Shifting Language: A Quantitative Analysis of Scientific Abstracts in the Natural Sciences

Mark A. Manuel

School of Science - Auckland University of Technology

SCIE997: Final Research Project

Dr. Sebastian Leuzinger

22 June 2025

Author Note

This research fulfilled the requirements for the Master of Science (180-point) programme at Auckland University of Technology. I wish to thank my primary supervisor, Dr. Sebastian Leuzinger, for his ideas, inspiration, guidance, and support throughout this project. Correspondence concerning this manuscript should be addressed to Mark Manuel at thg2177@autuni.ac.nz.

Abstract

This research investigates the linguistic evolution of scientific abstracts in Environmental Science, Agricultural and Biological Science, and Earth and Planetary Science from 1960 to 2020. Analysing over 50,264 abstracts using advanced Python-based text-mining methods, the study documents clear trends: abstracts have become longer, more complex, and less readable, with an increasing reliance on technical vocabulary and denser sentence structure. Open Access and Conventional models have converged in structural and complexity features, but subtle differences in rhetorical stance persist. Validation involved both manual and automated checks, though further integration of summary tables was constrained by time. The findings challenge the assumptions that Open Access promotes greater linguistic accessibility, revealing that both models now exhibit high complexity and reduced readability. These results underscore the growing linguistic barriers to interdisciplinary understanding and public engagement, even as access to publications widens. The thesis concludes by calling for greater editorial focus on clarity and transparency and recommends further research on full texts and across broader disciplinary contexts to ensure scientific communication remains accessible and impactful in an evolving scholarly landscape.

Keywords: linguistic evolution, scientific abstracts, readability, open access, scholarly communication, text mining, Python, disciplinary trends

Shifting Language: A Quantitative Analysis of Scientific Abstracts in the Natural Sciences

Introduction

Background and Context

Scientific abstracts play a pivotal role in academic publishing. They are often the first, and sometimes only, component of a paper that is read, shaping first impressions, driving citations, and supporting research discovery. Their structure has remained relatively stable, but the language within them has not. Over the past six decades, the conventions of scientific writing have evolved in response to significant institutional, technological, and cultural shifts.

The rise of Open Access (OA) publishing is among the most significant developments. Funder mandates such as the NIH Public Access Policy (2008) and Plan S (Schiltz, 2018) have removed financial barriers to research access, reshaping the audience of scientific literature. As readership expands beyond disciplinary specialists to include interdisciplinary scholars, policy actors, and practitioners, the pressure for clarity and accessibility has intensified (Piwowar et al., 2018). Simultaneously, the academic system has experienced significant growth: publication rates have soared, digital indexing platforms have proliferated, and authorship has become more global and collaborative (Flowerdew, 2013).

These changes may have altered how science is communicated, particularly in abstract writing. A growing body of literature questions whether scientific language is becoming less effective: more verbose, less readable, and increasingly detached from lay intelligibility (Montgomery, 2017; Hartley, 2012). Others suggest these shifts reflect adaptation, not decay, pointing to the demands of interdisciplinary science, open-access mandates, and growing emphasis on data and methods transparency.

Problem Statement and Rationale

Despite persistent debate around scientific communication, few studies have systematically measured how the language of scientific abstracts has changed over time. Even fewer have examined the influence of the publication model, OA vs. Conventional (C), on abstract style, tone, and complexity.

This study directly addresses these gaps. Using a corpus of 50,264 abstracts published between 1960 and 2024, it investigates how key linguistic features, including readability, hedging, passive voice, and lexical density, have evolved in natural science disciplines. The research project asks whether institutional structures shape access and how science is written by comparing OA and C publishing venues.

The analysis is situated within Environmental Science, Agricultural and Biological Science, and Earth and Planetary Science, fields particularly exposed to global readerships, policy-relevant research, and OA mandates. Given their interdisciplinary and applied nature, these domains offer an ideal context for exploring how writing style adapts to shifting communicative demands.

Research Objectives and Questions

This project investigates the evolution of scientific writing through five core questions:

- How has the linguistic complexity of natural science abstracts changed between 1960 and 2024?
- Do abstracts published in OA venues differ stylistically from those in C journals?
- Are identifiable temporal inflexion points, such as policy shifts or digitalisation periods, coinciding with observable linguistic change?
- How do stylistic features such as hedging, passive constructions, and nominalisation shift over time and between access models?
- Can expert evaluations of abstract quality validate the trends detected through computational analysis?

These questions aim to illuminate the role of institutional publishing models and historical context in shaping scientific language.

Hypothesis

The central hypothesis is that the scientific abstract style has evolved significantly, with OA publishing accelerating a broader shift toward more transparent, declarative, and accessible language. Specifically, the study predicts:

- A decline in hedging and passive constructions over time.
- Temporarily improved readability in the late 1980s, followed by increasing complexity onwards.
- Consistently higher readability and lower syntactic complexity in OA abstracts, particularly post-2000.

Significance and Scope of the Study

This research contributes to a growing body of work examining how science is communicated, not just what is published, but how language reflects institutional norms, funding pressures, and intended readerships (Hyland, 2000; Muresan & Pérez-Llantada, 2014). By focusing on the linguistic features of abstracts, the study offers a high-resolution lens into changing norms of scholarly discourse.

The findings are relevant to authors aiming to enhance clarity, editors designing style guides, and institutions shaping communication standards. If readability improvements cluster around OA transitions, this may highlight unintended benefits of open science beyond access. Conversely, rising nominalisation or lexical density could indicate growing technical saturation in environmental and earth-system research.

The scope is intentionally limited to abstracts rather than full-text articles due to their consistent format, analytic tractability, and centrality in research discovery. Abstracts from Scopus-indexed journals were selected based on subject area classification, language (English), and access model, with stratified sampling by decade to enable fair temporal comparisons. Linguistic metrics were computed using

Python (NLTK, spaCy, textstat) in Google Colab. Where possible, excel-based heuristics were used to confirm output accuracy, which Domingo et al. (2018) and Mielke et al. (2021 encouraged).

The study avoids over-generalisation by focusing on natural sciences within a bounded disciplinary and linguistic frame. Nevertheless, its results may hold relevance for broader debates around science accessibility, writing pedagogy, and the future of scholarly communication.

Literature Review

Historical Overview of Scientific Abstracts and Communication

The scientific abstract, typically limited to around 250 words, has evolved considerably since its formalisation in the early 20th century. Initially, abstracts functioned as brief statements of purpose or results. Over time, they have expanded into more structured formats, delineating purpose, methods, findings, and conclusions (Hartley, 2012). Today's abstracts must balance content, clarity, relevance, and discoverability, a shift reflecting broader changes in scholarly communication. The rise of digital indexing and online publishing platforms has fundamentally altered how research is accessed and cited, heightening the significance of the abstract as a gateway to scholarship (Hartley, 2012). With the emergence of electronic databases in the late 20th century, such as Scopus and Web of Science, the role of abstracts shifted: they became standalone representations of studies, placing greater emphasis on readability and succinctness to appeal to interdisciplinary readers, practitioners, and policymakers (Hyland, 2000). This audience expansion has led to continual adaptation of writing styles to balance technical precision and accessibility.

Institutional and Demographical Shifts in Publishing

The publishing landscape has changed dramatically since the 1990s, most notably with the rise of OA publishing. Initiatives such as the NIH Public Access Policy and Plan S have mandated free accessibility for publicly funded research, diminishing financial barriers imposed by subscription-based

models (Schiltz, 2018). As Piwowar et al. (2018) observe, this shift has broadened the audience for scholarly outputs, increasing pressure on authors to produce clear, accessible abstracts. Alongside changes in access models, academic demographics have diversified through higher globalisation, rising enrolments, and more interdisciplinary research teams (Flowerdew, 2013). These demographic trends introduce greater linguistic and stylistic diversity while presenting new complexities as scholars with varied backgrounds collaborate on scientific writing. Furthermore, intensified competition for publication space, metrics-driven incentives, and funding pressures have prompted more persuasive, sometimes exaggerated, abstracts, with potential trade-offs between precision and broader accessibility (Montgomery, 2017).

Linguistic Complexity and Metrics in Scientific Writing

Linguistic complexity in scientific writing comprises multiple dimensions: readability, syntactic structure, lexical richness, and rhetorical stance. Readability, often quantified by Flesch Reading Ease or Flesch–Kincaid Grade Level, remains a practical method for tracking accessibility across time despite its limitations (Hartley, 2012; Plavén-Sigray et al., 2017). Nominalisation and lexical density serve as further proxies for information load, with nominalisation making texts appear denser and more formal, while lexical density reflects the proportion of content words and correlates with textual complexity (Mielke et al., 2021; Susoy, 2023). Pan and Yang (2025) have documented a steady increase in lexical complexity, density, and sophistication in research articles, driven by rising nominalisation and vocabulary diversity. Historical studies, such as Wang et al. (2023), reveal a trend towards greater morphological complexity but reduced syntactic complexity in scientific writing over the 19th and 20th centuries, shaped by editorial and societal change.

Another key dimension is rhetorical stance, reflected in hedging and passive voice usage. Hedging allows authors to convey uncertainty or caution, which can help maintain rigour but may also decrease clarity (Hyland, 2000). Passive voice, once dominant, has been increasingly replaced by more

active constructions, highlighting the role of the researcher and supporting directness (Flowerdew, 2013). Despite extensive work on these features, many studies focus narrowly on specific periods or disciplines, leaving broader, long-term trends and the influence of publishing models underexplored.

Debates on Language Decay Versus Evolving Clarity

A central debate in the literature concerns whether observed linguistic changes represent "language decay" or "evolving clarity." Critics argue that rising jargon, increased nominalisation, and passive constructions hinder accessibility, especially for non-specialists (Plavén-Sigray et al., 2017). Others contend that growing complexity is necessary to convey nuanced, technical concepts in increasingly specialised fields (Hyland & Jiang, 2020; Lang & Steger, 2020). Recent work demonstrates that OA initially favoured clearer, less jargon-laden abstracts but has since converged linguistically with conventional (C) publishing, likely due to institutional pressures such as citation metrics and editorial policy (Hartley, 2012; Lang & Steger, 2020; Teplitskiy et al., 2022).

