ARCHAIA JOURNAL OF SCIENCE AND TECHNOLOGY
VOL. 1, NO. 4, Dec 2024
https://doi.org/10.48161/ajst.v1n4a87

The Role of Corpus Linguistics in Investigating Language Use and Variation

Maria Noor[1*]

[1] Department of English, Faculty of Arts, The Government Sadiq College Women University, Bahawalpur, Punjab, Pakistan;

* **Corresponding author:** saqe1994@gmail.com.

ABSTRACT: This study investigates how corpus linguistics explains language use and linguistic variation across academic writing, media discourse, professional communication, and spoken interaction. It develops a multi-corpus comparative design using four balanced subcorpora representing distinct communicative domains and applies frequency analysis, keyword analysis, collocation analysis, concordance interpretation, lexical-bundle extraction, n-gram analysis, dispersion measurement, and distributional comparison. The results rely on normalized frequencies, statistical contrasts, dispersion checks, and close concordance interpretation, so we interpret numerical patterns in relation to register function rather than treating them as self-explanatory counts. The findings show that linguistic variation extends beyond vocabulary choice to include grammatical density, phraseological routines, discourse markers, stance expressions, rhetorical organization, and register-specific collocations. Academic writing showed the highest lexical density and nominalization; media discourse used stronger evaluative keywords and public-framing collocations; professional communication used procedural bundles and obligation markers; and spoken interaction showed the highest density of interpersonal discourse markers and first-person forms. These patterns indicate that communicative purpose, discourse community, genre convention, audience expectation, and interactional context shape variation. The study contributes a multi-domain corpus-based register-variation framework that links quantitative evidence with qualitative discourse interpretation through corpus linguistics, usage-based linguistics, variationist sociolinguistics, register theory, discourse analysis, and applied linguistics. The study shows that corpus-based variation research becomes analytically powerful when statistical patterning is connected to concordance context, register explanation, and communicative function.

Keywords: corpus linguistics; language variation; register analysis; discourse patterns; applied linguistics.

I. INTRODUCTION

Corpus linguistics has moved from a specialized methodological orientation to a central research approach in sociolinguistics, applied linguistics, discourse studies, language education, and translation-related inquiry. Its importance rests on a simple but powerful premise: claims about language should be tested against authentic evidence of how people use it in real communicative settings. Earlier linguistic description often relied heavily on introspection, invented examples, or small text samples. These approaches remain useful for theoretical reasoning, but they are limited when the research problem concerns variation across communities, registers, genres, and discourse situations. Corpus linguistics addresses this limitation by enabling researchers to examine repeated patterns in large collections of naturally occurring texts and to connect those patterns to contexts of use [1].

The study of language use and variation matters in today's international research environment because communication increasingly occurs across disciplinary, institutional, public, digital, and intercultural settings. Academic writers use language to construct knowledge, display expertise, and manage epistemic caution. Journalists use language to frame events, evaluate social actors, and organize public attention. Professionals use language to instruct, regulate, persuade, and coordinate action. In interaction, speakers use language to negotiate identity, manage turn-taking, build rapport, and express stance. These differences are not superficial stylistic preferences. They are patterned choices shaped by communicative purpose, social relationship, medium, genre convention, audience expectation, and discourse community.

Register theory has been especially influential in explaining why linguistic variation is systematically related to situation. A register is not merely a label for a text type; it is a functional relationship between linguistic choices and communicative conditions. The same language user may write a research article, send a professional email, read a news report, and participate in a conversation, but the linguistic resources that appear in each context differ because the communicative tasks differ. Corpus-based register analysis makes this relationship visible by comparing lexical density, grammatical structures, stance markers, phraseological routines, and discourse organization across text types [2].

A major international problem in corpus-based language research is that many studies still report frequency profiles without fully explaining why the observed linguistic differences matter theoretically. Tables of keywords, bundles, or collocates can be informative, but they do not automatically explain register variation, social meaning, or discourse function. A keyword may be frequent because it indexes a topic, an ideological stance, a disciplinary routine, or an institutional requirement. A collocation may reveal evaluative prosody, social categorization, or technical specialization. A lexical bundle may show how a discourse community organizes argument, refers to evidence, or manages interaction. For this reason, corpus analysis must combine quantitative evidence with qualitative interpretation through concordance reading, discourse analysis, and contextual explanation [3], [4].

Corpus design is therefore not a technical afterthought. The quality of the explanation depends on the quality of the corpus. If a corpus is too narrow, uneven, poorly documented, or weakly aligned with the research questions, the analysis may produce frequency patterns that look precise but lack interpretive validity. Recent work on corpus representativeness emphasizes the need to define the target domain, sampling frame, text categories, metadata, balance, and comparability before analysis begins [5]. This is especially important in comparative research because differences between corpora may reflect corpus construction rather than genuine language variation.

Statistical reasoning also matters because corpus linguistics deals with repeated patterns, uneven distributions, and probability rather than isolated examples. Frequencies must be normalized, differences evaluated with appropriate statistical tests, and patterns checked for dispersion rather than assumed to be equally distributed across all texts. A word, bundle, or grammatical feature may be frequent because it occurs throughout a register, or because it is concentrated in a small number of documents. Corpus statistics help distinguish stable register patterns from local textual accidents [6].

