

Register Mimicry, Not Grammatical Deficiency: A Corpus-Based Analysis of It-Extraposition and Passive Voice in High-Impact versus National-Tier Research Abstracts

Waniya Maryam
waniamariam92@gmail.com
University of Agriculture, Faisalabad

Dr. Ayesha Asghar Gill
ayesha.asghar@uaf.edu.pk
Assistant Professor, University of Agriculture, Faisalabad

Corresponding Author: Waniya Maryam waniamariam92@gmail.com

Received: 30-11-2025

Revised: 14-12-2025

Accepted: 27-12-2025

Published: 10-01-2026

ABSTRACT

Non-canonical syntax is widely recognized as a strategic information-packaging resource in academic prose, yet no corpus-based study has compared how it-extraposition and passive voice function across institutionally stratified publication tiers within the compressed register of the research abstract. This study addresses that gap through a comparative corpus analysis of 80 HEC-recognized Pakistani linguistics and literature abstracts. A balanced corpus (40 W-category, 40 Y-category; 15,064 tokens) was analyzed in AntConc 3.5.9 using regular-expression queries and manual verification, with frequencies normalized per 1,000 words and inter-rater agreement of 85%. Qualitative analysis, grounded in Quirk et al.'s (1985) End-Weight/End-Focus principles, Prince's (1981) Given-New taxonomy, and Biber et al.'s (1999) register framework, interpreted how these structures organized clausal information. Y-category writers deploy passive voice 47.5% more frequently (17.02 vs. 11.54 per 1,000 words) and it-extraposition roughly three times more frequently (1.14 vs. 0.42) than W-category writers. Qualitative analysis reverses this surface impression: W-category extrapositions consistently involve clauses of 24–33 words against two- to three-word matrices, confirming genuine processing necessity, while Y-category instances average 11–15 words, with two cases driven by stylistic formality rather than cognitive demand. W-category modal passives cluster in speculative perimeters as disciplined epistemic hedges; Y-category modal passives cluster in the empirical core as rhetorical enforcement. Subject-Auxiliary Inversion yielded zero instances in both corpora, confirming a categorical genre-level register barrier. The study concludes that Y-category writers exhibit register mimicry — reproducing surface markers of academic formality at higher frequency without the functional precision that characterizes register mastery.

Keywords: non-canonical syntax; it-extraposition; passive voice; information packaging; corpus linguistics; register theory; End-Weight; End-Focus; academic abstracts; HEC journals

INTRODUCTION

Publication in high-impact international journals shapes academic careers, institutional rankings, and the global circulation of knowledge (Hyland, 2009, 2015). Within this competitive environment, the research abstract functions as a critical gatekeeping genre: a compressed text that must simultaneously establish a study's significance, methodology, and contribution within severe word constraints, while satisfying the stylistic expectations of editors and peer reviewers operating within elite discourse communities (Swales, 1990; Hyland, 2004). What distinguishes successful abstracts from rejected ones is rarely a matter of basic grammatical accuracy. Applied linguistics research consistently demonstrates that publication failure frequently persists even in manuscripts free of surface-level errors, pointing instead to deficiencies in the strategic management of information structure, clausal balance, and register-appropriate distance (Biber & Gray, 2016; Flowerdew, 2013; Swales, 2004).

Central to this strategic management is the deliberate use of non-canonical syntactic structures — configurations that depart from the default Subject-Verb-Object word order of English (Biber et al., 1999; Quirk et al., 1985). Three such structures are of particular significance in academic prose: prototypical passive voice, it-extraposition, and Subject-Auxiliary Inversion (SAI). These are not optional stylistic ornaments but purposeful information-packaging devices (Birner & Ward, 1998) grounded in Quirk et al.'s (1985) principles of End-Weight and End-Focus, which together hold that syntactically heavy and informationally new material should occupy the right periphery of the clause, preventing cognitive overload while assigning maximum salience to the proposition the writer most wishes the reader to register. Prince's (1981) Taxonomy of Assumed Familiarity adds the pragmatic dimension: effective discourse sequences information along a Given-before-New gradient, with non-canonical structures repositioning discourse-familiar material into backgrounded positions to reserve the high-salience clause-final slot for new propositional weight.