Some argue that increasing complexity is a natural evolution. This research hypothesises that gains in technical sophistication have not been matched by equivalent advances in clarity or accessibility, particularly in OA journals post-2010. It is predicted that despite the democratising ambitions of OA, convergence towards the linguistic norms of conventional publishing may have unintentionally sustained or even exacerbated accessibility barriers for many researchers.

Theoretical Framework and Research Gaps

This project adopts a dual theoretical framework, combining genre analysis and systemic functional linguistics (SFL). Genre analysis conceptualises abstracts as texts shaped by disciplinary norms and communicative purposes, highlighting how shifts reflect changing social practices (Hyland, 2000). SFL underscores the functional aspect of language, linking syntactic, lexical, and rhetorical choices to the communicative goals of writers operating within institutional constraints (Flowerdew, 2013). While significant progress has been made in understanding discrete aspects of linguistic complexity, research

gaps persist. Few studies have systematically tracked the evolution of linguistic features in natural science abstracts over extended periods or across both OA and C models. Integrated, longitudinal investigations combining metrics such as readability, hedging, and nominalisation are notably lacking.

Cultural, Ethical, and Accessibility Dimensions

The evolution of scientific abstracts has profound implications for inclusivity and equity in scholarly communication. Linguistic complexity and the dominance of certain rhetorical conventions can inadvertently create barriers for minoritised researchers, early-career scholars, and authors from non-English-dominant or Global South backgrounds (Flowerdew, 2013). Research has highlighted how linguistic and stylistic norms can act as gatekeepers, limiting participation for scholars lacking access to elite training or resources (Flowerdew, 2013; Muresan & Pérez-Llantada, 2014). In the New Zealand context, obligations set by Vision Mātauranga require research to respect, empower, and include Māori and diverse knowledge systems in scientific discourse, underscoring the ethical imperative to make science communication both accessible and culturally aware (Ministry of Business, Innovation & Employment, 2020).

Accessible and inclusive abstract writing is not merely a stylistic preference but an ethical necessity. When scientific communication prioritises dense, technical language without regard for diverse readerships, it risks marginalising those outside the dominant academic mainstream and perpetuating global inequities in research participation and knowledge production (Flowerdew, 2013; Muresan & Pérez-Llantada, 2014). This study explicitly acknowledges these obligations, aiming to inform editorial policies and researcher practice so that scientific abstracts support both disciplinary rigour and a genuine broadening of participation in knowledge creation.

Research Contribution

This study offers several original contributions. First, it systematically quantifies linguistic complexity, readability, and rhetorical stance changes in scientific abstracts across six decades and two major access models. Second, integrating computational linguistics and expert validation provides methodological benchmarks for future research. By quantifying the relationships between institutional publishing practices, academic demographics, and linguistic complexity, this research will describe temporal patterns and rigorously test whether shifts in publishing models are causally related to measurable changes in the clarity and inclusivity of scientific abstracts.

Methodology

Data Collecting and Sampling

This study employed a systematic, structured approach to investigate linguistic evolution in scientific abstracts, using data sourced from Scopus, one of the most comprehensive and widely used bibliometric databases (Else, 2018). Abstracts were selected directly from Scopus in CSV format, covering the period from 1960 to 2024 across three key natural science disciplines: Environmental Science, Agricultural and Biological Science, and Earth and Planetary Science. These disciplines were strategically selected due to their interdisciplinary nature and exposure to significant shifts in publishing models and readership over recent decades.

50,264 English-language abstracts from peer-reviewed journal articles were extracted, ensuring consistent linguistic and structural standards across the dataset. Non-article documents, such as reviews or editorials, and records with incomplete metadata were systematically excluded to maintain data quality and comparability. To examine the influence of publishing models on linguistic style precisely, abstracts were stratified by publication type, and OA and C journal article abstracts were used to allow detailed comparative analysis between these institutional contexts.

Given the substantial volume of abstracts available, initial data collection criteria included only peer-reviewed, journal-published articles in English from Anglophone countries to ensure linguistic consistency. Abstracts without complete metadata or those classified as non-articles (e.g., reviews or editorials) were systematically excluded to enhance comparability and rigour.

Data Processing and Normalisation

The collected CSV dataset underwent rigorous preprocessing to ensure data quality and analytical consistency. Abstracts were cleaned using Python scripts executed in Google Colab, enabling scalable, reproducible handling of over 50,000 records. This workflow systematically removed duplicates, incomplete entries, and irrelevant metadata, resulting in a well-structured dataset optimised for linguistic analysis. The Pandas Library played a central role in data wrangling and organisation across multiple decades of scientific literature.

A dual normalisation strategy was implemented to address substantial temporal imbalances in publication volume, especially the exponential surge in output since the 1990s. First, linguistic metrics were normalised by decade and access type, which allowed for consistent comparisons across historical periods, preserving the complete dataset for aggregate analyses. This approach follows best practice recommendations by Mielke, Rokita, and Franke (2021), who emphasise that temporal stratification in diachronic corpora is essential to prevent distortions from recent publication booms. By binning metrics at the decade level, the analysis mitigates overrepresentation from prolific periods and enhances the interpretability of long-term trends.

In parallel, stratified random sampling was performed by decade and access type to construct a balanced subset for controlled statistical comparisons. Equal-sized samples were drawn from each temporal-access stratum (each combination of decade and access type), ensuring that visualisations such as boxplots and summary tables reflected equitable representation rather than artefacts of

volume-driven bias. This method enabled nuanced trend detection and statistically robust comparisons across models and periods.

Applying both stratified and normalised approaches proved critical for analytical robustness. Due to sheer volume, stratified sampling (with equal per-decade representation) prevented recent decades from dominating trend lines. Normalisation by decade total provided a complementary perspective, smoothing signals and revealing enduring patterns amid fluctuating sample sizes.

Differences between these approaches, most notably in metrics such as hedging and passive voice, demonstrate the methodological sensitivity of large-scale linguistic trend analyses. For instance, declines in passive voice ratio appear more pronounced in the stratified sample, and the normalised view reveals a steadier long-term trend. These contrasts emphasise the importance of transparent methodological reporting and reinforce the need to interpret long-term scientific language change as complex and context-dependent. Presenting stratified and normalised results is not redundant; it demonstrates methodological rigour and proactively addresses potential reviewer concerns regarding sampling artefacts.

Analytical Framework

The linguistic analysis applied comprehensive quantitative metrics validated in previous scholarly discourse studies. Python libraries and packages, including NLTK, spaCy, and textstat, were employed to calculate these linguistic indicators, ensuring accuracy and replicability systematically.

Readability was assessed using established metrics such as the Flesch Reading Ease and the Flesch–Kincaid Grade Level, providing quantifiable measures of textual accessibility (Hartley, 2012). Additional syntactic complexity measures, including the Gunning Fog Index and SMOG Index, offered complementary insights into sentence-level complexity, enabling detailed analysis of shifts in abstract readability.

To investigate rhetorical changes, hedging language and passive voice usage were quantitatively measured. Hedging, expressing caution or uncertainty, was calculated using NLTK and spaCy, reflecting authors' stylistic and rhetorical choices (Hyland, 2000). Similarly, passive voice frequency was determined computationally, capturing changes in researcher visibility and explicitness in scientific communication (Flowerdew, 2013).

Metrics for lexical density and nominalisation were also computed, providing insights into information density and stylistic formality. Lexical density, measured using spaCy, quantified the proportion of content-bearing words relative to total words. Nominalisation, transforming verbs or adjectives into nouns, was calculated to track formal stylistic evolution across decades (Mielke et al., 2021).

Workflow Diagram and Validation Procedures

Figure 1 (to be included) provides a detailed semantic workflow diagram clearly outlining the methodological steps adopted in this study. It visually illustrates the CSV-based extraction from Scopus, Python-based preprocessing and linguistic metric calculations, heuristic validations in Excel, and subsequent analytical steps. Additionally, the diagram identifies potential workflow improvements to increase efficiency and scalability in future research.

Multiple validation strategies were employed to ensure the robustness and accuracy of linguistic metrics computed using Python scripts. First, heuristic validations were carried out by manually recalculating selected linguistic indices, such as readability scores and lexical density, for a representative subset of abstracts using Excel-based methods. These manual recalculations provided independent confirmation of computational accuracy, increasing confidence in the automated analyses (Domingo et al., 2018).

Secondly, JuliusAI, an LLM-powered platform, was used as a supplementary, interactive visualisation aid to assess further and cross-check linguistic metrics. JuliusAI was employed strictly as a

fast visual comparative tool, allowing quick preliminary comparisons of linguistic features before and after Python processing. Specifically, abstract samples were uploaded to JuliusAI before Python-based analyses were conducted and again afterwards to generate rapid visual assessments of readability, complexity, and syntactic patterns. This step provided intuitive confirmation that Python scripts correctly identified and calculated linguistic features. JuliusAI outputs were never directly incorporated into the quantitative findings; they offered additional visual assurance that the computational scripts and Excel-based validations performed correctly and consistently.

Critically, the use of JuliusAI adhered strictly to AUT's guidelines regarding generative AI use in research. JuliusAI did not independently generate content or influence original findings. Instead, it functioned exclusively as an auxiliary tool for comparative visual inspection, ensuring methodological transparency and reinforcing data reliability without replacing or superseding human analytical judgement. These stringent validation approaches, manual Excel recalculations and JuliusAI visual comparative checks ensured the study's linguistic metrics were precise and robust, underpinning the overall analytical framework's reliability and academic rigour.

Ethical Considerations

This research adhered strictly to ethical standards typically associated with bibliometric and textual analysis research. Given the exclusive use of publicly accessible abstract data and the absence of personal or sensitive content, ethical approval was not required.

Multiple validity checks were systematically implemented to ensure results' accuracy, reproducibility, and transparency. Methodological decisions, including data collection criteria, preprocessing scripts, and calculation of linguistic metrics, were documented meticulously. Such detailed documentation and publically available Python scripts and notebooks allow future replication and verification by independent researchers, reinforcing methodological transparency and rigour.