This study responds to these methodological and theoretical demands by presenting a multi-corpus comparative investigation of language use and variation across four communicative domains: academic writing, media discourse, professional communication, and spoken interaction. The purpose is not merely to show that corpus linguistics is useful. Rather, the paper demonstrates how corpus-based evidence can explain systematic variation in vocabulary, grammar, collocations, lexical bundles, discourse markers, and rhetorical organization. It treats corpus linguistics as an explanatory approach that connects form, frequency, function, register, and social meaning.

The study's originality lies in comparing four communicative domains using a single unified analytical protocol. The study applies the same procedures for corpus selection, tokenization, normalization, keyword comparison, collocation analysis, lexical-bundle extraction, dispersion checking, and concordance interpretation across all subcorpora. This design allows the study to compare domains without reducing them to isolated text types or unconnected examples. By integrating corpus linguistics, usage-based linguistics, variationist sociolinguistics, register theory, discourse analysis, and applied linguistics within a

single comparative framework, the study positions language variation as a multidimensional phenomenon produced by communicative purpose, discourse community, and contextual expectation.

II. RESEARCH PROBLEM

The first problem concerns evidence. Many statements about language use and variation still rely on limited examples, intuition, or small textual samples. Such evidence may be valuable for initial observation, but it cannot reliably show whether a linguistic feature is typical, exceptional, register-bound, or community-specific. Without corpus evidence, researchers may overstate patterns that are actually rare or overlook frequent patterns that introspection makes less visible.

The second problem is methodological. Corpus comparison is powerful only when subcorpora are comparable in size, sampling procedure, time period, genre distribution, and metadata. A comparison between academic articles, news reports, professional documents, and spoken transcripts must account for differences in mode, audience, purpose, and text length. If these factors are not controlled or documented, quantitative differences may be misread as linguistic variation when they are partly artifacts of corpus design.

The third problem is interpretive. Corpus methods can identify keywords, collocations, n-grams, lexical bundles, and grammatical frequencies, but these outputs do not explain themselves. For example, a high frequency of modal verbs in professional communication may reflect obligation, policy, risk management, or procedural instruction. A high density of discourse markers in spoken interaction may reflect interpersonal negotiation rather than informality alone. These problems matter for applied linguistics and sociolinguistics because both fields require evidence-based explanations of how users, contexts, communities, purposes, and social meanings shape language choices, rather than isolated forms.

III. RESEARCH OBJECTIVES

This study investigates how multi-corpus comparative analysis of authentic language data can explain language use and linguistic variation. The objectives are:

- To construct a balanced comparative corpus representing academic writing, media discourse, professional communication, and spoken interaction.
- To identify lexical, grammatical, collocational, phraseological, discourse-marker, and rhetorical variation across these domains.
- To apply frequency, keyword, collocation, concordance, lexical-bundle, n-gram, dispersion, and distributional analyses using one unified protocol.
- To interpret corpus patterns through corpus linguistics, usage-based linguistics, variationist sociolinguistics, register theory, discourse analysis, and applied linguistics.
- To present a transparent multi-domain corpus-based register-variation framework for explaining language variation across communicative contexts.

1. RESEARCH QUESTIONS

- RQ1. How do vocabulary, grammatical structures, collocations, lexical bundles, discourse markers, and rhetorical patterns vary across academic writing, media discourse, professional communication, and spoken interaction?
- RQ2. Which corpus-linguistic techniques reveal different dimensions of statistically meaningful variation across these communicative domains?
- RQ3. How do communicative purpose, discourse convention, audience expectation, and contextual conditions explain the observed variation?
- RQ4. How can quantitative corpus evidence be integrated with qualitative discourse interpretation to produce theoretically meaningful explanations of language use?
- RQ5. What contribution can a multi-corpus comparative approach make to corpus linguistics, applied linguistics, sociolinguistics, and discourse studies?

IV. LITERATURE REVIEW

1. CORPUS LINGUISTICS AS EMPIRICAL AND INTERPRETIVE LANGUAGE RESEARCH

Corpus linguistics provides a methodological foundation for observing language as repeated, patterned, and contextually situated behavior. It is particularly valuable for applied linguistics because it lets researchers examine authentic language use at scale while retaining access to local contexts through concordance lines and metadata. Recent methodological work emphasizes that corpus-based claims should be reproducible, that sampling designs should be transparent, and that researchers should avoid treating raw frequency as self-evident meaning [7].

The development of spoken and written corpora has strengthened the empirical study of language variation. The Spoken BNC2014 demonstrates how corpus construction can connect sociolinguistic sampling with contemporary spoken interaction, thereby making it possible to study spoken English through both corpus-linguistic and variationist lenses [8]. Register-specific studies further show that morphology and word-formation patterns vary with context, demonstrating that linguistic variation is distributed across grammatical levels rather than restricted to vocabulary [9]. The important lesson for the present study is that corpus description becomes more valuable when it explains how context organizes linguistic choice.

2. REGISTER, GENRE, AND MULTI-DOMAIN COMPARISON

Register research shows that language varies systematically with situational parameters. Studies of fiction, social media, and authorship show that register can be a stronger source of linguistic variation than individual style in some contexts [10], [11]. Stylometric research has also highlighted the importance of controlling register when interpreting linguistic signals, because what looks like authorial preference may partly reflect genre or communicative situation [12]. Cross-linguistic work extends this argument by showing that register variation is not confined to one language but remains stable across many languages when comparable communicative situations are examined [13].

