Linguistic Research 43(1): 329-349
DOI: 10.17250/khisli.43.1.202603.012

The relationship between abstract discourse features and citation impact: An integrated analysis using Coh-Metrix and Biblioshiny

Myoungho Ha
(Silla University)

Ha, Myoungho. 2026. The relationship between abstract discourse features and citation impact: An integrated analysis using Coh-Metrix and Biblioshiny. *Linguistic Research* 43(1): 329-349. This study investigates how abstract-level discourse features relate to citation impact when considered alongside topical structure. A harmonized set of 436 English-language articles and reviews (1995–2024) from the Web of Science (Linguistics, Language & Linguistics, Education & Educational Research) was analyzed. Topic structure was mapped using bibliometrix through the Biblioshiny interface with all keywords (DE + ID; minimum frequency = 5), producing a co-word network, a thematic map (centrality × density), and period-wise evolution. Discourse was profiled with five Coh-Metrix composites—narrativity, syntactic simplicity, concreteness, referential cohesion, and deep cohesion—standardized as z-scores. Citation impact was normalized as average citations per year (TCperYear). Annual production peaked in 2003–2006, whereas TCperYear peaked later, indicating lagged diffusion. Thematic mapping showed a Motor axis centered on language–meaning–cognition, a Basic infrastructure of corpus/syntax/English, and bridging roles for computational linguistics and machine learning. In OLS with HC3 errors, syntactic simplicity and referential cohesion had small but significant negative associations with TCperYear; other indices were not significant. Results were robust in auxiliary checks and showed low multicollinearity (all VIFs < 2). Findings suggest that authors should maintain information density, limit mechanical repetition, and use only as much syntactic complexity as needed for precision—especially in high-centrality topics. **(Silla University)**

Keywords abstracts, citation impact, Coh-Metrix, bibliometric mapping, English for Academic Purposes

1. Introduction

In contemporary scholarly communication, the abstract functions as the gateway to discovery, visibility, readability, and ultimately citation. A substantial body of evidence links linguistic choices in titles and abstracts to downloads and citations (Jamali and Nikzad 2011; Letchford, Moat, and Preis 2015). The association between title structure and publication or citation outcomes can vary across journals, fields, and editorial stages (Fox and Burns 2015). Large cross-sectional analyses further suggest that heavy jargon and complex syntax reduce comprehensibility (Plavén-Sigray et al. 2017). Beyond text-internal factors, integrative reviews emphasize that journal prestige, collaboration networks, and topical positioning also shape citation outcomes (Tahamtan, Safipour Afshar, and Ahamdzadeh 2016). Taken together, these strands motivate a unified test, within a single consistent sample, of how micro-level discourse features—style, syntax, and cohesion—interact with macro-level topic structure to relate to citation impact.

Quantitative assessment of textual discourse has been formalized by the Coh-Metrix family of tools. Coh-Metrix reports multidimensional indices—narrativity, syntactic simplicity, concreteness, referential cohesion, and deep cohesion—as z-scores relative to a reference corpus, thereby profiling discourse features on a comparable scale (Graesser et al. 2004; McNamara et al. 2014). These indices have been validated as sensitive to genre norms and reading difficulty and are widely used across education, cognition, and discourse studies.

At the field level, shifts in annual production, citation patterns, and topic structure are typically traced through science mapping. The bibliometrix package, accessed through the Biblioshiny interface, reproducibly generates production and citation indicators, keyword co-occurrence networks, thematic maps (centrality × density), and thematic evolution flows (based on an inclusion index) from Web of Science and Scopus data (Aria and Cuccurullo 2017). Callon's coordinates for centrality and density and the inclusion index extend the classic co-word mapping tradition (Callon, Courtial, and Laville 1991; Cobo et al. 2011, 2012). Co-word analysis has also been adapted to information retrieval and knowledge management (Ding, Chowdhury, and Foo 2001), and alternative software such as VOSviewer is widely adopted (van Eck and Waltman 2010). Recent methodological reviews set standards for design, reproducibility, and reporting (Donthu et al. 2021; Župič and Čater 2015). Policies

affecting access and dissemination, including open access, have repeatedly been shown to influence citation impact (Eysenbach 2006; Tahamtan et al. 2016).

This study analyzes a homogeneous sample from the Web of Science Core Collection (N = 436). Annual production, average citations per year (TCperYear), and topic structure are estimated using Biblioshiny (co-occurrence mapping, thematic map, thematic evolution). Coh-Metrix composite indices—PCNARz, PCSYNz, PCCNCz, PCREFz, and PCDCz—are then related to TCperYear via regression. Topic maps are generated under common settings (all keywords: DE + ID; minimum frequency = 5; stemming off; stopwords on), and evolution flows are linked by the inclusion index.

Three research questions guide the analysis. First, how do annual output and TCperYear change from 1995 to 2024, and where do inflection points or lags appear? Second, how is the topic landscape organized under all keywords (DE + ID), and how does it evolve over time when read through Callon's coordinates and the inclusion index? Third, how are abstract-level discourse indices (PCNARz, PCSYNz, PCCNCz, PCREFz, PCDCz) associated with TCperYear, as estimated by ordinary least squares with HC3 standard errors and complemented by median (quantile) regression? The contributions are threefold: an integrated, reproducible pipeline that links topic structure, discourse features, and citation impact within one sample; an empirical connection between central topical position and abstract-level stylistic strategies; and practice-oriented guidance grounded in the joint configuration of information density, cohesion strategy, and topic centrality.