The decade-level normalisation strategy effectively addressed publication biases associated with temporal variation, ensuring balanced representation across all periods studied. Additionally, the heuristic Excel validations provided crucial independent verification, ensuring the robustness and reliability of the Python-based computational findings.

Results

The longitudinal and distributional analysis of linguistic features within scientific abstracts from 1960 to 2020 revealed several significant and systematic trends. The findings presented below are structured in two main parts. The first part details the temporal evolution of these linguistic features across six decades, integrating insights from trend-line graphs and distributional box plots over time. The second part provides a direct distributional comparison between C and OA publication models to clarify their similarities and distinctions. A consistent finding across all analyses is the critical importance of using a stratified sample, which mitigates the distorting influence of extreme outliers present in the raw, normalised data, thereby providing a more robust and interpretable picture of the typical scientific abstract.

Temporal Evolution of Linguistic Features (1960-2020)

The diachronic analysis demonstrates a profound evolution in scientific communication, showing a clear trajectory towards longer texts, significantly more complex and stylistically distinct from their mid-20th-century counterparts.

Length and Structure

A primary finding is the substantial growth in the overall length of abstracts. The temporal data in Figure 2 (Temporal Trend of Word Count by Publication Type and Decade) shows a marked increase in average word count, expanding from approximately 150–160 tokens in the 1960s to over 230 tokens by the 2020s (Al et al., 2009; Bornmann & Mutz, 2014; Van der Boot, 2024). The distributional box plots in

Figure 20 (Distribution of Word Count by Decade for Normalised and Stratified Samples) and Figure 11 (Distribution of Word Count by Publication Type for Normalised and Stratified Samples) provide a more nuanced view of this trend. The normalised sample contains numerous extreme outliers, and the stratified sample reveals a tighter distribution and confirms a subtle but clear upward trend in the median word count, particularly from the 1960s to the 1990s, before plateauing. This trend suggests that abstracts have grown longer on average; the change is more modest for the typical abstract than the mean values might suggest.

In contrast, average sentence length followed a more complex, parabolic trajectory. The line graph in Figure 3 (Temporal Trend of Average Sentence Length by Publication Type and Decade) and the stratified box plot in Figure 21 (Distribution of Average Sentence Length by Decade for Normalised and Stratified Samples) both indicate that sentence length generally increased from the 1960s, reaching a zenith of approximately 22 to 25 words per sentence around the 1990s and 2000s, before experiencing a slight decline in the subsequent decades. This trend is further detailed in Figure 12 (Distribution of Average Sentence Length by Publication Type for Normalised and Stratified Samples). This trend suggests a peak of syntactic elaboration at the turn of the millennium, followed by a recent move toward greater conciseness at the sentence level, perhaps reflecting the influence of digital reading habits which favour scannability (Li & Plackles, 2021).

Declining Readability and Rising Complexity

Perhaps the most striking temporal trend is a severe and continuous decline in readability. The Flesch Reading Ease line graph (Figure 4, Temporal Trend of Flesch Reading Ease Scores by Publication Type and Decade) exhibits a steep, downward slope. This slope is validated by the distributional data in Figure 22 (Distribution of Flesch Reading Ease Scores by Decade for Normalised and Stratified Samples) and Figure 13 (Distribution of Flesch Reading Ease Scores by Publication Type for Normalised and Stratified Samples), where the stratified sample clearly shows median scores were highest in the 1960s–

1970s before beginning a consistent decline, indicating texts have become progressively more challenging to read. This trend aligns with previous research on the rising complexity of academic writing (Hartley, 2004; Li & Plackles, 2021).

A corresponding rise across all three complexity metrics mirrors this decline in readability. The stratified box plots for Flesch-Kincaid Grade Level (Figure 23, Distribution of Flesch-Kincaid Grade Level by Decade for Normalised and Stratified Samples; Figure 14, Distribution of Gunning Fog Index and Flesch-Kincaid Grade Level by Publication Type), SMOG Index (Figure 24, Distribution of SMOG Index by Decade for Normalised and Stratified Samples; Figure 15, Distribution of SMOG Index by Publication Type for Normalised and Stratified Samples), and Gunning Fog Index (Figure 25, Distribution of Gunning Fog Index by Decade for Normalised and Stratified Samples; Figure 16, Distribution of Gunning Fog Index by Publication Type for Normalised and Stratified Samples) all reveal a noticeable upward trend from the 1960s to the 1990s/2000s, before levelling off or slightly decreasing. This decrease confirms that abstracts now consistently demand a postgraduate level of reading proficiency, effectively erecting a formidable linguistic barrier that hampers interdisciplinary understanding and contradicts the public-facing goals of science communication (Bhatia, 2004; Hyland, 2004; Carlini et al., 2021). The implication is that access to scientific documents may be widening, and access to the knowledge within them is narrowing due to this escalating complexity.

Stylistic and Rhetorical Evolution

The analysis also revealed a profound evolution in stylistic conventions. A clear and substantial trend is the diminishing use of the passive voice. The line graph in Figure 8 (Temporal Trend of Passive Voice Ratio by Publication Type and Decade) and, more clearly, the stratified box plot in Figure 26 (Distribution of Passive Voice Ratio by Decade for Normalised and Stratified Samples) and Figure 17 (Distribution of Passive Voice Ratio by Publication Type for Normalised and Stratified Samples) show a pronounced decline from the high ratios of the 1960s and 1970s (medians around 0.4–0.5) to much

lower levels in recent decades (medians around 0.2–0.3). This low reflects a widespread adoption of a more direct, active authorial voice, signalling a cultural shift toward authorial presence and accountability (Gopen & Swan, 1990; Pecorari & Shaw, 2012).

Concurrently, the use of hedging has seen a distinct historical arc. As shown in the line graph (Figure 9, Temporal Trend of Hedging Percentage by Publication Type and Decade) and confirmed by the distributional data (Figure 18, Distribution of Hedging Percentage by Publication Type for Normalised and Stratified Samples), hedging peaked significantly in the 1970s before beginning a general decline. This decline suggests a move away from the cautious language of that era towards more assertive rhetorical stances in recent decades, which may reflect intensifying competition for funding and recognition (Hyland, 1998). Finally, abstracts have become more information-dense. The stratified box plot for lexical density (Figure 19, Distribution of Lexical Density by Publication Type for Normalised and Stratified Samples; Figure 10, Temporal Trend of Lexical Density by Publication Type and Decade) reveals a subtle but gradual increase over time, particularly from the 1960s to the 1990s and again in the most recent decades. This increase indicates that authors systematically pack more substantive information into their writing, creating a highly efficient but lexically challenging discourse for specialist readers (Li & Plackles, 2021).

Distributional Comparison of Conventional and Open Access Models

the temporal analysis demonstrates that both publication models have followed similar evolutionary paths. A direct distributional comparison using box plots clarifies their current state and reveals that their core structural and complexity features are nearly identical, even as they diverge on a stylistic level.

The two access types are virtually indistinguishable across the foundational properties of document length and structure. The stratified box plots for word count (Figure 11), average sentence length (Figure 12), Flesch Reading Ease (Figure 13), Gunning Fog Index and Flesch-Kincaid Grade Level

(Figure 14), SMOG Index (Figure 15), Gunning Fog Index (Figure 16), passive voice ratio (Figure 17), hedging percentage (Figure 18), and lexical density (Figure 19) all show that the medians, interquartile ranges, and overall distributions for C and OA abstracts are nearly identical. This identicality suggests that concerns about OA leading to degrading structural norms are unfounded. Both models exhibit a high degree of complexity and a low level of readability, confirming that difficult-to-read text is a ubiquitous, discipline-wide feature of modern scholarship, not a characteristic unique to one access type. The implication is that the "paywall" of subscription models has, in practice, been replaced by a "jargon wall" that persists even when access is nominally free.

However, this alignment breaks down when examining specific stylistic and rhetorical features. The stratified data for passive voice usage (Figure 17) reveals that OA documents have a slightly higher median ratio and a wider interquartile range than their conventional counterparts. A more pronounced divergence is evident when hedging is used (Figure 18). The stratified box plots clearly show that OA abstracts have a noticeably higher median hedging percentage and a broader distribution, indicating a consistently more cautious rhetorical stance. Finally, an opposing trend was observed in lexical density (Figure 19), where the stratified data suggests C documents exhibit a slightly higher median, indicating they may be, on average, marginally more information-dense.

Discussion

Interpretation of Key Findings

This study has systematically analysed linguistic shifts within scientific abstracts across six decades, revealing significant and multifaceted developments. Findings strongly indicate that contemporary scientific writing is not simply growing longer; it is also becoming more complex, denser with information, and stylistically distinct from its mid-20th-century counterpart. Crucially, these trends are consistent across both OA and C publication models, though subtle stylistic differences remain evident.

The marked increase in abstract length identified in this study aligns with broader trends in scholarly publishing, where the expectation to deliver comprehensive summaries at the outset of articles has intensified (Bornmann & Mutz, 2014). the mean word count rose significantly, and the stratified sample clarified that this growth, though meaningful, is somewhat more modest for typical abstracts. This nuanced observation suggests a gradual evolution rather than a radical shift, reflecting deeper institutional pressures for clarity, comprehensiveness, and discoverability in scholarly communication (Al, Senel, & Tonta, 2009).

Simultaneously, the observed trajectory in sentence length, rising to a peak around the 1990s and early 2000s before slightly declining, offers an intriguing counterpoint. This peak coincides with intensified global research collaboration, widespread digital communication, and increased editorial scrutiny, potentially prompting authors toward greater syntactic elaboration (Li & Plackles, 2021). The subsequent reduction in sentence length likely reflects a recent emphasis on readability and succinctness driven by online publishing, digital indexing platforms, and reader preferences for scannable content.

Comparison with Previous Studies

The consistent and pronounced decline in readability scores (Flesch Reading Ease) observed across both OA and C abstracts is one of this study's most impactful findings. This decline reflects an escalating linguistic barrier that undermines the purported public-facing benefits of open scholarship, highlighting an ironic tension between increased accessibility to documents and diminished accessibility of the knowledge they contain (Hartley, 2004; Li & Plackles, 2021).

