G. J. (2006). Network properties of written human language. *Physical Review E*, *74*(2), 026102. https://doi.org/10.1103/PhysRevE.74.026102

45. Amancio, D. R. (2015). Authorship recognition via fluctuation analysis of network topology and word intermittency. *Journal of Statistical Mechanics: Theory and Experiment*, *2015*(03), P03005. https://doi.org/10.1088/1742-5468/2015/03/P03005

46. Myers, G. (1990). *Writing biology: Texts in the social construction of scientific knowledge*. Boulder, CO: The WAC Clearinghouse. https://wac.colostate.edu/books/landmarks/myers/

Appendix

A1. Syntactic Complexity Indices

We used 14 syntactic complexity measures from Lu's (2010) L2 Syntactic Complexity Analyzer (L2SCA). These indices represent five categories: sentence length, T-unit length, subordination, coordination, and phrasal complexity.

Index	Name	Definition
MLS	Mean Length of Sentence	Average number of words per sentence
MLT	Mean Length of T-unit	Average number of words per T-unit
MLC	Mean Length of Clause	Average number of words per clause
C/S	Clauses per Sentence	Total number of clauses divided by sentences
VP/T	Verb Phrases per T-unit	Total verb phrases divided by T-units
C/T	Clauses per T-unit	Total number of clauses divided by T-units
DC/C	Dependent Clauses per Clause	Proportion of clauses that are dependent
DC/T	Dependent Clauses per T-unit	Total dependent clauses divided by T-units
T/S	T-units per Sentence	Total T-units divided by sentences
CT/T	Complex T-units per T-unit	Proportion of T-units containing at least one dependent clause
CP/T	Coordinate Phrases per T-unit	Total coordinate phrases divided by T-units
CP/C	Coordinate Phrases per Clause	Total coordinate phrases divided by clauses
CN/T	Complex Nominals per T-unit	Total complex nominals divided by T-units
CN/C	Complex Nominals per Clause	Total complex nominals divided by clauses

A2. Code and data access

All texts, preprocessing scripts, and textual analysis code are available at: https://github.com/bayesianMeow/Multidimensional-Analysis-Of-Linguistic-Differences-Sinha-Ranjan