Despite sustained scholarly attention to academic discourse, a significant empirical gap persists at the intersection of corpus linguistics, information structure, and publishing-tier differentiation. Existing research has typically examined syntactic features in isolation, without systematically investigating how passive voice, it-extraposition, and SAI function as an integrated repertoire of information-packaging strategies across contrasting publication categories (Flowerdew, 2015; Lillis & Curry, 2010). No corpus-based study has yet quantified and compared this distributional pattern directly across high-impact W-category and national-tier Y-category research abstracts under Pakistan's Higher Education Commission (HEC) journal recognition system. The practical consequence of this gap is that regional scholars seeking to enter international publishing networks lack a concrete, empirically grounded model of the specific syntactic strategies that distinguish elite from national-tier academic prose.

This study investigates how publication category conditions the deployment of non-canonical syntax in research abstracts. Specifically, the study aims to:

- 1) Quantify and compare the frequency and distributional patterns of passive voice, it-extraposition, and SAI between W-category and Y-category abstracts;
- 2) Evaluate how publication category influences sentence weight distribution according to End-Weight and End-Focus principles;
- 3) Analyze how authors across both tiers manage information flow according to Given-before-New sequencing; and
- 4) Determine the extent to which macro-level register constraints condition these micro-level stylistic choices.

These objectives correspond to four research questions:

(RQ1) How do W-category and Y-category abstracts differ in the frequency and distributional patterns of passive voice, it-extraposition, and SAI?

(RQ2) How does publication category affect how these structures satisfy End-Weight and End-Focus?

(RQ3) How does deployment reflect Given-before-New sequencing under Prince's (1981) framework?

(RQ4) To what extent do register constraints condition these stylistic choices?

The study adopts a complementary qualitative and quantitative corpus design. A purposively sampled corpus of 80 abstracts (40 W-category, 40 Y-category; 15,064 tokens) drawn from HEC-recognized Pakistani linguistics and literature journals (2020–2026) was analyzed in AntConc 3.5.9 through regular-expression concordance queries and manual structural verification, with frequencies normalized per 1,000 words. Qualitative analysis interpreted each verified instance through Quirk et al.'s (1985) End-Weight/End-Focus principles, Prince's (1981) Taxonomy of Assumed Familiarity, and Biber et al.'s (1999) functional-register framework. Full methodological detail is provided in Section 3.

The remainder of the article proceeds as follows. Section 2 reviews relevant literature and establishes the research gap. Section 3 details the corpus, data collection, and analytical procedures. Section 4 presents the quantitative and qualitative findings with discussion. Section 5 concludes, outlines the study's contributions, pedagogical implications, and limitations, and proposes future research directions.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

The Functional-Register Framework (Biber et al., 1999)

Biber's functional-register model (Biber et al., 1999) holds that the distribution of grammatical features is never arbitrary but is a direct linguistic response to the situational and communicative demands of a specific register (Biber, 2012). Academic abstracts are situated at the extreme informational pole of Dimension 1 in Biber's (1988, 1995) Multi-Dimensional Analysis framework, producing a processing bottleneck that requires non-canonical packaging strategies to move heavy informational load toward the right clause periphery. The Longman corpus baseline establishes that agentless passives account for more than 85% of passive instances in academic prose, it-extraposition occurs at approximately 2.5 to 4.5 instances per 1,000 words, and SAI returns a near-zero frequency profile, since its affective, emphatic function belongs almost exclusively to involved or fictional registers (Biber et al., 1999, pp. 475, 915, 976). This baseline provides an objective metric against which W-category and Y-category deployment can be evaluated (Biber & Gray, 2016). Critically, Biber (1999, p. 485) identifies heavy reliance on passive in the absence of corresponding nominal elaboration as a feature of procedural rather than argumentative academic prose, a distinction directly relevant to the Y-category findings of this study.