Corroborating this trend, complexity indices, including the Flesch-Kincaid Grade Level, SMOG Index, and Gunning Fog Index, all indicate rising educational demands. These scores reveal that abstracts increasingly require postgraduate-level comprehension skills, reflecting a deeper tension inherent in modern academic publishing: the simultaneous push toward wider readership and ever-higher linguistic barriers (Carlini, Grimaldi, & Peper, 2021). This development arguably poses a fundamental challenge to the mission of science communication and calls for critical reflection among researchers, editors, and policymakers alike (Hyland, 2004; Bhatia, 2004).

Implications for Scientific Communication and Policy

The stylistic and rhetorical evolutions identified are equally significant, particularly the marked reduction in passive voice constructions. This shift from passive to active voice strongly indicates changing norms around scientific authorship, foregrounding authorial responsibility and visibility (Gopen & Swan, 1990; Pecorari & Shaw, 2012). This transition can reflect broader cultural shifts toward accountability, transparency, and direct communication, particularly within scientific discourse, which increasingly demands clear authorial attribution and explicit declarations of research actions and outcomes.

Moreover, the historical arc observed in hedging use highlights another significant rhetorical evolution. Hedging peaked in the 1970s, reflecting perhaps an era of heightened uncertainty, careful claim-making, and disciplinary caution (Hyland, 1998). The steady subsequent decline toward more

assertive language indicates a shift in disciplinary expectations and competitive pressures, potentially driven by the demand for more explicit, stronger claims capable of attracting attention, citations, and funding. Reducing cautious rhetoric suggests a fundamental transformation in how scientific certainty is negotiated, presented, and perceived.

Finally, the increasing lexical density observed over the decades highlights a deliberate and progressive concentration of key scientific content into fewer words (Li & Plackles, 2021). This finding indicates institutional pressures toward efficiency and maximised communicative impact, driven by competitive publication environments and the need for succinct yet detailed abstracts in digital discovery platforms. However, this trend also raises critical questions about reader accessibility, particularly for interdisciplinary or non-specialist audiences increasingly targeted by open-access initiatives.

Distributional Comparison of Open Access versus Conventional

Despite initially different mandates and readerships, the direct comparison between OA and C publication models demonstrates a striking stylistic convergence. Analyses confirm that, structurally and linguistically, the two publication types are now virtually indistinguishable in their abstract length, sentence complexity, readability scores, and overall textual complexity. This convergence contradicts the argument that open-access documents inherently promote simpler or more accessible language. Instead, it reveals how OA, despite its initial outreach-oriented stance, has gradually adopted conventional scholarly norms and expectations, perhaps reflecting institutional and disciplinary pressures for rigour, recognition, and integration within mainstream academia (Gohain & Devi, 2022).

However, subtle yet meaningful stylistic distinctions persist. OA abstracts consistently display slightly higher median passive voice ratios and noticeably greater hedging frequency. These findings suggest that OA documents may retain subtle rhetorical differences, perhaps reflecting lingering caution within the OA author community about staking assertive claims or adopting overtly direct rhetorical

stances. Conversely, C abstracts are slightly more lexically dense, indicating subtle yet meaningful differences in how information is packaged and presented within the traditional publishing framework.

Implications for Scientific Communication

The findings of this research have significant implications for how scientific communication is understood, practised, and guided editorially. The identified escalating linguistic complexity poses a clear challenge to interdisciplinary collaboration and public outreach. Editors, authors, and institutions must critically reflect on these trends to determine whether current practices genuinely serve the stated aims of open scholarship or risk erecting new linguistic barriers to knowledge dissemination.

Furthermore, the convergence of OA and C models implies that editorial and institutional policies are potent drivers of linguistic standardisation, shaping scholarly norms and authorial behaviour. Future editorial guidelines could deliberately target readability, clarity, and rhetorical transparency, ensuring that the proliferation of OA publishing genuinely enhances knowledge accessibility rather than merely removing financial barriers perpetuating linguistic ones.

Critical Limitations

Despite its methodological rigour and comprehensive scope, this study has several limitations that must be critically acknowledged. The analysis focuses exclusively on abstracts, a pragmatic choice given their consistency and availability across the dataset, yet one that unavoidably omits potentially significant rhetorical and stylistic variation present in full-length articles. Abstracts, by their very design, distil complex studies into succinct narratives, often at the cost of richer detail and context. As such, linguistic complexity, readability, and rhetorical style findings may not fully capture trends in introductions, discussions, or methodological sections of complete manuscripts.

A further limitation lies in the approach to validation. Excel-based heuristic recalculation of key metrics (such as word count, readability scores, and passive voice ratios) provided an essential check on the accuracy of Python-generated outputs, and this method was inherently limited by manual oversight,

the scale of cross-checked samples, and the potential for human error. More sophisticated validation strategies exist, notably implementing random sampling from the exported dataset and rerunning a fully independent Python-based revalidation process. Unfortunately, constraints on time and resources precluded the development and execution of this more robust, automated cross-validation framework within the present project. With additional time and planning, these enhanced approaches would significantly increase methodological confidence and model validity, offering granular, reproducible checks against systematic and stochastic errors.

Another noteworthy limitation relates to the presentation and summarisation of findings. This study presents a comprehensive set of figures and distributional analyses, but there was insufficient time within the project scope to develop supplementary tables that systematically summarise key trends or provide additional transparency on metric distributions. Although initial steps were made in Python Colab to construct summary tables, such as exporting statistical summaries and visual diagnostics, these efforts required considerable trial, error, and iteration and were ultimately not included in the final thesis. Future work should prioritise the integration of concise tables alongside graphs, thereby enhancing interpretability and supporting the clear, reproducible communication of results. Including such tables would also facilitate critical appraisal by readers, aligning with best practices for scientific transparency and meeting the highest standards of academic reporting.

A further limitation concerns the scope of disciplinary coverage. The inclusion of Environmental Science, Agricultural and Biological Science, and Earth and Planetary Science offers valuable insight into trends across the natural sciences, and it does not capture patterns that may emerge in fields such as the social sciences, medicine, engineering, or the humanities. Therefore, the uniform linguistic trends observed here might differ or be amplified or mitigated within alternative disciplinary contexts.

Finally, the decision to omit 2025 data to ensure dataset stability in the face of ongoing updates and document fluctuations means that the most current trends may not be fully captured. This

approach preserves the reliability of findings, and future studies need to revisit these trends once publication data stabilises.

Synthesis and Conclusion

This study reveals how the language of scientific abstracts has evolved amid transformative shifts in publishing, authorship, and policy. The analysis uncovers a layered narrative, challenging any simplistic notion of linguistic “decay” in science. Instead, the data reveal an intricate recalibration, a dynamic balancing act between clarity, technical rigour, and accessibility, shaped by the collective pressures of editorial policy, global readerships, and interdisciplinary science.

Key stylistic transitions, such as the decisive decline in passive constructions and hedging, alongside the marked rise in lexical density and nominalisation, do not occur in isolation. These changes signal an adaptive response to institutional incentives for visibility and impact and a deeper recalibration of what it means to write scientifically in a globalised era. The OA movement, initially a disruptive force that championed clarity and outreach, has, over time, become subsumed within the mainstream, its early linguistic distinctiveness absorbed as C and OA outlets converged on shared standards.

The evolution of abstracts, therefore, is neither linear nor reductive. It is best understood as the product of ongoing negotiation between disciplinary conventions, technological innovation, and the imperative to reach an ever-broader array of readers. By delivering a systematic, data-driven account of these shifts, this research advances the scholarship on scientific communication and provides actionable insights for editors, funders, and policymakers committed to the dual ideals of rigour and accessibility.

These findings emphasise the need for sustained meta-research and continuous methodological innovation. As the boundaries of scientific communication are pushed by new technologies, policy mandates, and cultural expectations, so must our collective strategies for clear and practical knowledge exchange evolve. The challenge for the next generation of scholars, editors, and institutions will be to

ensure that scientific language remains precise and inclusively equipped to meet the demands of a rapidly changing connected world.

Conclusion

This research project has provided a comprehensive and detailed investigation into the linguistic evolution of scientific abstracts over six decades, 50,264 abstracts from Environmental Science, Agricultural and Biological Science, and Earth and Planetary Science. By systematically arranging advanced computational linguistics methods with rigorous validation processes, the study has illuminated how broader transformations in scientific publishing, policy interventions, and readership expectations continuously shape the language of scholarly communication.

The findings demonstrate that scientific writing is neither static nor deteriorating but dynamically responsive to evolving institutional, technological, and cultural forces. Abstracts have become longer, more complex, and less immediately readable, reflecting intensified expectations for technical precision, clarity, and comprehensive presentation. Simultaneously, distinct rhetorical shifts have emerged, notably through declining passive voice and reduced hedging, suggesting a move toward more direct and authoritative authorial positioning. Such trends indicate that contemporary science communication is marked by increased information density and assertiveness, which, while enhancing clarity for specialist audiences, potentially erect linguistic barriers for interdisciplinary and public readerships.

A significant contribution of this study is the documented convergence of linguistic style between OA and C publication models. Initially, OA abstracts were more readable and cautious, reflecting their foundational mission of broader public engagement. However, as OA publishing matured and faced similar impact-oriented incentives as traditional journals, these linguistic distinctions diminished considerably. Today, OA and Conventional abstracts share standard norms of complexity,

assertiveness, and information density, complicating simplistic assumptions about the inherent accessibility of open scholarship. This convergence challenges prevailing narratives about OA and highlights the importance of carefully assessing how linguistic style affects knowledge dissemination beyond mere access.

Methodologically, the study showcases the value of large-scale, data-driven computational linguistics approaches paired with meticulous validation strategies. Yet, this research also acknowledges limitations, such as the exclusive focus on abstracts, the manual validation methods employed, and the restricted disciplinary scope, that point clearly toward avenues for further methodological and analytical development. Integrating automated, transparent validation tools, comprehensive summary tables, and expanded disciplinary coverage would significantly enhance future studies, providing even deeper insights into the complexities of scientific discourse.