2. Literature review

2.1 Abstract discourse and citation diffusion: Readability and information compression

The abstract is the first gateway for search, display, and reading, and extensive evidence links linguistic choices in titles and abstracts to downloads and citations (Jamali and Nikzad 2011; Letchford et al. 2015). Large cross-sectional analyses report that heavy jargon and complex syntax reduce comprehensibility (Plavén-Sigray et al. 2017), while concise titles have been associated with higher citation counts (Letchford

et al. 2015). The strength—and even the direction—of these associations varies across journals, fields, and editorial stages (Fox and Burns 2015). Accordingly, integrated assessments should consider both text-internal factors and exogenous determinants such as journal influence, topical centrality, and access policies (Tahamtan et al. 2016). Beyond average readability effects, two issues are central: fit to field norms and the balance between information compression and content density.

2.2 Quantifying discourse features with Coh-Metrix

Coh-Metrix provides multidimensional indices—narrativity, syntactic simplicity, concreteness, referential cohesion, and deep cohesion—reported as *z*-scores relative to a reference corpus, enabling comparable profiles of textual discourse (Graesser et al. 2004; McNamara et al. 2014). Prior validation shows that these indices capture genre norms and reading difficulty; in particular, syntactic simplicity (higher values indicate simpler sentence structure) and cohesion indices (referential and deep) are closely tied to processing load and comprehension. Because interpretation is sensitive to the composition of the reference corpus and to genre mixing, designs that compare documents within a single sample and apply multivariate controls are recommended. The present study therefore examines five composite indices at the abstract level—PCNARz, PCSYNz, PCCNCz, PCREFz, and PCDCz—to test their independent links with citation outcomes. From a cognitive-processing perspective, these indices jointly capture how efficiently readers can allocate limited attentional resources to high-value propositions in an abstract, linking discourse configuration to information density and processing effort.

2.3 Science mapping of topic structure: The role of Biblioshiny

Macro-level field dynamics—changes in annual production and citation, and the formation and evolution of topic structure—are captured through science mapping. Using Biblioshiny as the interface to bibliometrix, the analysis generates production and citation indicators, keyword co-occurrence networks, thematic maps, and thematic evolution from Web of Science or Scopus data (Aria and Cuccurullo 2017). Thematic maps position clusters in Callon's centrality × density space, and thematic evolution

links clusters over time via the inclusion index, extending the classic co-word mapping tradition (Callon et al. 1991; Cobo et al. 2011, 2012). Co-word analysis has also been adapted in information retrieval and knowledge management (Ding et al. 2001), and alternative tools such as VOSviewer are widely used (van Eck and Waltman 2010). Because preprocessing choices—minimum keyword frequency, stemming, and stopword handling—can affect cluster stability and interpretive granularity, transparent reporting of settings is essential.

2.4 Bridging micro-discourse and macro topic landscapes: Gaps and positioning

Comprehensive reviews have codified standard procedures and reporting for bibliometric studies (Župič and Čater 2015; Donthu et al. 2021), yet designs that jointly analyze—within the same sample—topic structure (co-occurrence, strategic diagram, evolution) and abstract-level discourse indices (Coh-Metrix) to explain citation impact remain scarce. Citation counts are also shaped by exogenous factors—open access and dissemination policies, journal standing, and visibility—which must be considered when interpreting text-based effects (Eysenbach 2006; Tahamtan et al. 2016). The present study addresses this gap by linking Biblioshiny-derived temporal and topical landscapes with Coh-Metrix indices in a single sample to estimate the relative contributions of topical centrality and stylistic/cohesion strategies to citation impact. A further premise is that readability and compression effects are unlikely to be uniform across topics; the analysis therefore lays groundwork for exploring interactions between macro structures and micro discourse choices.

3. Methodology

3.1 Data source, preprocessing, and citation linkage

Materials were retrieved from the Web of Science (Core Collection) using the Advanced Search interface. Records were restricted to the Web of Science Categories Linguistics, Language & Linguistics, and Education & Educational Research. Topic terms in the TS field included: abstract, research abstract*, journal abstract*, academic writing, second language writing, English for Academic Purposes, and English for

Specific Purposes. The time span was 1995–2024 (PY), the language filter was English (LA), and document types were Article and Review (DT). After verifying filters on the results page, metadata were exported in the Full Record and Cited References format.

The initial pull (n = 450) was standardized under UTF-8 encoding with NFC normalization. Field names and formats were harmonized for UT, DOI, TI, AB, DE, ID, PY, SO, DT, and TC. Keywords (DE/ID) were delimiter-parsed, trimmed, and deduplicated; PY was cast to integer; DOI whitespace and casing were normalized. Duplicate removal followed a multi-step rule (UT → DOI → normalized title, TI_norm). Core fields (AB, DE/ID, PY, TC) were flagged when missing to enable downstream exclusion decisions. Abstracts were stored both as raw text (AB_raw) and as lightly cleaned text (AB) after removing extraneous spaces and page markers. The analytical sample comprised 436 documents, and a trial import into bibliometrix via Biblioshiny confirmed completeness and readable formatting.

To mitigate cohort differences and indexing lag, citation impact was expressed as TCperYear. The indicator was defined as total citations (TC) divided by years available, where years available denotes the elapsed time (in years) from the publication year to the citation-count date. When counts are taken at year-end, years available can be computed as (citation-count year − publication year + 1). Abstract texts and citation indicators matched one-to-one for all 436 records.