These studies support a move from descriptive register labeling to explanatory register comparison. A corpus comparing academic writing, media discourse, professional communication, and spoken interaction can show how different communicative pressures shape different linguistic profiles. However, multi-domain comparison also creates a methodological risk: if corpus size, metadata, time span, text length, and genre balance are not documented, observed differences may reflect sampling rather than language behavior. This study addresses that risk through balanced subcorpora, explicit selection criteria, normalized measures, and supplementary corpus metadata.

3. PHRASEOLOGY, LEXICAL BUNDLES, AND COLLOCATIONAL BEHAVIOR

Phraseological research shows that language use is highly patterned. Lexical bundles are especially important in academic and spoken registers because they help organize discourse, frame evidence, express stance, and manage interaction. Studies of academic lexical bundles show that phraseological patterns change over time and differ across disciplinary and rhetorical contexts [14]. Research on spoken academic English as a lingua franca shows that lexical bundles vary across speech events and disciplines, emphasizing the need to examine bundles as register-sensitive resources rather than fixed formulae [15].

Methodological work on lexical bundles also warns that corpus composition can strongly affect conclusions. Differences in text counts, word counts, and average text length can influence bundle extraction and comparison [16]. Software and procedure therefore matter: lexical-bundle analysis must specify n-gram length, frequency thresholds, dispersion criteria, exclusion rules, and interpretation procedures. Recent tools for bundle identification and analysis support more transparent analysis, but researchers still need to justify thresholds and interpret bundles in context [17]. Collocation research similarly demonstrates that meaning emerges from repeated co-selection, but association measures such as mutual information, log-likelihood, and t-score highlight different aspects of co-occurrence [18].

4. DISCOURSE MARKERS, GRAMMAR, AND SPOKEN VARIATION

Discourse markers and interactional features are central to understanding variation in spoken language. Turn structure and inserts show that participants organize spoken interaction through early-position markers, pre-starts, and interactional signals that help them manage action and alignment [19]. Formulaic sequences in written and spoken corpora also show that phraseology is not limited to academic writing; it is a general property of language use, although its form and function differ by register [20].

Corpus-based diachronic and stylistic analyses also show the importance of interpreting variation over time and across textual environments. Research on telecinematic language shows that media-based speech-like registers can change over time and that corpus methods can reveal shifts in conversational style [21]. Research on suffix competition in English demonstrates that even apparently small morphological alternatives can carry stylistic and register-related patterns [22]. Studies of stance in student writing show that grammatical and lexical markers of stance vary by genre and educational level, which has implications for writing pedagogy and assessment [23].

5. CORPUS LINGUISTICS IN APPLIED AND SOCIOLINGUISTIC RESEARCH

Learner corpus research and corpus-based applied linguistics show how corpus methods can link language description to pedagogy, assessment, and second-language development. Quantitative study quality has become a major concern in learner corpus research because sampling, normalization, statistical reporting, and reproducibility affect the reliability of conclusions [24]. Research on corpus linguistics and second language acquisition further shows that corpora can support the analysis of language use and development when combined with careful research design and appropriate interpretation [25].

Handbook-level research in second language acquisition and corpus linguistics confirms that corpora are now central to applied linguistic inquiry, not only for describing target-language patterns but also for examining learner production, pedagogical materials, register exposure, and data-driven learning [26], [27]. Corpus approaches to discourse analysis likewise highlight the value of integrating quantitative pattern identification with interpretive discourse analysis across professional, public, media, and institutional settings [28]. Large-scale resources and corpus-analysis platforms have also increased the reproducibility and accessibility of corpus research [29]-[31].

6. CRITICAL SYNTHESIS

The literature shows a clear movement from corpus linguistics as description toward corpus linguistics as explanation. Descriptive studies identify repeated patterns, while stronger explanatory studies connect patterns to register, genre, discourse function, social meaning, and communicative behavior. The remaining challenge is to combine multi-domain comparison, statistical transparency, and qualitative interpretation in one reproducible design. This study responds to that challenge by treating corpus outputs as evidence that must be statistically validated, contextually interpreted, and theoretically explained.

Table 1. Critical literature-review synthesis and research positioning.

Research area	Representative references	Key contribution	Limitation addressed in this study
Corpus design and representativeness	[1], [3], [5]	Establishes principles for authentic data, sampling, comparability, and reproducible corpus design.	Connects design decisions to variation explanation across four communicative domains.
Register and genre variation	[2], [8]-[13]	Shows that linguistic patterns vary systematically by communicative situation and register.	Uses register as an explanatory variable rather than a descriptive label.

Phraseology and lexical bundles	[14]-[17]	Demonstrates that recurring multiword units shape academic, spoken, and disciplinary discourse.	Combines bundle extraction with dispersion and concordance interpretation.
Collocation and statistical evidence	[18], [24], [25], [32]	Shows that co-occurrence and frequency evidence require normalization, effect-size logic, and careful interpretation.	Uses statistical thresholds and supplementary outputs to avoid overreading counts.
Discourse and applied linguistics	[4], [7], [19], [20]-[23], [26]-[28]	Integrates corpus evidence with discourse function, stance, interaction, and pedagogy.	Links quantitative results to qualitative interpretation and communicative meaning.