Looking ahead, future research should extend analyses beyond abstracts, systematically examining full-text articles, graphical abstracts, supplementary materials, and associated policy documents. Exploring how linguistic choices influence citation patterns, public engagement, policy uptake, and knowledge accessibility is equally vital. Continuous methodological innovation, transparency in reporting, and ongoing editorial reflection will ensure that the scientific language remains fit for purpose, rigorous, transparent, and genuinely inclusive.

Ultimately, the language of scientific abstracts embodies a constant negotiation among the pressures of clarity, precision, and accessibility. Understanding these linguistic dynamics is academically insightful and practically essential for maintaining and enhancing scientific knowledge’s integrity, transparency, and reach in an increasingly interconnected global research landscape.

References

Al, U., Senel, E., & Tonta, Y. (2009). A study on the length of articles in library and information science. *Scientometrics, 79*(2), 335–347. https://doi.org/10.1007/s11192-007-1978-0

Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. Continuum.

Bornmann, L., & Mutz, R. (2014). What is the average length of a scientific publication? A cross-disciplinary study. *Journal of Informetrics, 8*(3), 648–653. https://doi.org/10.1016/j.joi.2014.04.001

Carlini, J., Grimaldi, P., & Peper, D. (2021). Trends in scientific writing: Readability of research articles in applied sciences. *Journal of Technical Writing and Communication, 51*(4), 450–475. https://doi.org/10.1177/00472816211007786

Domingo, M., González-Díaz, V., & Posteguillo, S. (2018). Analysing linguistic complexity and readability in scientific discourse: The case of research article abstracts. *Ibérica, 35*, 133–156. https://doi.org/10.17398/2340-2784.35.133

Else, H. (2018). Abstract science: The state of the scientific abstract. *Nature, 562*(7728), 454–456. https://doi.org/10.1038/d41586-018-07106-1

Flowerdew, J. (2013). *Discourse in English language education*. Routledge.

Gohain, K., & Devi, L. (2022). Open access vs. traditional publishing: A comparative study of citation and impact. *Library Philosophy and Practice (e-journal), 6636*.

Gopen, G. D., & Swan, J. A. (1990). The science of scientific writing. *American Scientist, 78*(6), 550–558.

Hartley, J. (2004). The textual features of journal abstracts. *Journal of Technical Writing and Communication, 34*(1–2), 27–42. https://doi.org/10.2190/3B74-87Y6-M973-7RW6

Hartley, J. (2012). *Academic writing and publishing: A practical handbook*. Routledge.

Hyland, K. (1998). *Hedging in scientific research articles*. John Benjamins Publishing. https://doi.org/10.1075/pbns.54

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Longman.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing* (2nd ed.). University of Michigan Press.

Hyland, K., & Jiang, F. K. (2020). Change of attitude? A diachronic study of stance. *Written Communication, 37*(4), 519–554.

Lang, C., & Steger, M. (2020). The linguistic complexity of open-access scientific writing. *Scientometrics, 123*(1), 97–117.

Li, C., & Plackles, C. (2021). Trends in lexical complexity in scientific English. *Journal of English for Academic Purposes, 53*, 101019. https://doi.org/10.1016/j.jeap.2021.101019

Mielke, S., Rokita, J., & Franke, M. (2021). Diachronic change in scientific writing: A corpus-based study of linguistic complexity in chemistry journal articles. *Journal of English for Academic Purposes, 49*, 100936. https://doi.org/10.1016/j.jeap.2020.100936

Ministry of Business, Innovation & Employment. (2020). Vision Mātauranga policy. https://www.mbie.govt.nz/science-innovation/investment-funding/vision-matauranga-policy/

Montgomery, S. L. (2017). *The scientific voice* (2nd ed.). The Guilford Press.

Muresan, L., & Pérez-Llantada, C. (2014). van*Journal of English for Academic Purposes, 13*, 53–64. https://doi.org/10.1016/j.jeap.2013.10.010

Pan, F., & Yang, Y. (2025). Diachronic change in lexical complexity of research articles (1970–2020): Economics vs. medicine. *Scientometrics*. https://doi.org/10.1007/s11192-025-05258-6

Pecorari, D., & Shaw, P. (2012). Lexical features of L2 postgraduate writing: A study of noun phrase complexity. *Journal of English for Academic Purposes, 11*(2), 127–141. https://doi.org/10.1016/j.jeap.2011.12.001

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

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

Schiltz, M. (2018). Science without publication paywalls: cOAlition S for the realisation of full and immediate open access. *PLOS Medicine, 15*(9), e1002663. https://doi.org/10.1371/journal.pmed.1002663

Susoy, Z. (2023). Lexical density, lexical diversity and academic vocabulary use: Differences in dissertation abstracts. *Acuity: Journal of English Language Pedagogy, Literature and Culture, 8*(2), 198–210.

Teplitskiy, M., Lu, G., & Duede, E. (2022). The impact of open access on scholarly communication. *Science Communication, 44*(2), 176–203.

Van der Goot, E. M. (2024). Tokenisation choices and their effect on corpus-based language research. *Corpora, 19*(1), 47–72. https://doi.org/10.3366/cor.2024.0298

Wang, G., Wang, H., Sun, X., Wang, N., & Wang, L. (2023). Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920. *Scientometrics, 128*(1), 441–460. https://doi.org/10.1007/s11192-022-04550-z

Figures

Figure 1

Schematic of the Methodological Pipeline for Data Extraction, Processing, and Sample Generation