3.2 Topic mapping procedure (bibliometrix via Biblioshiny)

Topic structure and positioning were derived using bibliometrix through the Biblioshiny interface under a common specification: field = all keywords (DE + ID), minimum frequency = 5, stemming = off, stopwords = on. The co-occurrence network used a force-directed layout; interpretation emphasized relative distances and linkage patterns rather than absolute coordinates. The thematic map applied Callon's centrality × density framework to identify the Motor, Basic, Niche, and Emerging/Declining quadrants. The thematic evolution analysis segmented the timeline into three windows (1995–2003, 2004–2006, 2007–2024) and linked clusters across periods using the inclusion index as edge weight. For recent years, potential effects of indexing delay and policy changes were noted during interpretation.

3.3 Discourse indices, regression model, diagnostics, and robustness

Abstract-level discourse features were measured with five Coh-Metrix composite indices: PCNARz (narrativity), PCSYNz (syntactic simplicity), PCCNCz (concreteness), PCREFz (referential cohesion), and PCDCz (deep cohesion). All indices were standardized as *z*-scores relative to a reference corpus. Pairwise relationships were summarized with Spearman correlations to screen for potential multicollinearity; variance inflation factors (VIFs) were also computed as a supplementary diagnostic.

Regression analysis used TCperYear as the dependent variable and the five indices as predictors in an ordinary least squares (OLS) model. HC3 heteroskedasticity-consistent standard errors were applied. Coefficients were interpreted as the change in TCperYear associated with a one-standard-deviation increase in each index, holding the others constant. Diagnostics included standardized residual–leverage plots and Cook's distance, which were used to identify influential observations; indications of heteroskedasticity were documented.

Robustness checks assessed the stability of results under alternative specifications: (a) replacing the dependent variable with total citations (TC) in an OLS–HC3 model; (b) recalculating co-word and thematic diagrams with keyword minimum frequencies of 4 and 6; and (c) shifting thematic evolution window boundaries by ±1 year. The main patterns remained stable across these variations. To bound interpretation conservatively, possible residual confounding due to journal influence, subfield, or year effects, as well as sensitivity of *z*-scores to the reference corpus, is noted. A median (τ = 0.5) quantile regression was additionally estimated as a sensitivity check.

4. Analysis and discussion

This section links bibliometric mapping (bibliometrix via Biblioshiny) with discourse metrics (Coh-Metrix) on the same sample (N = 436). Results are presented in the following order: (i) temporal dynamics (annual production and TCperYear), (ii) the strategic layout of the topic landscape (thematic map), and (iii) the association between abstract-level discourse indices and citation impact (OLS–HC3 with a supplemental median regression). Topic maps were generated under a common setting (field = all keywords (DE + ID), minimum frequency = 5, stemming = off, stopwords

= on). Detailed figures for the co-occurrence network, period-wise evolution, and the inter-index correlation heatmap appear in Appendix Figures A1–A3.

4.1 Overview of production and citation time series

To contextualize subsequent topic and regression results, Figure 1 traces annual output (number of items per publication year) and Figure 2 plots TCperYear, defined as total citations divided by years available. Read together, the series distinguish production volume from impact dynamics and help identify inflection points and lags. Because recent years may be affected by indexing delay and a not-yet-open citation window, values at the right tail of both series should be interpreted cautiously.

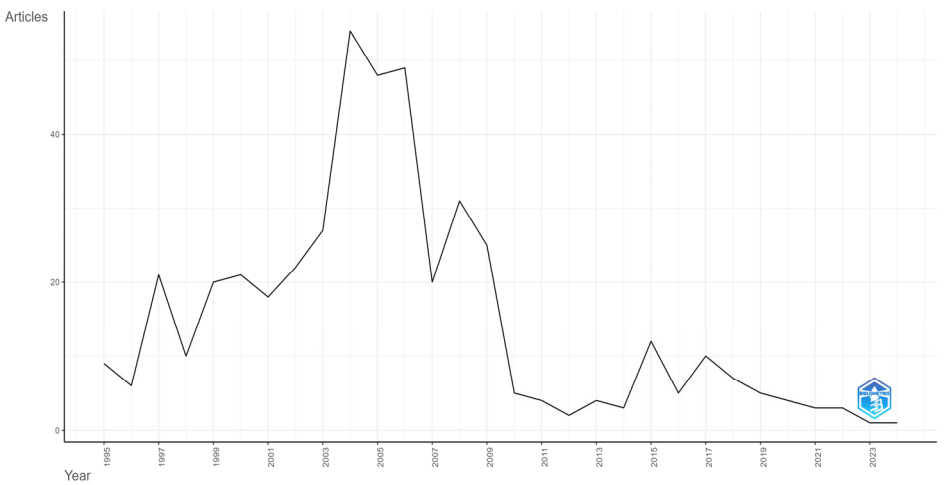

Figure 1. Annual scientific production

Annual output increases steadily after 1995, peaks sharply during 2003–2006, and then stabilizes at roughly 40–50 items per year. Around 2007 the trajectory shifts into a mild decline or leveling phase. Short rebounds appear near 2015 and 2017, followed by a gentle downward drift with minor fluctuations in the most recent years. This pattern likely reflects a mix of external factors—special-issue or project bursts, shifts in database indexing policy, and redistribution of attention as the topic landscape evolves. Values in the terminal years should be interpreted cautiously due to indexing lag and occasional coverage gaps. Detailed network, period evolution, and correlation

diagrams are provided in Appendix Figures A1–A3 for reference.