The Structural-Processing Framework (Quirk et al., 1985)

Quirk et al.'s (1985) comprehensive grammar provides the internal logic for how English sentences are organized to facilitate cognitive processing. The Principle of End-Weight holds that syntactically heavy constituents must be postponed toward clause-final position to prevent 'top-heaviness,' while the Principle of End-Focus holds that the most informationally salient material must occupy that same slot to receive maximum cognitive prominence (Quirk et al., 1985, pp. 1361–1362). End-Focus operates as the primary processing rule, while End-Weight functions as a secondary, compounding tendency. This framework provides the precise diagnostic terminology needed to determine whether a given non-canonical structure reflects genuine processing necessity or superficial stylistic preference, operationalized in the present study through a canonical-transformation test that calculates the word-count ratio between each matrix predicate and its corresponding extraposed or passivized constituent.

The Cognitive-Pragmatic Framework (Prince, 1981)

Prince's (1981) Taxonomy of Assumed Familiarity rejects a simple Old-versus-New binary, instead categorizing information into New (Brand-new or Unused), Inferable, and Evoked status. Successful discourse sequences sentence-initial positions with discourse-familiar or inferable material while reserving sentence-final position for genuinely new, cognitively demanding content. Within this framework, non-

canonical structures function as cognitive runways: an extraposed or passivized clause-initial slot offers the reader a familiar structural anchor before introducing brand-new propositional content. The pragmatic lens explains why two texts can be equally grammatical yet differ sharply in how cohesively they guide an expert reader through dense disciplinary content, and provides the framework through which this study diagnoses the specific information-packaging failures visible in Y-category writing.

Empirical Research on Non-Canonical Syntax and Publishing Tiers

A substantial empirical literature documents the functional motivation of non-canonical structures in academic and learner writing. Section-specific genre research (Carter-Thomas & Rowley-Jolivet, 2001; Szczygłowska, 2025; Prado-Alonso, 2011) establishes that abstracts in particular require exceptional informational density within severe word limits, making them the most demanding site for non-canonical deployment. Learner corpus research (Callies, 2005; Larsson, 2017; Looock, 2023) documents that developing writers exhibit a characteristic pattern of structural over-use without functional integration, applying formal register markers at high frequency without the precision that distinguishes expert from novice prose. Yoon's (2019) concept of lexico-grammatical 'teddy bears' demonstrates that developing writers lock onto frozen formulaic structures (it can be concluded that, it has been found that) to signal academic formality, producing a flat, monotone stance that fails to satisfy advanced stylistic expectations. Neither the section-specific genre research nor the learner corpus literature, however, has directly compared how high-impact and national-tier publications deploy passive voice and it-extraposition as an integrated information-packaging repertoire within the abstract genre specifically.

Research Gap

The reviewed literature establishes that non-canonical structures are functionally motivated information-packaging devices governed by register constraints (Biber et al., 1999), structural processing principles (Quirk et al., 1985), and pragmatic familiarity assumptions (Prince, 1981), and that developing writers frequently exhibit formulaic over-reliance on a narrow set of these structures (Yoon, 2019). However, no corpus-based study has directly stratified an academic corpus according to HEC institutional publishing tier as the primary comparative variable, nor examined passive voice and it-extraposition jointly as an integrated repertoire within the compressed abstract genre. The present study fills this gap by combining corpus-assisted quantitative measurement with close qualitative analysis grounded in a unified three-framework theoretical model, applied specifically to the abstract as a discrete analytical site.

METHODOLOGY

Research Design

This study adopts a complementary qualitative and quantitative corpus-assisted design (Creswell & Creswell, 2018). The quantitative strand captures structural frequencies and baseline distributional patterns across the dataset, while the qualitative strand serves as the primary analytical engine for interpreting the contextual and stylistic motivations behind those patterns (Sandelowski, 2000); the two strands operated sequentially and descriptively rather than through deep systemic convergence (Tashakkori & Teddlie, 2010). The corpus linguistics framework moves stylistic analysis away from impressionistic readings into verifiable, systematic observation (McEnery & Hardie, 2011), enabling the objective tracking of syntactic patterns that are frequently invisible to macroscopic qualitative reading alone.