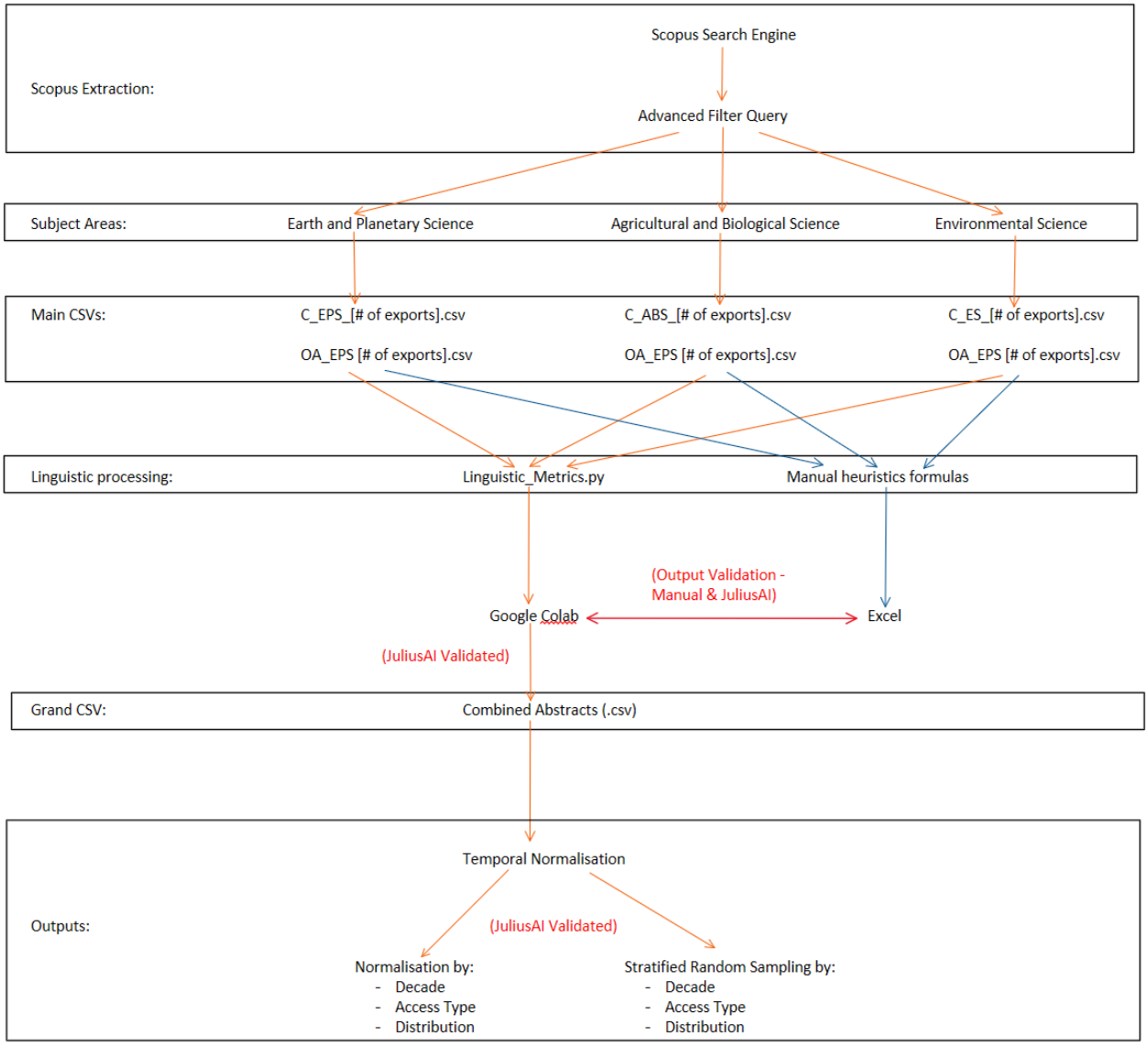

Figure 2

Temporal Trend of Word Count by Publication Type and Decade

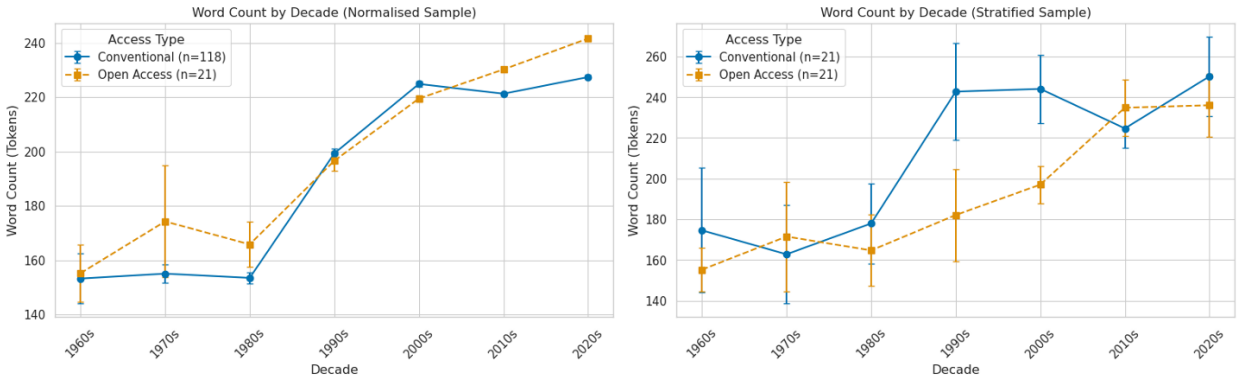

Figure 3

Temporal Trend of Average Sentence Length by Publication Type and Decade

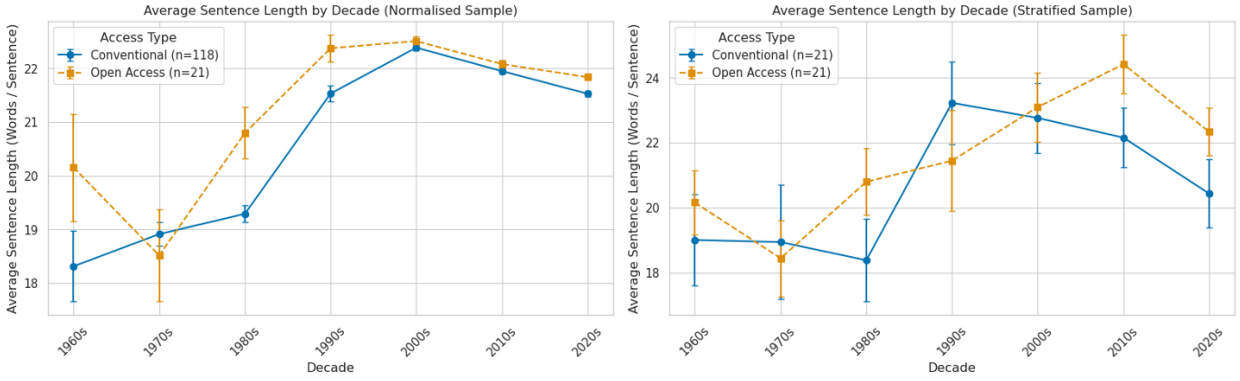

Figure 4

Temporal Trend of Flesch Reading Ease Scores by Publication Type and Decade

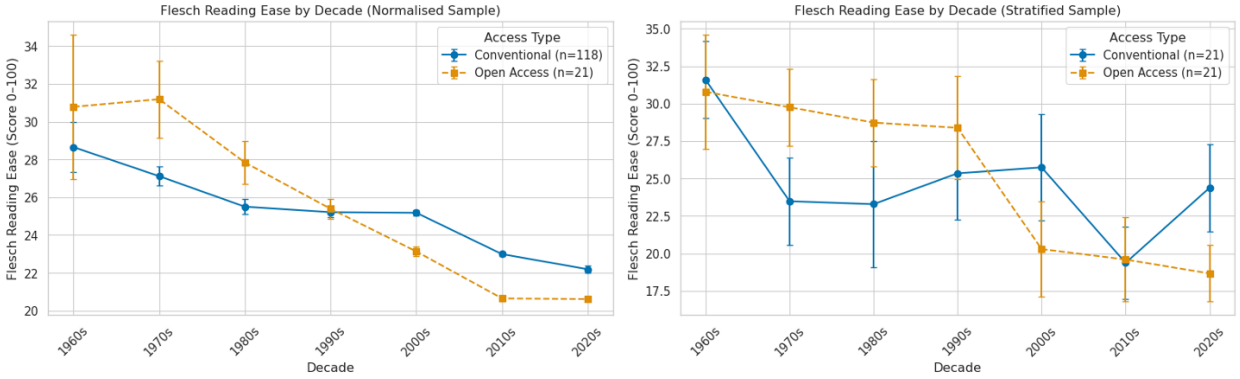

Figure 5

Temporal Trend of Flesch-Kincaid Grade Level by Publication Type and Decade

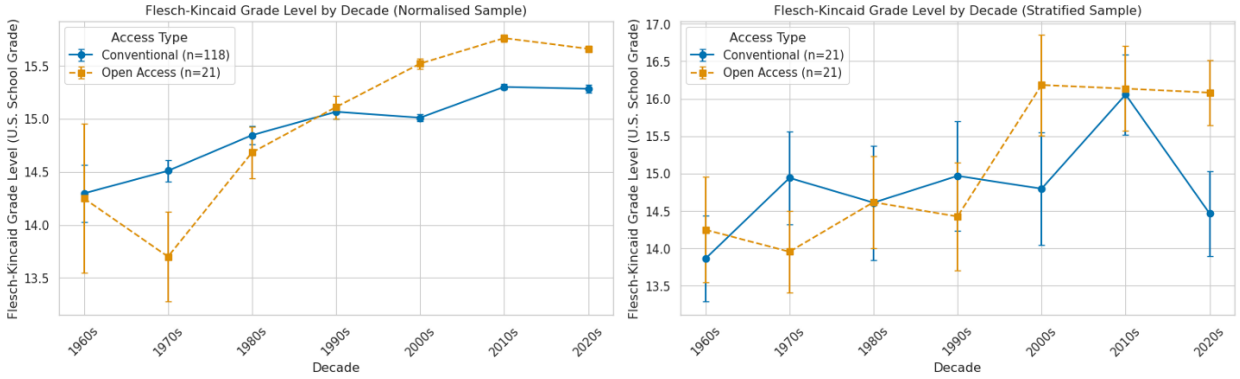

Figure 6

Temporal Trend of SMOG Index by Publication Type and Decade

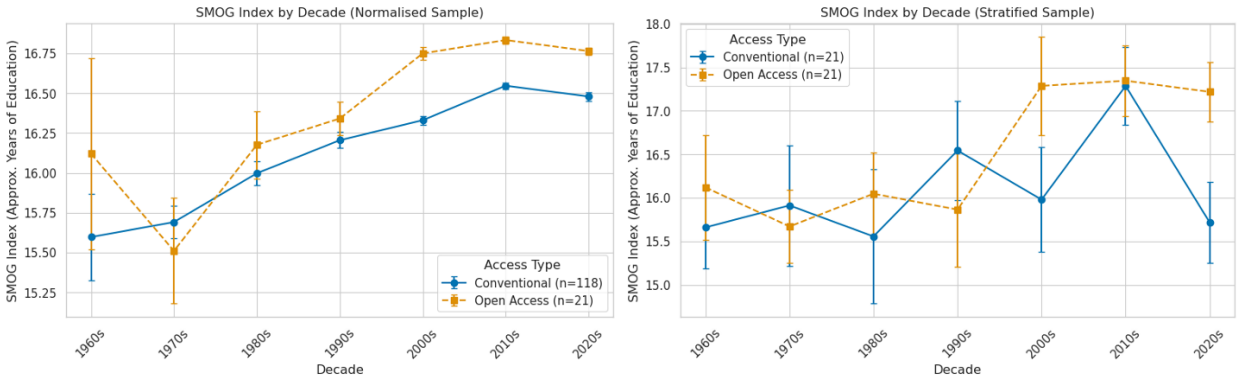

Figure 7

Temporal Trend of Gunning Fog Index by Publication Type and Decade

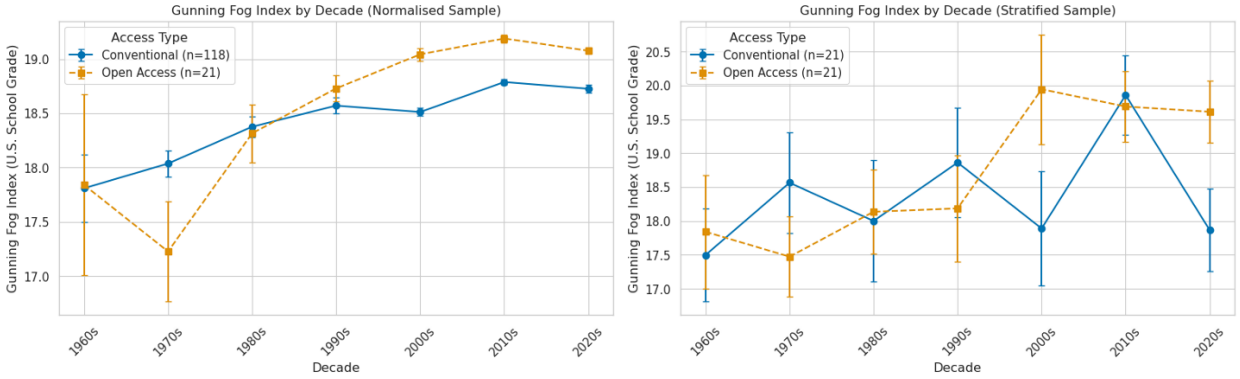

Figure 8

Temporal Trend of Passive Voice Ratio by Publication Type and Decade

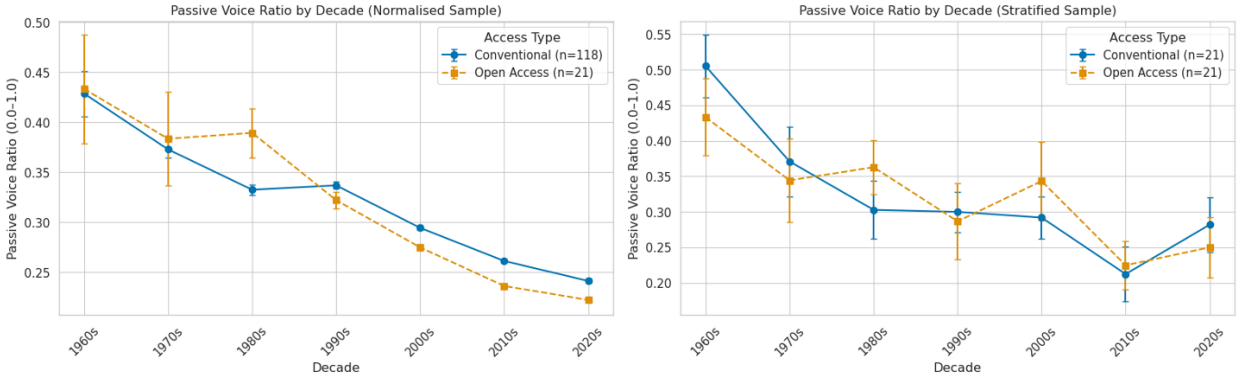

Figure 9

Temporal Trend of Hedging Percentage by Publication Type and Decade

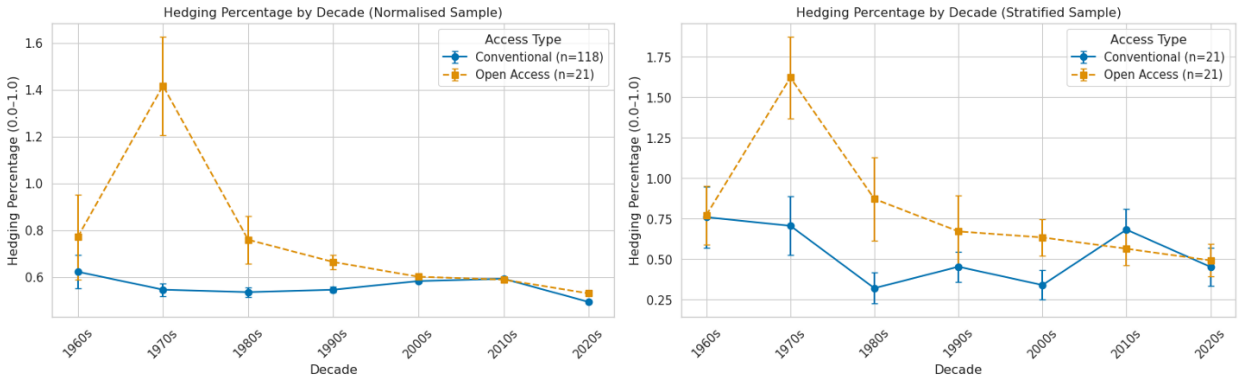

Figure 10

Temporal Trend of Lexical Density by Publication Type and Decade

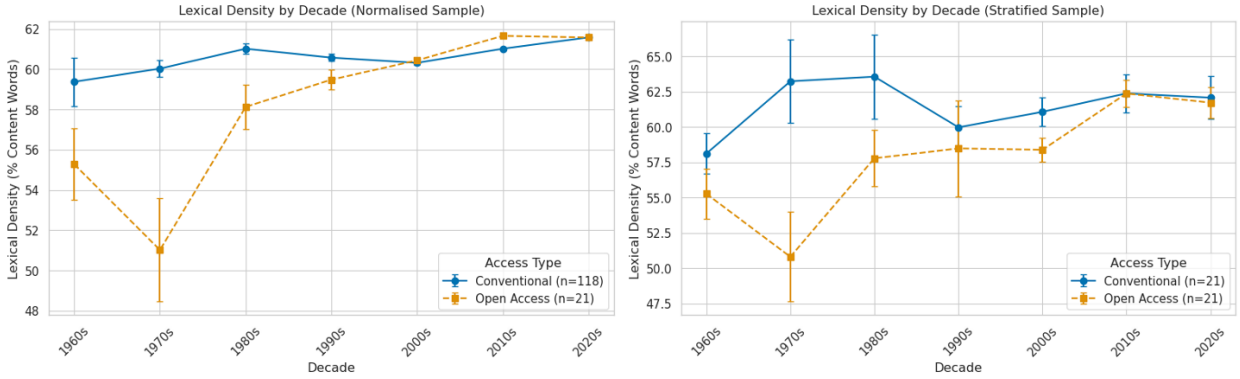

Figure 11

Distribution of Word Count by Publication Type for Normalised and Stratified Samples

Figure 12