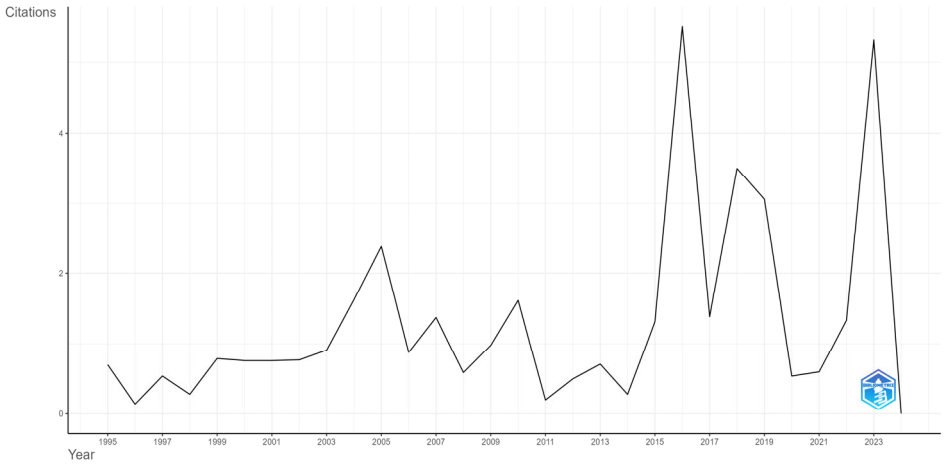

Figure 2. Average citations per year

The citation trajectory diverges from production. In the early years it hovers at roughly one citation per year, then bends upward around 2005 as cohorts begin to accumulate citations more rapidly. A clear peak appears near 2016, followed by intermittent spikes. The apparent downturn in the most recent years is likely mechanical, reflecting a not-yet-open citation window. In short, the lag between the production peak (mid-2000s) and the citation peak (mid-2010s) is consistent with familiar citation dynamics in which highly cited work diffuses over time.

4.2 Strategic layout of topics: Thematic map

To provide a compact view of role structure in the field, a thematic map was produced in Biblioshiny under a common specification (field = all keywords (DE + ID), minimum frequency = 5, stemming = off, stopwords = on). On Callon's plane, centrality reflects the breadth and strength of a cluster's links to other clusters (external importance), whereas density indicates within-cluster development (internal maturity). Accordingly, quadrant membership and relative distances—rather than absolute coordinates—guide interpretation.

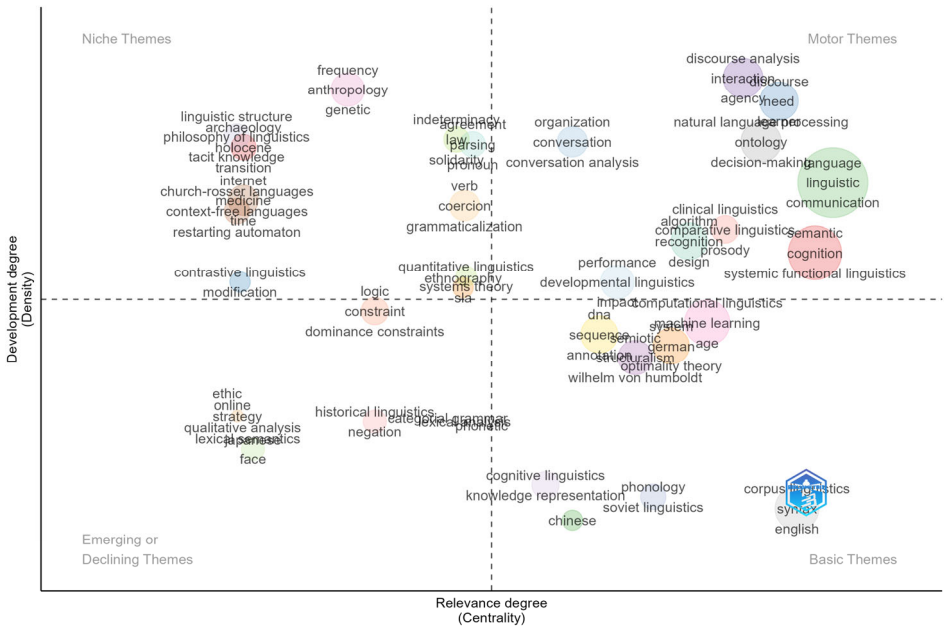

Figure 3. Thematic map (centrality × density)

On Callon's plane (x = centrality, y = density), the language–linguistic–communication cluster anchors the Motor quadrant. It is flanked by semantic, cognition, and systemic functional linguistics, collectively forming a language–meaning–cognition axis that is both well connected and internally developed. Corpus linguistics, syntax, and English occupy the Basic quadrant: these themes provide infrastructural vocabulary and methods, showing broad ties to the rest of the map but lower internal cohesion, consistent with their role as shared resources across subareas. Computational linguistics and machine learning sit on the Motor–Basic boundary and act as bridges from core discourse to infrastructural techniques; their boundary position suggests ongoing consolidation and cross-talk rather than a fully crystallized module.

The Niche quadrant (upper left) concentrates formal-theoretical, philosophical, and algorithm-specific modules. High density but limited external ties implies specialized conversations—valuable within their subcommunity yet less central to the field's traffic patterns. The Emerging/Declining quadrant (lower left) contains qualitative analysis,

historical linguistics, and phonetics/phonology, indicating topics either in early formation (not yet well connected) or in relative dispersion. Borderline points near quadrant edges should be read as transitional: small shifts in keyword frequency or preprocessing can move them across boundaries.

Overall, the map depicts a core propelled by language–meaning–cognition, supported by a Basic infrastructure (corpus/syntax/English), with computational approaches bridging core and infrastructure. This structure aligns with the co-occurrence network (Appendix Figure A1) and the period-wise lineages (Appendix Figure A3), where trajectories from computational/corpus toward cognitive/contrastive foci are visible. In the next subsection, distributional properties and inter-index correlations of the Coh-Metrix measures are summarized to connect this macro positioning with micro-level discourse choices.