Corpus and Sampling

The target population comprised research-article abstracts published in HEC-recognized Pakistani journals in Linguistics and Literature between 2020 and 2026. A purposive sampling technique (Patton, 2015; Palinkas et al., 2015) was used to construct a balanced corpus of 80 abstracts divided evenly into two sub-corpora: 40 W-category (high-impact, internationally indexed; 7,191 tokens) and 40 Y-category (national-tier, HEC-recognized; 7,873 tokens), yielding a combined corpus of 15,064 tokens. All selected abstracts were restricted to Linguistics and Literature to control for cross-disciplinary register variation, and only complete, officially published abstracts accompanying full peer-reviewed research articles were included. W-category texts were drawn from SAGE Journals and comparable international platforms; Y-category texts were drawn from official HEC-recognized national journal hosting sites.

Data Processing and Analytical Tools

Corpus texts were compiled into plain-text sub-corpus files and analyzed in AntConc (Version 3.5.9) using tested regular-expression queries, with grammatical categorization supported by the CLAWS part-of-speech tagging pipeline. Passive voice was identified using the be + past participle formula, restricted to prototypical be-passives (base, perfect-aspect, and modal forms); stative or adjectival passives were excluded. This search yielded 83 verified passive instances in the W-corpus and 134 in the Y-corpus. For it-extraposition, the search term it was queried, yielding 23 concordance hits in W-category and 44 in Y-category; after rigorous manual verification filtering out cataphoric, referential, and other non-extraposition uses of it, 3 true instances were confirmed in W-category and 9 in Y-category. A pilot scan for Subject-Auxiliary Inversion yielded zero true instances in either sub-corpus, confirming the strict register restrictions imposed by the abstract genre (Biber, 1988). Raw frequencies were normalized per 1,000 words to enable fair comparison across the two unevenly sized sub-corpora.

Qualitative Analytical Framework

Each verified instance was evaluated through three complementary theoretical lenses applied in parallel. Quirk et al.'s (1985) End-Weight and End-Focus principles were operationalized through a canonical-transformation test: each extraposed instance was rewritten in its canonical form to assess whether structural weight genuinely demanded the construction, and word-count ratios between matrix predicate and extraposed clause were calculated. Prince's (1981) Taxonomy of Assumed Familiarity tracked the familiarity status of clause-initial and clause-final material across each instance. Biber et al.'s (1999) register framework assessed how constructions were leveraged to satisfy institutional expectations of objectivity, agent erasure, and epistemic stance calibration, with each instance additionally mapped onto Swales and Feak's (2009) five-move abstract structure to determine its rhetorical move.

For passive voice, a nested purposive sampling design (Collins et al., 2007; Onwuegbuzie & Leech, 2007) selected six representative sentences per sub-corpus for intensive qualitative analysis, following selection controls of Miles and Huberman (1994) based on structural configuration, verb semantic category, and rhetorical move placement. Modal passives were analyzed separately to examine how auxiliary operators calibrate epistemic and deontic stance across different rhetorical moves.

Reliability

A randomly selected 20% sample of concordance lines, drawn proportionally from both sub-corpora, was independently reviewed by a second coder against the structural diagnostic criteria of Biber et al. (1999) and Quirk et al. (1985). Overall inter-rater agreement reached 85% (87% for be-passive variants, 83% for it-extraposition structures), within the range considered highly acceptable for manual corpus coding

(Krippendorff, 2004; Neuendorf, 2002). Disagreements, concentrated in complex clausal boundaries and adverbial modifier scope within dense abstract prose, were resolved through collaborative discussion and re-application of the structural criteria before finalizing the dataset.

FINDINGS AND DISCUSSION

It-Extraposition: Frequency, Structure, and Functional Precision (RQ1, RQ2)

Quantitative analysis (Table 1) reveals a striking disparity in it-extraposition deployment between the two corpora.