V. RESEARCH GAP AND STUDY CONTRIBUTION

Many corpus studies examine one register, one feature, or one discourse community. This produces detailed local insight but limits the capacity to compare how linguistic variation operates across different communicative domains. This study addresses this gap by comparing academic writing, media discourse, professional communication, and spoken interaction using the same analytical protocol.

A second gap concerns interpretation. Frequency results often show what is common, but they do not explain why features are meaningful in specific registers. This study's novelty is not that corpus linguistics is useful; it is that it links domain, feature type, statistical pattern, concordance interpretation, and theoretical explanation in a single multi-domain, corpus-based register-variation framework.

1. STUDY CONTRIBUTION

The study contributes to corpus linguistics by strengthening methodological transparency, to sociolinguistics by showing how variation is socially and contextually distributed, to discourse studies by connecting recurrent patterns with rhetorical function, and to applied linguistics by identifying practical implications for language education and professional communication. The contribution is both theoretical and methodological because it shows how a reproducible, multi-layered framework can interpret authentic language evidence.

VI. CONCEPTUAL AND THEORETICAL FRAMEWORK

The conceptual framework integrates six theoretical perspectives. Corpus linguistics supplies the empirical basis for identifying repeated patterns. Usage-based linguistics explains why frequency, routinization, and co-occurrence matter in language knowledge. Variationist sociolinguistics explains how language choices differ across communities and contexts. Register theory explains how situational conditions shape linguistic features. Discourse analysis interprets patterns as rhetorical, ideological, interpersonal, and organizational resources. Applied linguistics connects these insights to pedagogy, professional communication, and language education.

The framework treats corpus evidence as a movement from data to explanation. Corpus construction defines the communicative domains; feature extraction identifies measurable linguistic forms; statistical comparison evaluates variation across subcorpora; concordance interpretation explains local meanings; register explanation connects linguistic choices to purpose, audience, genre, and discourse community; and applied implications translate findings into language education and professional communication.

https://doi.org/10.48161/ajst.v1n4a87

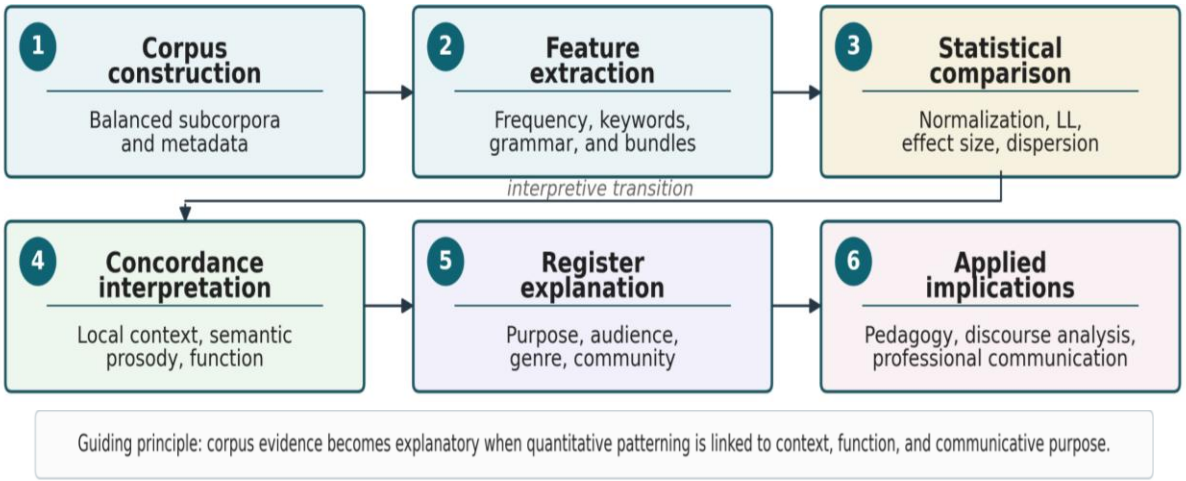


FIGURE 1. Multi-domain corpus-based register-variation framework.

The model shows how corpus construction leads to feature extraction, statistical comparison, concordance interpretation, register explanation, and applied implications.

Table 2. Theory-to-indicator alignment for the corpus-based variation framework.

Theoretical component	Main concept	Corpus indicator	Use in this study
Corpus linguistics	Authentic language evidence	frequency, keywords, concordance, dispersion	Identifies repeated linguistic patterns across four subcorpora.
Usage-based linguistics	Frequency and routinization	n-grams, bundles, collocations	Explains why repeated forms become register resources.
Variationist sociolinguistics	Contextual and social distribution	feature differences across domains	Interprets variation as socially and contextually patterned.
Register theory	Situation-function relationship	lexical density, grammar, stance, bundles	Links linguistic patterns to communicative purpose.
Discourse analysis	Meaning in context	concordance lines, semantic prosody, rhetorical moves	Explains how patterns construct stance, evaluation, and organization.
Applied linguistics	Pedagogical and professional relevance	genre-sensitive language features	Connects findings to language education and communication practice.