4.3 Descriptive statistics and correlation summary for Coh-Metrix indices

Table 1 reports distributional summaries (count, mean, standard deviation, range, and quartiles) for five Coh-Metrix composites—PCNARz (narrativity), PCSYNz (syntactic simplicity), PCCNCz (concreteness), PCREFz (referential cohesion), and PCDCz (deep cohesion)—expressed as *z*-scores relative to the reference corpus. Values are rounded to three decimals.

Table 1. Distribution of Coh-Metrix composite indices (436 abstracts, z-scores)

	count	mean	std	min	25%	50%	75%	max
PCNARz	436	-1.260	0.572	-3.289	-1.623	-1.267	-0.830	0.355
PCSYNz	436	-0.877	0.945	-4.219	-1.489	-0.882	-0.305	1.978
PCCNCz	436	-0.408	1.014	-4.986	-1.109	-0.377	0.263	3.381
PCREFz	436	-0.109	1.259	-4.085	-0.985	-0.235	0.802	3.008
PCDCz	436	-0.340	1.165	-3.135	-1.158	-0.354	0.404	3.069

Notes. Indices are *z*-scores relative to the reference corpus (mean = 0, *SD* = 1); values > 0 indicate higher levels than the reference mean. All values are rounded to three decimals.

Medians lie below zero for all indices, indicating that—relative to the reference corpus—these abstracts are, on average, less narrative, syntactically more complex, and lexically more abstract. Dispersion patterns differ by construct. PCNARz is tightly

clustered (Q1–Q3 ≈ −1.623 to −0.830; IQR ≈ 0.793), consistent with a stable low-narrativity norm in abstracts. PCSYNz shows a long upper tail (max = 1.978), suggesting a minority of abstracts adopting relatively simple sentence structures while the majority remain complex. PCCNCz centers slightly below zero yet spans a wide range, consistent with methods- or materials-heavy texts raising concreteness in some cases. Cohesion varies most: PCREFz and PCDCz have broad interquartile and overall ranges, implying heterogeneous use of lexical repetition/anaphora (referential cohesion) and causal/conditional/connective signaling (deep cohesion).

The rank-correlation structure (Appendix Figure A2) is moderate overall. The strongest negative association is PCSYNz–PCREFz ($\rho \approx -0.514$), and the strongest positive is PCNARz–PCREFz ($\rho \approx +0.451$). Several pairs (e.g., PCNARz–PCCNCz) are near zero, indicating partial independence. Variance inflation factors are low (all VIFs < 2; Appendix Table A4), so multicollinearity is limited and all five indices are jointly included in the subsequent regression. These distributional and correlational patterns calibrate expected effect sizes and provide context for the results reported in Section 4.4.

4.4 Discourse indices and citation impact: OLS–HC3 results

Below, an OLS model with HC3 robust standard errors links five Coh-Metrix indices (*z*-scores) to TCperYear. Coefficients are interpreted as the expected change in citations per year for a 1-*SD* increase in each index, holding the others constant. Table 2 reports mean effects at the center of the distribution; distributional heterogeneity is examined via a median ($\tau = 0.5$) quantile regression in the Appendix.

Table 2. OLS–HC3 coefficients for TCperYear by Coh-Metrix indices ($N = 436$)

	coef	se(HC3)	*t*	*p*
Intercept	1.926	0.209	9.234	< .001
PCNARz	0.043	0.134	0.319	0.750
PCSYNz	-0.185	0.084	-2.186	0.029
PCCNCz	-0.097	0.068	-1.423	0.155
PCREFz	-0.172	0.061	-2.827	0.005
PCDCz	-0.073	0.057	-1.282	0.200

Notes. OLS with HC3 robust standard errors. All predictors are *z*-scores of Coh-Metrix composite indices. Coefficients indicate the change in TCperYear (citations/year) for a 1-*SD* increase in each predictor, holding others constant. Intercept denotes expected TCperYear when all predictors equal zero. [$R^2 = 0.036$; $Adj.\ R^2 = 0.025$]; $N = 436$. Values are rounded to three decimals; $p < .001$ is reported as "< .001."

The intercept (≈ 1.926) is statistically significant, indicating that abstracts at the reference-mean level on all indices are cited, on average, about 1.93 times per year. Two indices show small but reliable negative associations with TCperYear: syntactic simplicity (PCSYNz; $\beta = -0.185$, $p = .029$) and referential cohesion (PCREFz; $\beta = -0.172$, $p = .005$). Narrativity, concreteness, and deep cohesion are not significant at conventional levels. Taken together, these results suggest that, net of other features, stronger surface repetition and aggressive simplification are linked to lower annual citation rates, whereas other discourse dimensions play a limited role in the mean of the distribution.

The overall explanatory power of the model is modest ($R^2 = 0.036$; $Adj.\ R^2 = 0.025$). This pattern is expected, because citation impact is strongly shaped by external factors such as journal influence, collaboration networks, subfield, and access conditions, which are not explicitly modeled here. The coefficients should therefore be interpreted as marginal adjustments in citation rates associated with abstract-level discourse features, rather than as a comprehensive predictive model. Extending the analysis with additional article- and journal-level covariates or multilevel structures is left for future work.

Median ($\tau = 0.5$) quantile regression yields attenuated or null effects, consistent with influences that are stronger in the tails of the citation distribution. A parallel specification using total citations (TC) produces the same sign pattern (Appendix Tables A1–A3), and multicollinearity remains limited (all VIFs < 2; Appendix Table A4).