Table 1. Comparative quantitative summary of it-extraposition

Metric	W-Category	Y-Category	Difference
Total corpus tokens	7,191	7,873	+682 (Y larger)
Raw it-extraposition count	3	9	+6 (Y higher)
Normalized freq. (per 1,000 words)	0.42	1.14	~3× higher (Y)
Structural subtypes	2	8	Y more varied
Primary abstract move	Conclusion (66.7%)	Results (66.6%)	Functional divergence
Average extraposed clause length	24–33 words	11–15 words	W substantially heavier
End-Weight: structurally necessary?	Yes, all 3 instances	Partial — 2 of 9 stylistic only	W functionally precise

Despite being only slightly larger, the Y-category corpus shows a normalized extraposition frequency approximately three times that of W-category (1.14 vs. 0.42 per 1,000 words), with substantially greater structural variety across eight distinct subtypes compared to two in W-category. Move-level analysis reveals a fundamental rhetorical divergence: W-category instances concentrate in the Conclusion move (66.7%), functioning as devices for synthesizing interpretive significance, while Y-category instances concentrate in the Results move (66.6%), functioning primarily as formulaic reporting templates.

Applying Quirk et al.'s (1985) canonical-transformation test reveals the critical distinction. All three W-category instances involve extraposed clauses of 24–33 words set against matrix predicates of two to three words — asymmetries so severe that canonical alternatives are effectively unreadable. The W-11 instance, *“It is upheld that her ingenious hybridity originates from her position as an off-center postcolonial female writer who critiques colonial stereotypes and mimicry by revisiting/challenging colonial discourse,”* places a 24-word theory-dense clause against a two-word matrix predicate; the W-10 instance, *“It is concluded that avid texters, while appearing to greatly deviate from more traditional, standard written English, are a rich source for studying creative and spontaneous language adaptation,”* deploys a 33-word clause with an embedded concessive. In contrast, two Y-category instances (Y-20: *“It has been observed that syntax is the core component of linguistics,”* 7-word clause; Y-29: *“It is evident from the proposed study that the majority of the seminary students have a positive attitude,”* 11-word clause) are demonstrably driven by

stylistic formality rather than processing necessity, since their canonical alternatives remain fully readable. This directly addresses RQ2: W-category exploits End-Weight and End-Focus as functional processing tools; Y-category partially achieves the formal effect without consistently serving its cognitive rationale.

Prince's (1981) framework reveals a parallel divergence in Given-New management: W-category extrapositions consistently introduce Brand-New or Discourse-New propositional content — fresh theoretical arguments and macro-syntheses not previously established in the abstract — with the dummy it functioning as a genuine cognitive runway. Y-category extrapositions predominantly carry Inferred or Situationally Evident information, with the dummy it functioning as a conventional reporting frame rather than a cognitive scaffolding device.

Passive Voice: A Shared Resource, Differently Calibrated (RQ1, RQ3)

Passive voice quantitative results (Table 2) initially appear to favour the Y-category corpus, which deploys passives 47.5% more frequently (17.02 vs. 11.54 per 1,000 words).

Table 2. Comparative quantitative summary of passive voice

Metric	W-Category	Y-Category	Difference
Total passive count	83	134	+51 (Y higher)
Normalized freq. (per 1,000 words)	11.54	17.02	+47.5% (Y)
Agentless passives (% of total)	90.4%	91.0%	Comparable; matches Biber baseline
Modal passives (% of total)	7.2%	3.7%	W higher
Perfect passives (% of total)	9.6%	14.2%	Y higher
Modal passive distribution	Move 1 & 5: epistemic hedging	Move 2 & 3: rhetorical enforcement	Critical functional divergence
Tense consistency	Consistent past/present mapping	Frequent oscillation within abstracts	Y register violation

Both corpora align with Biber et al.'s (1999) global baseline regarding agentless passive dominance (90.4% W, 91.0% Y) and be-passive primacy, but diverge sharply in modal passive distribution (7.2% vs. 3.7%) and functional placement. W-category modal passives concentrate in the speculative perimeters of the abstract (Move 1 and Move 5), where low-probability epistemic hedges such as *may* be associated and *may* be triangulated signal disciplined claim-bounding. Y-category modal passives concentrate instead in the empirical core (Move 2 and Move 3), where high-certitude auxiliaries such as *cannot* be ignored and can be clearly observed function as tools of rhetorical enforcement rather than negotiated claim-bounding.