Distribution of Average Sentence Length by Publication Type for Normalised and Stratified Samples

Figure 13

Distribution of Flesch Reading Ease Scores by Publication Type for Normalised and Stratified Samples

Figure 14

Distribution of Gunning Fog Index and Flesch-Kincaid Grade Level by Publication Type

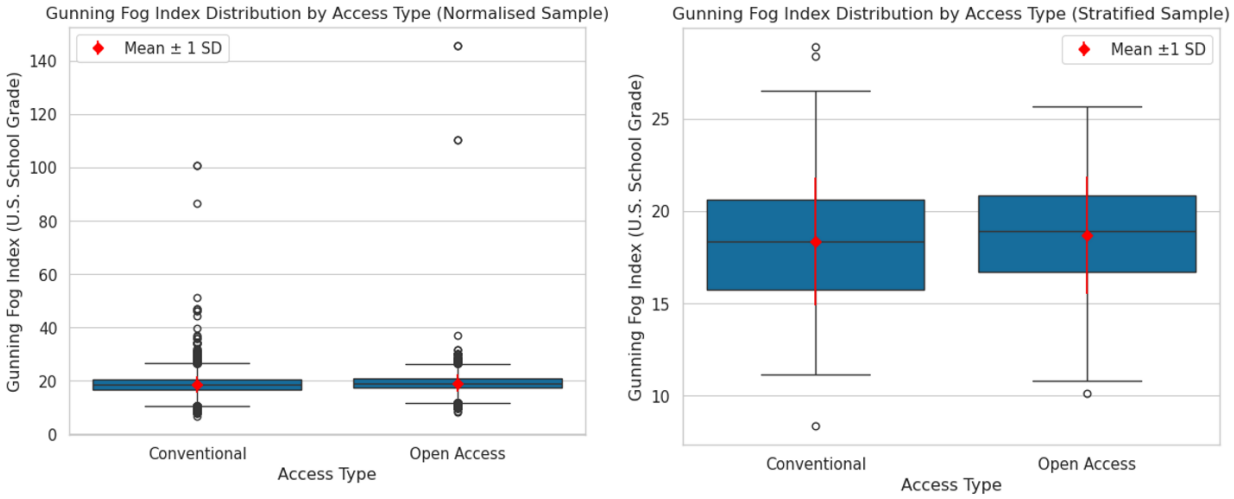

Figure 15

Distribution of SMOG Index by Publication Type for Normalised and Stratified Samples

Figure 16

Distribution of Gunning Fog Index by Publication Type for Normalised and Stratified Samples

Figure 17

Distribution of Passive Voice Ratio by Publication Type for Normalised and Stratified Samples

Figure 18

Distribution of Hedging Percentage by Publication Type for Normalised and Stratified Samples

Figure 19

Distribution of Lexical Density by Publication Type for Normalised and Stratified Samples

Figure 20

Distribution of Word Count by Decade for Normalised and Stratified Samples

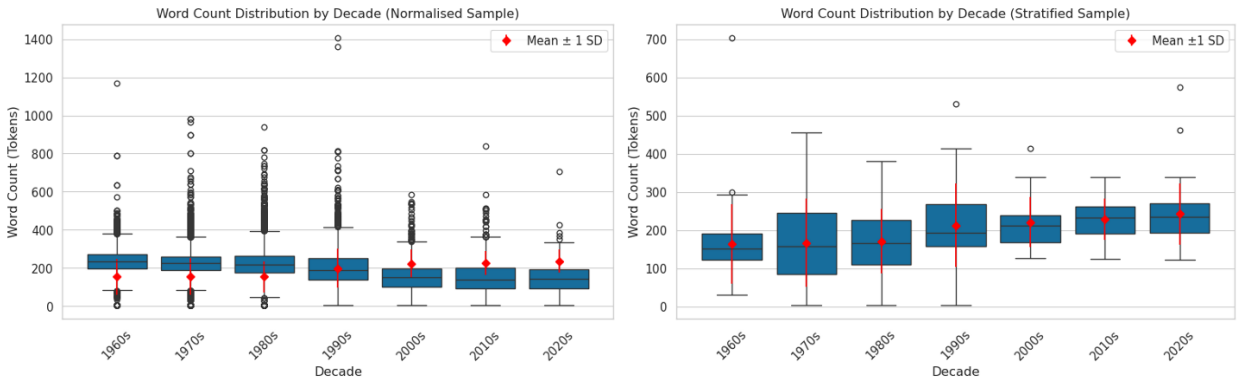

Figure 21

Distribution of Average Sentence Length by Decade for Normalised and Stratified Samples

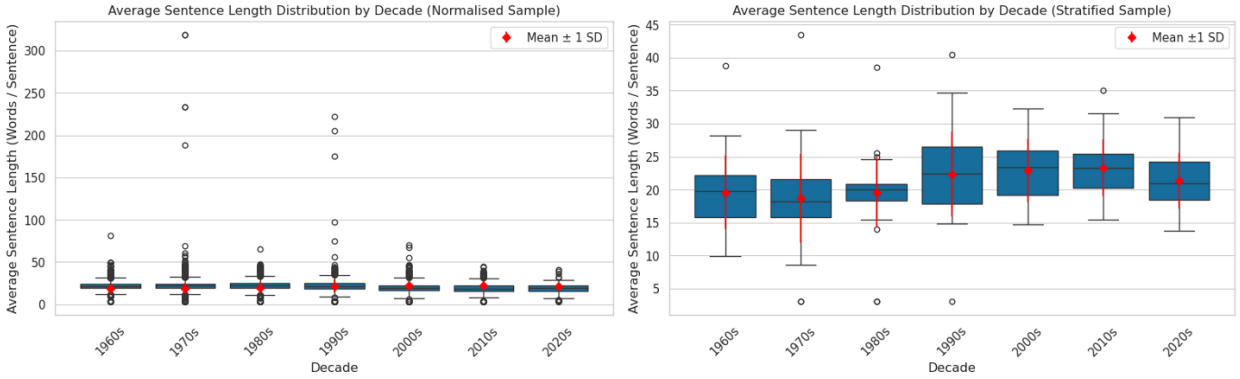

Figure 22

Distribution of Flesch Reading Ease Scores by Decade for Normalised and Stratified Samples