4.5 Integrating time, topics, and discourse

Taken together, the three strands align. For temporal dynamics (RQ1), annual output and citation impact peak at different times, indicating lagged diffusion: production rises sharply in the mid-2000s, whereas TCperYear crest later—consistent with the time needed for high-impact work to gain recognition. For topic structure

and evolution (RQ2), the field centers on a Motor axis of language–meaning–cognition, with a Basic infrastructure of corpus linguistics, syntax, and English; computational linguistics and machine learning bridge core and infrastructural themes. Period lineages indicate a shift from computational/corpus emphases toward cognitive/contrastive orientations, while several specialized modules remain niche or transient.

For discourse–impact links (RQ3), OLS–HC3 estimates suggest that abstracts associated with higher citation rates tend to preserve information density, avoid unnecessary surface repetition, and use only the syntactic complexity needed for precision. In highly central topical areas, this balance appears especially consequential, implying that topic position and rhetorical strategy should be interpreted jointly rather than in isolation. Supporting materials appear in the Appendix: co-occurrence network (Figure A1), period-wise thematic evolution (Figure A3), inter-index correlation heatmap (Figure A2), auxiliary models with total citations (Tables A1–A3), and VIF diagnostics (Table A4). All numeric values are rounded to three decimals; p values below .001 are reported as $p < .001$.

The evidence suggests that abstracts associated with stronger citation performance typically maintain high information density by foregrounding core results and contributions rather than relying on paraphrastic restatement. From a cognitive-processing perspective, concentrating semantic content in a compact but well-structured abstract reduces unnecessary working-memory load and allows readers to identify the main contribution more quickly. Surface repetition—whether through reiterated terminology or extended anaphoric chains—adds length without commensurate gains in clarity and should be minimized. Syntactic complexity is most effective when calibrated to necessity: structures should be sufficiently elaborate to preserve precision, but not so nested as to obscure the main claim. Logical relations are conveyed more transparently when a small set of high-information connectives is used selectively, rather than dense strings of discourse markers. Because topical centrality appears to condition these relations, the balance between compression and clarity is especially consequential in highly central themes. These points are best treated as conditional guidelines aligned with the genre norms of the target journal; in all cases, key information should appear early and be explicitly linked to the study's contribution.

5. Conclusion

This study integrated topic mapping with Biblioshiny and discourse measurement with Coh-Metrix on a single, harmonized sample of 436 abstracts to examine how macro-level topic structure (time series of output and citation, co-word networks, thematic map/evolution) and micro-level discourse features (narrativity, syntactic simplicity, concreteness, referential cohesion, deep cohesion) relate to TCperYear. Topic maps were generated under a common all keywords configuration, and citation counts were normalized as TCperYear to mitigate cohort and indexing effects. The outcome is a transparent, reproducible pipeline for combined bibliometric–discourse analysis.

Three findings stand out. First, production and citation peak at different times: the 2003–2006 production plateau is followed by a lagged peak in TCperYear. Second, in the topic landscape, the language–linguistic–communication axis functions as a Motor cluster, whereas corpus linguistics, syntax, and English serve as Basic infrastructure; thematically, a clear lineage runs from computational/corpus toward cognitive/contrastive foci. Third, in OLS with HC3 robust standard errors, syntactic simplicity (PCSYNz) and referential cohesion (PCREFz) show small but statistically significant negative coefficients, while narrativity, concreteness, and deep cohesion are not significant at conventional levels.

These patterns carry practical implications for abstract writing. Over-simplifying sentence structure or relying on mechanical surface repetition may be associated with lower citation impact by reducing information density and precision. By contrast, selectively linking core information and maintaining logical cohesion with only the complexity needed for precision appears more compatible with higher impact—especially in high-centrality topical areas where the balance between compression and clarity is critical.

Scholarly contributions are threefold. First, an analysis procedure is presented that links topic centrality (centrality × density) and discourse indices to citation outcomes within the same sample, narrowing the gap between bibliometrics and discourse analytics. Second, beyond generic claims about readability, the results point to conditional style strategies: sustaining information compression while curbing superficial repetition can be advantageous. Third, by reading regression estimates alongside time-series and evolutionary context, the study offers an interpretive frame

that emphasizes the context dependence of style effects.

Limitations and directions for future work are clear. External validity is constrained by the focus on English-language abstracts indexed in WoS, and residual external factors (e.g., journal influence, open-access policies) may remain. In addition, the relatively low R^2 values of the citation models indicate that much of the variance in citation impact remains unexplained by abstract-level features alone, underscoring the need to incorporate richer contextual covariates in future designs. Additional precision could come from quantile regressions targeting upper and lower tails, models with interactions between topic centrality and discourse indices, and fixed effects for journal, year, and subfield. Re-calibrating Coh-Metrix standardization against a domain-specific reference corpus may also improve measurement accuracy. Such extensions would enable a stricter test of the conclusions reported here.

References

Aria, Massimo and Corrado Cuccurullo. 2017. Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11(4): 959-975. https://doi.org/10.1016/j.joi.2017.08.007.

Callon, Michel, Jean-Pierre Courtial, and Françoise Laville. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics* 22(1): 155-205. https://doi.org/10.1007/BF02019280.

Cobo, Manuel José, Antonio Germán López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. 2011. Science mapping software tools: Review, analysis, and cooperative study. *Journal of the American Society for Information Science and Technology* 62(7): 1382-1402. https://doi.org/10.1002/asi.21525.