Qualitative analysis further reveals that Y-category abstracts exhibit systematic tense inconsistency within individual texts, oscillating between simple present and past tense without principled motivation. A representative Y-corpus abstract deploys *is sought*, *is observed*, and *is noted* to report completed findings

(specific numerical decreases in a finished experimental study), creating an ambiguity in the epistemic status of claims that disrupts the Given-New chain: the reader cannot reliably track whether passivized propositions present new claims or recap established findings. This pattern is absent from the W-category corpus, where past passives consistently mark completed methodological procedures and present passives mark claims with ongoing theoretical relevance. Additionally, the Y-corpus exhibits several instances of syntactically non-standard dangling participial passives that introduce Brand-New inferential gaps without structural support, a failure of the information-packaging function the passive is theoretically capable of serving (Prince, 1981).

Consolidated Cross-Corpus Analysis (RQ3, RQ4)

Tables 3 and 4 consolidate the cross-corpus comparative evidence across all analytical dimensions, directly addressing RQ3 and RQ4.

Table 3. Consolidated cross-corpus summary: it-extraposition — all analytical dimensions

Analytical Dimension	W-Category	Y-Category
Freq. (per 1,000 words)	0.42 — restrained, selective	1.14 — ~3× higher, formulaic
Structural variety	2 subtypes: passival + adjectival	8 subtypes: passival, modal, perfect, adverbial, prepositional, infinitival
Extraposed clause length	24–33 words; canonical alternative unreadable in all 3 cases	11–15 words; 2 of 9 instances (Y-20, Y-29) stylistically driven, not structurally necessary
Prince information status	Consistently Brand-New / Discourse-New content at focus position	Predominantly Inferrable / Expected; dummy it as reporting convention
Matrix verb selection	High-argument verbs: vital, upheld, concluded	Routine reporting verbs: found, observed, determined, revealed, examined
Register deployment	Precise functional tool among several strategies	Primary surface signal of academic register identity

Table 4. Register function mapping: passive voice and modal distribution by tier

Register Feature	W-Category Function	Y-Category Function
Modal passive placement	Move 1 & 5 (perimeters); low-probability hedges: may be associated, may be triangulated	Move 2 & 3 (empirical core); high-certitude enforcement: cannot be ignored, can be clearly observed
Passive clustering	Functionally differentiated across moves; varied matrix verbs	Formulaic clustering in Results move: it has been found, it was observed, it was determined

Tense-move mapping	Consistent: past = methodology, present = claims/conclusions	Oscillation within single abstracts; present tense used to report completed past findings
Given-New management	Brand-New in focus position; Evoked/Inferable at sentence-initial slot	Brand-New entities introduced without discourse anchoring; Inferable content in focus position
Overall pattern	Register mastery: functional integration of form and purpose	Register mimicry: surface conformity at higher frequency without functional depth

The consolidated picture is unambiguous. SAI yielded zero true instances across all 80 abstracts in both corpora, a finding interpreted not as a limitation but as an empirically significant confirmation that the compressed abstract genre imposes a categorical register barrier against this structure (Biber et al., 1999, p. 915). For the two structures that do appear, the evidence converges on a single overarching conclusion: Y-category writers are not less syntactically active than W-category writers; if anything, they deploy passive voice and it-extraposition at substantially higher raw and normalized frequencies. What distinguishes the two tiers is not the presence or frequency of non-canonical structures but the functional precision with which they are deployed — whether extraposition reflects genuine processing necessity, whether passive subjects and clause-final slots honour the Given-before-New contract, and whether register-marking structures are integrated into their communicative context or reproduced as formulaic surface conventions. This pattern is best characterized as register mimicry: Y-category writers reproduce the register's most visible surface conventions at higher frequency without full functional integration into the communicative purposes those conventions are designed to serve. This directly addresses RQ4 and constitutes the study's central theoretical contribution.