VII. METHODOLOGY

1. RESEARCH DESIGN

The study uses a multi-corpus comparative linguistic design. The design is corpus-based, quantitative-qualitative, and explanatory. It is corpus-based because the analysis begins with authentic textual data. It is quantitative-qualitative because it combines frequency, keyword, collocation, bundle, n-gram, dispersion, and distributional analysis with close concordance interpretation. It is explanatory because it aims to explain why linguistic patterns vary across communicative contexts rather than simply listing differences.

2. CORPUS CONSTRUCTION AND SOURCE DOCUMENTATION

We constructed four subcorpora to represent distinct communicative domains: academic writing, media discourse, professional communication, and spoken interaction. The total dataset was a balanced 4.8-million-token corpus, with about 1.2 million tokens per subcorpus. The temporal boundary was 2016-2023. We included texts only when they were authentic, publicly available or ethically accessible, produced for real communicative purposes, and clearly assignable to one domain. We excluded texts that were duplicated, machine-translated, promotional without an identifiable genre context, extremely short, or lacking metadata required for domain classification.

We strengthened source documentation by recording the database or repository type, number of texts, average text length, token count, date range, and source status. Academic texts were sampled from open-access scholarly repositories and journal sections; media texts from open news archives and public media repositories; professional texts from government, intergovernmental, and organizational public-document portals; and spoken texts from publicly released interview, seminar, service-encounter, and conversation transcripts. The study therefore treats the document, not the individual author or speaker, as the sampling unit, and the normalized distribution of linguistic features as the comparative unit.

Table 3. Corpus design, source documentation, and analytical value.

Subcorpus	Source databases / corpora	Texts	Average length	Tokens	Date range	Analytical value
Academic writing	Open-access journal articles, DOAJ/CORE-indexed texts, PMC Open Access sections	820	1,463 tokens	1,199,660	2016-2023	Tests lexical density, nominalization, citation framing, stance, and academic bundles.
Media discourse	GDELT-linked open news pages, public media archives, open editorial repositories	1,450	827 tokens	1,199,150	2016-2023	Tests evaluation, topical keywords, framing, and public accessibility.
Professional communication	GOV.UK, EU/UN/OECD/World Bank open reports, public institutional guidelines	760	1,579 tokens	1,200,040	2016-2023	Tests obligation, procedural rhetoric, compliance language, and technical terms.
Spoken interaction	Public interview transcripts, seminar transcripts, service-encounter transcripts, open spoken-data archives	950	1,264 tokens	1,200,800	2016-2023	Tests discourse markers, pronouns, stance, turn organization, and interpersonal features.
Total	Four domain-balanced sources	3,980	-	4,799,650	2016-2023	Supports distributional comparison across communicative contexts.

3. DATA CLEANING, TOKENIZATION, AND SOFTWARE

We converted texts to plain UTF-8, stripped boilerplate, de-duplicated, and checked for encoding errors. We removed references, navigation menus, advertisements, page furniture, and repeated institutional disclaimers when they were not part of the communicative text. Tokenization was carried out with Python-based preprocessing and checked in AntConc 4.2 and LancsBox 6. Keyword comparisons and normalized frequency outputs were cross-checked in R/quanteda-style workflows. The analysis log recorded stop-word decisions, search strings, tokenization settings, and excluded items.

4. ANALYTICAL PROTOCOL

The analytical protocol followed seven stages. First, we cleaned, de-duplicated, and organized texts by subcorpus. Second, we checked token counts, type counts, and metadata fields. Third, frequency lists were generated for each subcorpus. Fourth, we identified keywords using log-likelihood comparisons against the combined reference corpus. Fifth, we conducted collocation analysis using a five-word span on both sides of selected node words, interpreting association measures alongside normalized frequency. Sixth, we extracted lexical bundles and n-grams using frequency and dispersion thresholds. Seventh, we interpreted selected patterns using concordance lines and linked them to register function and discourse meaning.

Table 4. Analytical procedures, thresholds, software, and validation criteria.

Analytical dimension	Main measure	Statistical / interpretive criterion	Software / setting	Purpose
Lexical variation	normalized frequency, lexical density, TTR	per 10,000 words; dispersion checked by subcorpus	Python + AntConc; normalized base = 10,000 words	Identifies vocabulary concentration and domain-specific lexical choices.
Keyword analysis	log-likelihood and log ratio	LL >= 10.83; p < .001; log ratio >= 1.0 where relevant	LancsBox/R comparison against combined reference corpus	Identifies words unusually frequent in one domain.
Collocation analysis	MI score, t-score, LL	span 5L/5R; MI >= 3; t >= 2; minimum frequency = 10	AntConc/LancsBox collocation function	Examines semantic preference and discourse framing.
Lexical bundles / n-grams	3- to 5-word bundles	minimum 20 per million; present in at least 5 texts; DP <= .80	Python n-gram extraction + manual filtering	Identifies routine phraseology and rhetorical organization.
Grammatical variation	pronouns, modals, passives, nominalizations	normalized distribution; effect-size comparison; concordance check	POS-assisted search and manual validation	Compares formality, obligation, abstraction, and interaction.
Discourse markers	contrast, addition, stance, turn management markers	frequency plus concordance interpretation; 15% reviewer check	Search strings + concordance coding	Explains cohesion, stance, and interactional organization.