Cobo, Manuel José, Antonio Germán López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. 2012. SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology* 63(8): 1609-1630. https://doi.org/10.1002/asi.22688.

Ding, Ying, Gobinda G. Chowdhury, and Schubert Foo. 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management* 37(6): 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0.

Donthu, Naveen, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of*

Business Research 133: 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070.

Eysenbach, Gunther. 2006. Citation advantage of open access articles. *PLOS Biology* 4(5): e157. https://doi.org/10.1371/journal.pbio.0040157.

Fox, Charles W. and C. Sean Burns. 2015. The relationship between manuscript title structure and success: Editorial decisions and citation performance for an ecological journal. *Ecology and Evolution* 5(10): 1970-1980. https://doi.org/10.1002/ece3.1480.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 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.

Jamali, Hamid R. and Mahsa Nikzad. 2011. Article title type and its relation with the number of downloads and citations. *Scientometrics* 88(2): 653-661. https://doi.org/10.1007/s11192-011-0412-z.

Letchford, Adrian, Helen Susannah Moat, and Tobias Preis. 2015. The advantage of short paper titles. *Royal Society Open Science* 2(8): 150266. https://doi.org/10.1098/rsos.150266.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664.

Plavén-Sigray, Pontus, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *Elife* 6: e27725. https://doi.org/10.7554/eLife.27725.

Tahamtan, Iman, Abbas Safipour Afshar, and Kousha Ahamdzadeh. 2016. Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics* 107(3): 1195-1225. https://doi.org/10.1007/s11192-016-1889-2.

van Eck, Nees Jan and Ludo Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84(2): 523-538. https://doi.org/10.1007/s11192-009-0146-3.

Župič, Ivan and Tomaž Čater. 2015. Bibliometric methods in management and organization. *Organizational Research Methods* 18(3): 429-472. https://doi.org/10.1177/1094428114562629.

Appendix

Note. This appendix compiles figures and ancillary estimates omitted from the main text for brevity. All numeric values are rounded to three decimal places; p values smaller than .001 are reported as "$p < .001$." All predictors are z-scores of the Coh-Metrix composite indices (standardized against the reference corpus). Coefficients represent the expected change in the dependent variable for a 1-SD increase in the given index, holding other variables constant. The intercept/const reflects the expected outcome when all indices equal 0. Sample size N = 436. TCperYear is defined as total citations (TC) divided by years available (elapsed years from publication year to the citation-count date).

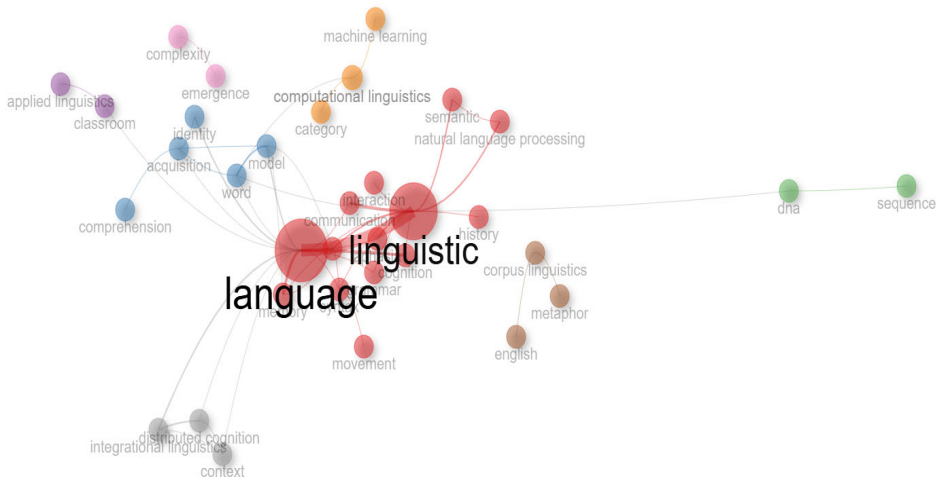

Figure A1. Keyword co-occurrence network (all keywords, minimum frequency = 5, stemming = off, stopwords = on)

Description. Force-directed layout of the co-occurrence network. Node size reflects term frequency; edge thickness reflects co-occurrence strength. The language/linguistic family functions as the hub, with radial branches along meaning–computational/applied–corpus–discourse/cognitive axes. Complexity and emergence appear as thinly connected niche modules; dna/sequence sits at the outer periphery as an outlier.

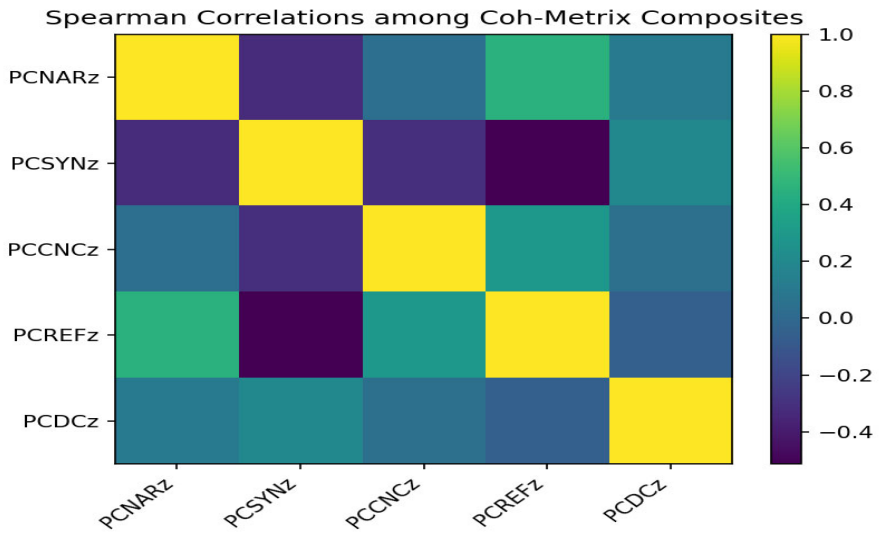

Figure A2. Spearman correlation heatmap (PCNARz, PCSYNz, PCCNCz, PCREFz, PCDCz)