Synthesis

Taken together, the quantitative and qualitative findings overturn the naive interpretation that higher frequency of non-canonical structures signals greater syntactic sophistication. W-category writers deploy passive voice and it-extraposition with structural necessity, functional precision, and move-appropriate stance calibration. Y-category writers deploy the same structures at higher frequency but with shorter extraposed clauses, lower-efficiency Given-New management, formulaic matrix-verb clustering, tense inconsistency, and modal passives positioned in the empirical core rather than the speculative perimeters. The distinction between the two tiers is a distinction of functional deployment rather than grammatical knowledge, and this finding has direct implications for how EAP and ERPP curricula should conceptualize and teach non-canonical syntax.

CONCLUSION

This study investigated how publication category conditions the deployment of it-extraposition and passive voice in research abstracts, comparing 40 W-category and 40 Y-category HEC-recognized Pakistani linguistics and literature abstracts through a complementary quantitative and qualitative corpus design. The overarching finding overturns the naïve interpretation that higher frequency signals greater syntactic sophistication: Y-category writers deploy passive voice 47.5% more frequently (17.02 vs. 11.54 per 1,000 words) and it-extraposition roughly three times more frequently (1.14 vs. 0.42 per 1,000 words) than W-category writers, yet the qualitative analysis consistently shows that W-category writers deploy both structures with greater functional precision, structural necessity, and communicative integration.

Addressing RQ1, SAI yielded zero true instances across all 80 abstracts, confirming a categorical register barrier. For the two structures that do appear, Y-category writers show greater raw frequency and structural variety in it-extraposition (eight subtypes vs. two) but substantially shorter extraposed clauses (11–15 words vs. 24–33 words), with two instances demonstrably driven by stylistic formality rather than processing necessity. Addressing RQ2, W-category instances satisfy End-Weight and End-Focus in every case, with canonical alternatives being effectively unreadable; Y-category instances partially achieve the formal effect without consistently serving its cognitive rationale. Addressing RQ3, W-category extrapositions consistently introduce Brand-New Discourse-New content, with the dummy it functioning as a genuine cognitive runway; Y-category extrapositions predominantly carry Inferred or Expected content, with it functioning as a genre-conventional reporting frame. Addressing RQ4, W-category modal passives cluster in the speculative perimeters of the abstract as disciplined epistemic hedges, while Y-category modal passives cluster in the empirical core as tools of rhetorical enforcement; Y-category abstracts additionally exhibit systematic tense inconsistency and dangling participial constructions that disrupt the Given-New chain, patterns entirely absent from the W-category corpus.

These four strands of evidence converge on a single theoretical conclusion: the syntactic gap between W-category and Y-category abstract writing is a gap of functional deployment rather than grammatical knowledge, and is most accurately characterized as register mimicry. Y-category writers demonstrate clear awareness of non-canonical structures and deploy them at higher frequency, confirming that the barrier to international publication is not structural ignorance. The barrier is the gap between knowing that a structure signals academic formality and knowing precisely when, why, and how to deploy it so that it genuinely serves cognitive, pragmatic, and register demands.

Three pedagogical implications follow directly. EAP and ERPP curricula must move beyond teaching passive voice and it-extraposition as markers of formal register and explicitly incorporate Quirk et al.'s (1985) End-Weight and End-Focus as diagnostic frameworks, training writers to calculate the matrix-to-clause ratio of any extraposition candidate before deploying it. Writing workshops should implement Given-New sequencing revision protocols, training writers to audit the familiarity status of clause-initial and clause-final material across every passive and extraposed construction. Formulaic passive clustering must be explicitly deconstructed, with writers offered functionally differentiated alternatives that achieve equivalent impersonality without structural monotony, alongside a tense-to-move mapping protocol addressing the tense oscillation identified in the Y-category corpus.