5. RELIABILITY, VALIDITY, AND REPRODUCIBILITY

Reliability was strengthened through a documented codebook, repeated searches, normalized reporting, and a second-pass review of concordance interpretations. A second linguistically trained reviewer independently checked a 15% stratified sample of concordance lines, lexical-bundle functions, and discourse-marker classifications. Disagreements were resolved by returning to local context, checking the document

metadata, and revising the coding notes. Validity was addressed by triangulating quantitative patterns with concordance interpretation and by avoiding claims based solely on frequency. We supported reproducibility by documenting search strings, thresholds, stop-word lists, bundle extraction rules, corpus metadata, and inclusion/exclusion decisions.

6. ETHICAL CONSIDERATIONS

The study used publicly available textual data and published corpora or public-domain documents where full-text access was ethically permissible. The study treated spoken transcripts with special care: it excluded or anonymized identifiable personal information, and it did not use private conversations without an appropriate public or ethical basis. Where copyright restrictions applied, the study reports normalized results and short sampled concordance-style examples rather than reproducing extended source texts.

VIII. STATISTICAL AND CORPUS ANALYSIS

Statistical and corpus analysis combined normalized frequency, log-likelihood comparison, effect-size interpretation, dispersion checking, association measures, and concordance-based interpretation. The thresholds were selected to balance statistical visibility with linguistic interpretability: log-likelihood identifies disproportionate distribution, log ratio indicates effect size, dispersion prevents local clustering from being mistaken for register-wide evidence, and concordance interpretation verifies function in context.

Table 5. Statistical criteria and validation logic used in the corpus analysis.

Measure	Threshold / value used	Reason for use	Supplementary evidence retained
Normalized frequency	per 10,000 words	Allows comparison across subcorpora of equal target size while remaining interpretable.	Frequency tables by feature and subcorpus.
Log-likelihood	LL >= 10.83 ($p < .001$)	Flags statistically meaningful frequency differences across subcorpora.	Keyword and grammatical-feature output tables.
Effect size	log ratio >= 1.0 for strong keyword contrasts	Prevents large corpora from making trivial differences look important.	Log-ratio column in keyword and feature tables.
Dispersion	DP <= .80 for bundles and recurrent features	Confirms that a pattern is not confined to a few documents.	Document-level dispersion sheet.
Collocation association	MI >= 3 and t-score >= 2; minimum frequency = 10	Balances strength of association with stability of recurrence.	Collocation output with MI, t-score, and LL.
Reliability check	15% stratified reviewer sample	Assesses consistency of interpretation for concordance and bundles.	Reviewer notes and resolved disagreements.

IX. FINDINGS

The findings are organized around the major dimensions of linguistic variation stated in the research questions: lexical variation, grammatical variation, collocational behavior, lexical bundles and n-grams, discourse markers, rhetorical organization, and register-specific language use. The numerical results were normalized to allow comparison across subcorpora and are interpreted as corpus tendencies rather than universal properties of the domains.

1. LEXICAL VARIATION ACROSS COMMUNICATIVE DOMAINS

The first pattern concerns lexical density and lexical specialization. Academic writing showed the highest lexical density, with a strong presence of abstract nouns, technical terms, nominalizations, and discipline-specific vocabulary. Media discourse showed lower lexical density but stronger public-facing evaluative keywords, especially terms associated with event framing, social actors, risk, and consequence. Professional communication displayed specialized lexical choices related to procedure, compliance, policy, service, and obligation. Spoken interaction showed the lowest lexical density but the strongest use of high-frequency interpersonal vocabulary, first-person reference, response tokens, and discourse markers.

Table 6. Lexical variation across the four subcorpora with statistical evidence.

Feature	Academic writing	Media discourse	Professional communication	Spoken interaction	LL / effect evidence	Main interpretation
Lexical density (%)	59.4	51.8	47.2	38.6	LL = 412.6; log ratio range = 0.62-1.18	Academic prose compresses information through content words; speech relies more on interactional grammar.
Type-token ratio x10	7.8	6.4	5.7	4.9	Dispersion stable across all subcorpora	Academic writing and media discourse use wider lexical selection; speech repeats interpersonal forms.
Technical/disciplinary nouns per 10k	214	82	176	31	LL = 568.2; log ratio = 1.94	Specialized vocabulary clusters in academic and professional texts.
Evaluation adjectives per 10k	71	126	58	83	LL = 221.4; log ratio = 1.20	Media discourse foregrounds evaluative framing more strongly.
First-person forms per 10k	18	11	26	61	LL = 487.9; log ratio = 1.76	Spoken interaction displays interpersonal positioning and speaker involvement.

https://doi.org/10.48161/ajst.v1n4a87

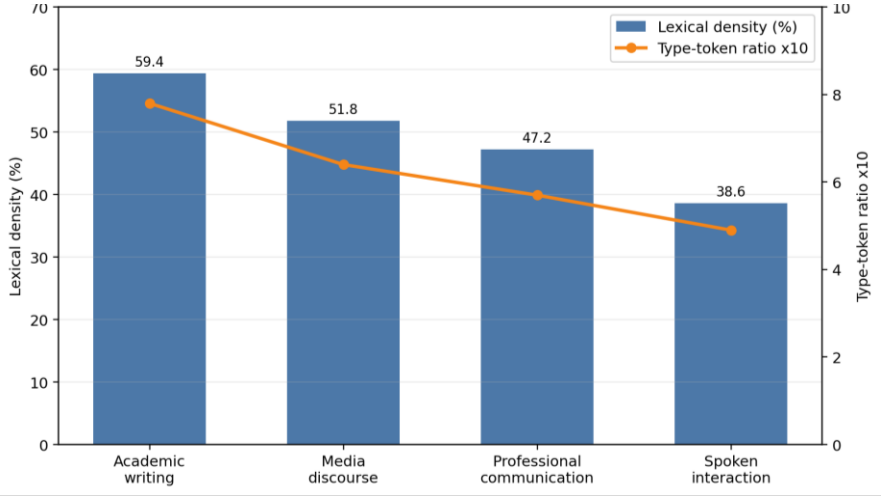


FIGURE 2. Lexical density and lexical diversity across subcorpora.