Description. The diagonal shows self-correlations (1.000). Color intensity encodes |ρ|; hue encodes the sign (positive/negative). In this sample, the largest negative correlation is PCSYNz–PCREFz (ρ = −0.514), and the largest positive correlation is PCNARz–PCREFz (ρ = 0.451). Overall collinearity is limited, with caution warranted only for a few pairs.

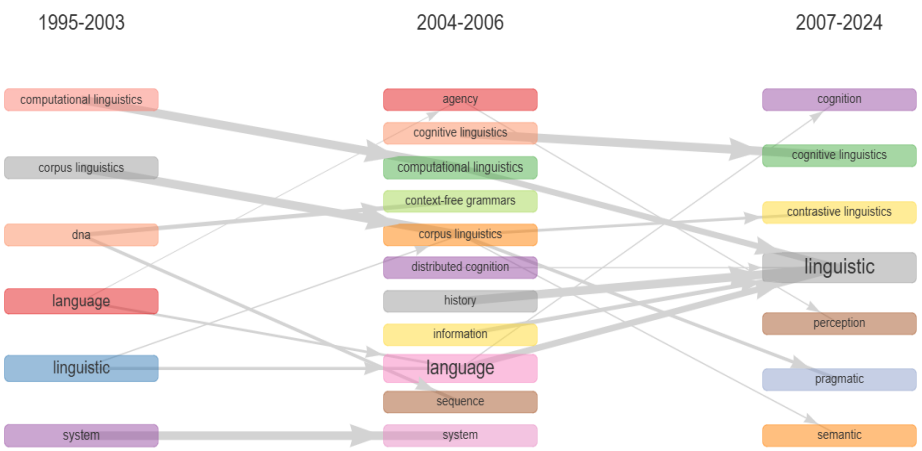

Figure A3. Thematic evolution by period (1995–2003 / 2004–2006 / 2007–2024)

Description. Bars denote period-specific co-word clusters; arrows connect clusters across periods weighted by the inclusion index. A dominant lineage runs from computational/corpus to cognitive/contrastive, alongside branches into methodological/meta themes and the dissolution/absorption of some specialized modules.

Table A1. OLS coefficients (dependent variable: total citations, TC; HC3 not applied)

Item	coef	se	*t*	*p*
Intercept	38.578	3.953	9.760	< .001
PCNARz	0.651	2.509	0.259	.795
PCSYNz	-3.654	1.605	-2.276	.023
PCCNCz	-1.910	1.319	-1.448	.148
PCREFz	-3.357	1.217	-2.759	.006
PCDCz	-1.468	1.112	-1.321	.187

Note. Ordinary least squares (mean effects) without robust standard errors. Predictors are *z*-scored Coh-Metrix composites. $N = 436$.

Table A2. OLS–HC3 coefficients (dependent variable: total citations, TC)

Item	coef	se(HC3)	*t*	*p*
Intercept	38.578	4.156	9.282	< .001
PCNARz	0.651	2.678	0.243	.808
PCSYNz	-3.654	1.685	-2.168	.031
PCCNCz	-1.910	1.365	-1.399	.162
PCREFz	-3.357	1.208	-2.778	.006
PCDCz	-1.468	1.140	-1.287	.199

Note. OLS with HC3 heteroskedasticity-consistent standard errors. Predictors are *z*-scored Coh-Metrix composites. $N = 436$.

Table A3. Median regression (QR, $\tau = 0.5$) coefficients (dependent variable: total citations, TC)

Item	coef	se	*z*	*p*
Intercept	60.000	2.576	23.296	< .001
PCNARz	0.000	1.635	0.000	1.000
PCSYNz	0.000	1.046	0.000	1.000
PCCNCz	0.000	0.859	0.000	1.000
PCREFz	0.000	0.793	0.000	1.000
PCDCz	0.000	0.724	0.000	1.000

Note. Median ($\tau = 0.5$) quantile regression. Coefficients with absolute values below 0.0005 are rounded to 0.000 (underlying estimates are near zero in scientific notation). $N = 436$.

Table A4. VIF (Variance Inflation Factor) summary (auxiliary diagnostic)

Variable	VIF	Tolerance	R^2_aux
PCNARz	1.358	0.737	0.263
PCSYNz	1.519	0.658	0.342
PCCNCz	1.179	0.849	0.151
PCREFz	1.548	0.646	0.354
PCDCz	1.107	0.903	0.097

Note. VIFs are computed for the predictor set (PCNARz, PCSYNz, PCCNCz, PCREFz, PCDCz); by definition, VIF = 1/Tolerance and R^2_aux = 1 − Tolerance. All VIFs are below 2 (well under common thresholds of 5 or 10), indicating low multicollinearity risk. N = 436.

Myoungho Ha
Assistant Professor
Silla University
140 Baegyang-daero(Blvd), 700beon-gil(Rd),
Sasang-gu, Busan 46958, Korea
E-mail: hadash21@silla.ac.kr

Received: 2025. 10. 22.
Revised: 2025. 11. 13.
Accepted: 2025. 12. 05.