Several limitations qualify these conclusions. The corpus is restricted to 80 abstracts in Linguistics and Literature, limiting generalizability to STEM and other disciplines. The zero-finding on SAI reflects the genre constraints of the abstract specifically and may not hold across full research-article sections. The study documents what Y-category writers do but cannot explain why, since no first-language transfer data was collected. Future research should replicate this methodology across STEM and social science corpora to test whether the register-mimicry pattern is discipline-specific; extend analysis to full research articles with section-by-section comparison to determine whether the functional deployment gap persists or intensifies outside the abstract's extreme informational density; and undertake contrastive interlanguage analysis pairing these corpus findings with first-language Urdu and Punjabi text samples from the same authors to determine whether the Given-New sequencing violations identified here reflect L1 transfer, insufficient exposure to target-register models, or pedagogical gaps in formal academic writing instruction.

REFERENCES

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

- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.
- Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.
- Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1), 2–20.
- Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- Birner, B. J., & Ward, G. (1998). *Information status and noncanonical word order in English*. John Benjamins.
- Callies, M. (2005). *Information highlighting in advanced learner English: The syntax-pragmatics interface in second language acquisition*. John Benjamins.
- Carter-Thomas, S., & Rowley-Jolivet, E. (2001). Syntactic differences in oral and written scientific discourse: The role of information structure. *ASp*, 31–34, 19–37.
- Collins, K. M. T., Onwuegbuzie, A. J., & Jiao, Q. G. (2007). A mixed methods investigation of mixed methods sampling designs in social and health science research. *Journal of Mixed Methods Research*, 1(3), 267–294.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). SAGE Publications.
- Flowerdew, J. (2013). Some thoughts on English for research publication purposes (ERPP) and related issues. *Language Teaching*, 48(2), 250–262.
- Flowerdew, J. (2015). Some thoughts on English for research publication purposes (ERPP): Issues and challenges. *Journal of English for Academic Purposes*, 18, 95–101.
- Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. University of Michigan Press.
- Hyland, K. (2009). *Academic discourse: English in a global context*. Continuum.

- Hyland, K. (2015). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing*, 31, 58–69.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). SAGE Publications.
- Larsson, T. (2017). A constructional approach to topicalization in advanced learner writing. *Journal of English Linguistics*, 45(2), 116–139.
- Lillis, T., & Curry, M. J. (2010). *Academic writing in a global context: The politics and practices of publishing in English*. Routledge.
- Loock, R. (2023). *Corpus-based translation studies: Research and practice*. Bloomsbury.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McKim, C. A. (2017). The value of mixed methods research: A mixed methods study. *Journal of Mixed Methods Research*, 11(2), 202–222.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). SAGE Publications.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. SAGE Publications.
- Onwuegbuzie, A. J., & Leech, N. L. (2007). A call for qualitative power analyses. *Quality & Quantity*, 41(1), 105–121.
- Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), 533–544.
- Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). SAGE Publications.
- Prado-Alonso, C. (2011). *Full-verb inversion in written and spoken English*. Peter Lang.
- Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–255). Academic Press.
- Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. In W. C. Mann & S. A. Thompson (Eds.), *Discourse description* (pp. 295–325). John Benjamins.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.
- Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in Nursing & Health*, 23(3), 246–255.
- Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Swales, J. M., & Feak, C. B. (2009). *Abstracts and the writing of abstracts*. University of Michigan Press.

Szczygłowska, T. (2025). Part-genre variation in research article sections: A corpus-based study. *Journal of English for Academic Purposes*, 73, 101–119.

Tashakkori, A., & Teddlie, C. (2010). *SAGE handbook of mixed methods in social & behavioral research* (2nd ed.). SAGE Publications.

Yoon, H. (2019). Lexico-grammatical ‘teddy bears’ in academic writing: Formulaic language and second-language writer development. *Journal of Second Language Writing*, 44, 30–42.