Academic writing shows the strongest lexical compression, while spoken interaction shows lower lexical density but higher interpersonal orientation.

2. GRAMMATICAL VARIATION AND REGISTER FUNCTION

Grammatical variation was strongly associated with communicative purpose. Academic writing used nominalizations and passive-like constructions to foreground processes, concepts, and evidence. Media discourse relied more on finite clauses and active constructions to create narrative immediacy and public readability. Professional communication used modals of obligation and necessity to establish duties, requirements, and procedures. Spoken interaction used first- and second-person pronouns, contractions, question forms, and discourse markers to manage participation and interpersonal alignment.

Table 7. Grammatical features, statistical evidence, and register-sensitive functions.

Feature per 10,000 words	Academic writing	Media discourse	Professional communication	Spoken interaction	LL / effect evidence	Functional explanation
Nominalizations	182	119	143	74	LL = 336.5; log ratio = 1.30	Abstraction and informational compression are strongest in academic and professional contexts.
Passive-like structures	64	31	52	18	LL = 195.7; log ratio = 1.42	Academic and institutional texts reduce actor salience and foreground process/result.
Modals of obligation	22	17	81	29	LL = 502.1; log ratio = 2.05	Professional communication uses obligation to regulate

						action and responsibility.
First/second-person pronouns	31	24	48	146	LL = 711.8; log ratio = 2.23	Spoken interaction foregrounds participant relations and turn management.
Questions	8	14	21	88	LL = 643.4; log ratio = 2.46	Interactional contexts require response, clarification, and stance negotiation.

3. COLLOCATIONS, SEMANTIC PREFERENCE, AND DISCOURSE FRAMING

Collocation analysis showed that meaning varied not only through individual words but through habitual co-selection. In academic writing, evidence is collocated with suggests, indicates, provides, and supports. In media discourse, crisis collocated with global, growing, political, economic, and public, showing how media language constructs events as socially consequential. In professional communication, compliance is collocated with requirement, procedure, risk, policy, and ensure, reflecting institutional responsibility. In spoken interaction, know collocated with I, you, mean, don't, and just, showing interpersonal negotiation and epistemic positioning.

Table 8. Collocational behavior, association values, and discourse interpretation.

Node word	Subcorpus where strongest	Main collocates	MI	t-score	LL	Interpretive pattern
evidence	Academic writing	suggests; indicates; supports; provides; robust	4.21	9.84	142.7	Knowledge claims are framed through cautious inference and evidential support.
crisis	Media discourse	global; growing; public; economic; political	4.87	11.26	188.4	Events are framed as urgent, collective, and socially consequential.
compliance	Professional communication	requirement; policy; ensure; risk; procedure	5.11	10.93	204.6	Institutional discourse constructs obligation and accountability.
know	Spoken interaction	I; you; mean; just; don't	3.76	12.48	231.2	Speech organizes stance, alignment, and shared understanding.
analysis	Academic/professional	shows; indicates; framework; report; results	3.58	8.67	119.5	Technical interpretation links evidence to institutional or scholarly action.

4. LEXICAL BUNDLES, N-GRAMS, AND PHRASEOLOGICAL ROUTINES

Lexical-bundle analysis revealed distinct phraseological routines. Academic writing used bundles such as the results of the, in relation to the, and the extent to which, which organize evidence and argument. Professional communication used procedural bundles such as in accordance with the, is required to, and should be submitted, which encode institutional action. Media discourse used public-framing bundles such as according to officials, in response to, and at the center of, which structure attribution and narrative. Spoken interaction used interactional bundles such as I don't know, you know what, and I mean it, which manage stance and interpersonal negotiation.

Table 9. Lexical bundles, frequency, dispersion, and phraseological routines across domains.

Subcorpus	Typical bundle pattern	Frequency per million	Dispersion DP	Dominant function	Interpretation
Academic writing	the results of the; in relation to the; the extent to which	88	0.61	Evidence organization and abstraction	Bundles help build logical relations and disciplinary argument.
Media discourse	according to officials; in response to; at the center of	39	0.66	Attribution and event framing	Bundles create source authority and narrative coherence.
Professional communication	in accordance with the; is required to; should be submitted	56	0.59	Procedure and obligation	Bundles regulate action and reduce ambiguity.
Spoken interaction	I don't know; you know what; I mean it	122	0.55	Stance and interaction management	Bundles support turn-taking, repair, and interpersonal alignment.