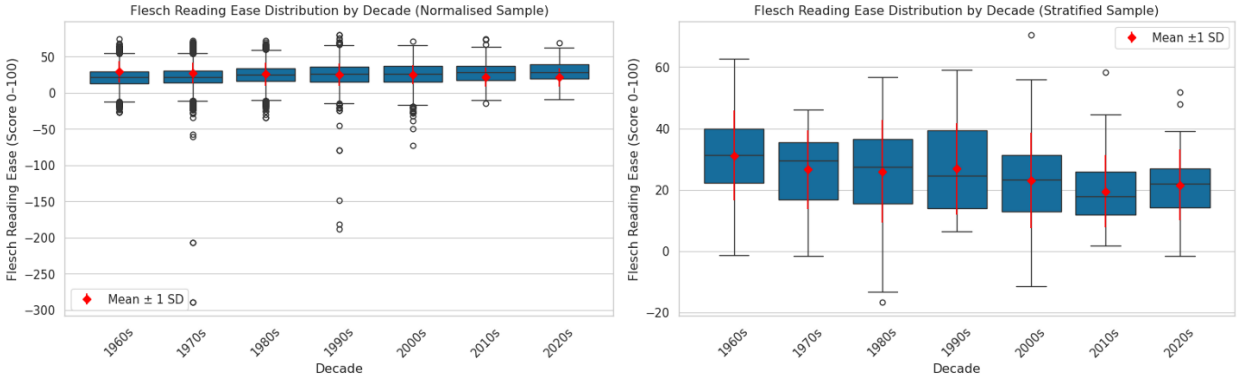

Figure 23

Distribution of Flesch-Kincaid Grade Level by Decade for Normalised and Stratified Samples

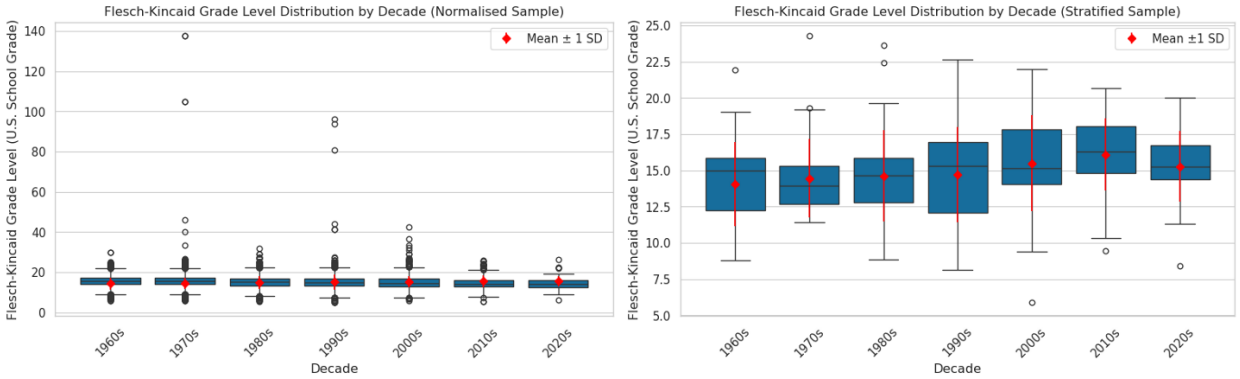

Figure 24

Distribution of SMOG Index by Decade for Normalised and Stratified Samples

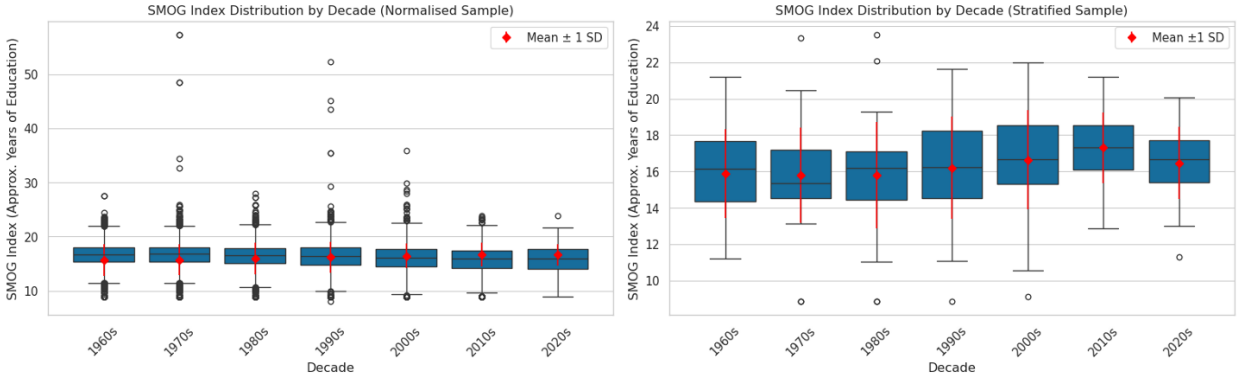

Figure 25

Distribution of Gunning Fog Index by Decade for Normalised and Stratified Samples

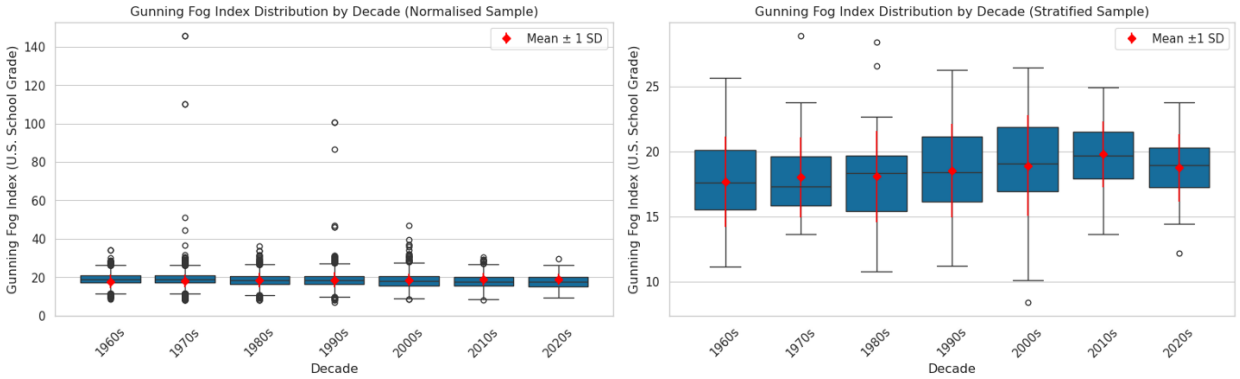

Figure 26

Distribution of Passive Voice Ratio by Decade for Normalised and Stratified Samples

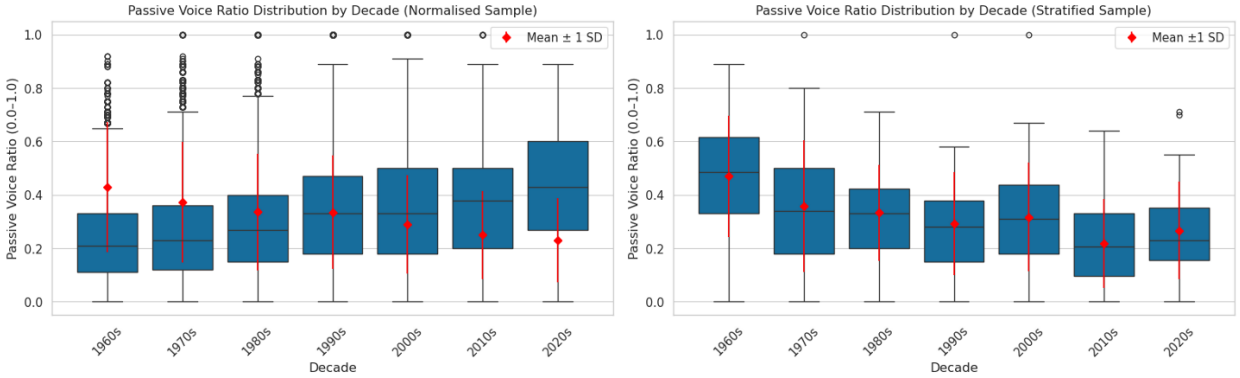

Figure 27

Distribution of Hedging Percentage by Decade for Normalised and Stratified Samples

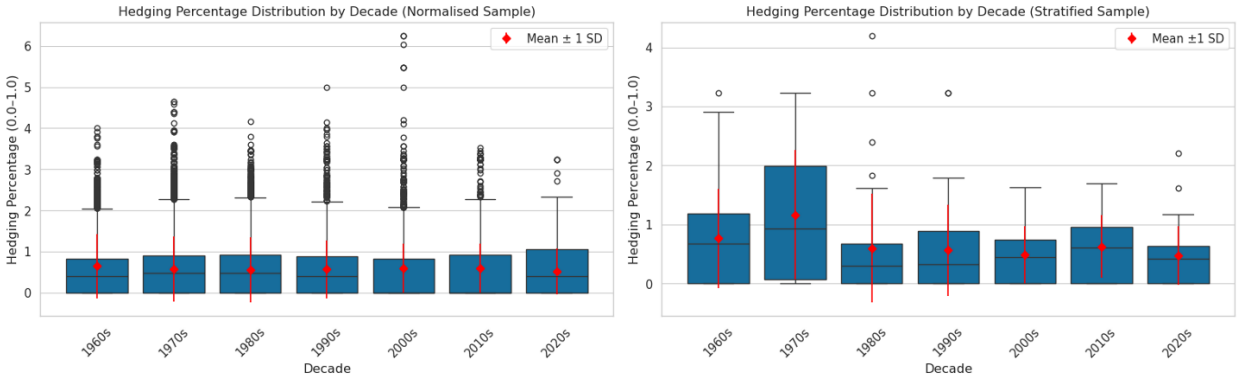

Figure 28

Distribution of Lexical Density by Decade for Normalised and Stratified Samples