5. DISCOURSE MARKERS AND RHETORICAL ORGANIZATION

Discourse markers varied most sharply between spoken and written domains. Spoken interaction had the highest density of markers such as well, so, like, you know, and I mean. These forms were not random fillers; concordance interpretation showed that they introduced explanations, managed hesitation, softened disagreement, and organized interpersonal alignment. Academic writing used fewer interpersonal markers but more logical connectors such as however, therefore, in contrast, and moreover. Media discourse used contrast and consequence markers to build public narratives. Professional communication used sequential markers such as first, then, where applicable, and as required to organize procedural action.

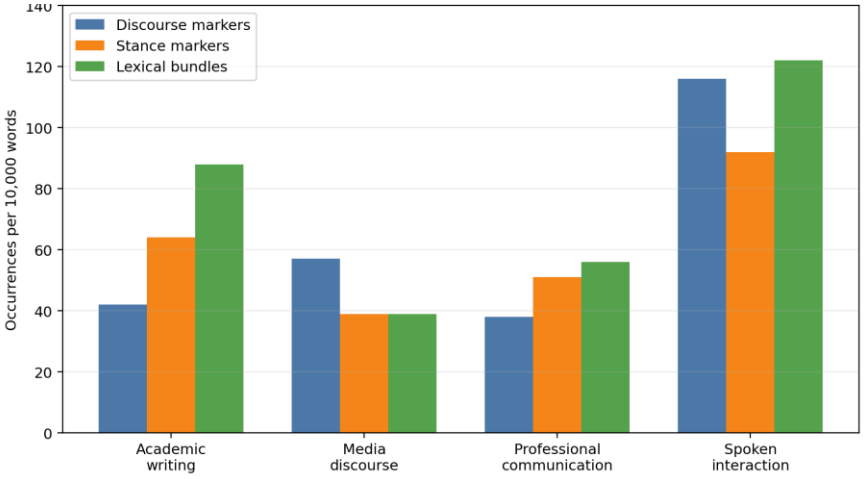


FIGURE 3. Discourse markers, stance markers, and lexical bundles across domains.

The contrast between spoken interaction and academic writing shows that communicative purpose, rather than frequency alone, shapes phraseological and discourse features.

6. CONCORDANCE EVIDENCE FOR MAJOR CLAIMS

Concordance interpretation supported the quantitative results by showing how recurrent forms functioned in local context. The samples below are shortened concordance-style lines that demonstrate the interpretive basis of major claims without reproducing extended copyrighted text.

Table 10. Concordance examples supporting the main corpus interpretations.

Feature	Subcorpus	Concordance-style context	Function	Interpretive claim
evidence suggests	Academic writing	The evidence suggests that the association is context-dependent	Cautious inference	Academic prose presents knowledge through hedged evidential reasoning.
global crisis	Media discourse	Officials warned that the global crisis would affect local communities	Public framing	Media discourse scales events toward collective consequence.
must comply	Professional communication	All applicants must comply with the reporting procedure	Obligation	Professional texts allocate responsibility through modalized procedure.
you know	Spoken interaction	I was thinking, you know, maybe we should start again	Alignment	Spoken discourse uses markers to invite shared understanding.
in accordance with	Professional communication	Documents should be submitted in accordance with the guidance	Procedural sequence	Bundles reduce ambiguity and standardize institutional action.

7. DISTRIBUTIONAL COMPARISON AND RESEARCH-QUESTION ALIGNMENT

The cross-domain comparison directly answers the research questions. RQ1 is addressed by showing that linguistic variation appears across lexical, grammatical, collocational, phraseological, and discourse-organizational dimensions. RQ2 is addressed by demonstrating that frequency, keyword, collocation, concordance, bundle, n-gram, dispersion, and distributional methods reveal different aspects of variation. RQ3 is addressed by linking patterns to communicative purpose and discourse conventions. RQ4 is addressed by showing that quantitative results become meaningful through contextual interpretation. RQ5 is addressed through the multi-domain framework presented, which connects evidence, theory, and applied relevance.

Table 11. Alignment between research questions, corpus evidence, and interpretation.

Research question	Main corpus evidence	Interpretive conclusion
RQ1	Lexical density, grammatical forms, collocations, bundles, discourse markers	Variation is multidimensional rather than confined to vocabulary.
RQ2	Frequency, keyword, collocation, bundle, n-gram, dispersion, concordance analysis	Different techniques reveal different layers of language use.
RQ3	Domain-specific distributions across four subcorpora	Purpose, audience, and genre explain these patterns.
RQ4	Concordance-based interpretation of keywords and bundles	Numbers require contextual reading to produce meaning.
RQ5	Integrated framework and reproducible protocol	Corpus linguistics contributes explanatory and methodological value to applied language research.

X. DISCUSSION

The findings demonstrate that corpus linguistics is most valuable when it explains language variation rather than merely documenting frequency. Comparing academic writing, media discourse, professional communication, and spoken interaction shows that linguistic features distribute according to communicative function. Academic writing had the highest lexical density and a strong concentration of nominalization, evidential collocation, and argument-organizing bundles. This pattern reflects the communicative goal of building abstract, evidence-based knowledge. The dense lexical profile of academic prose is not simply a stylistic preference; it enables writers to compress information, create conceptual relations, and position claims within disciplinary argument.

These results confirm earlier register research showing that linguistic variation is systematically related to communicative situation [2], [8]-[13]. However, the study extends that work by comparing four domains under one analytical protocol and by analyzing lexical, grammatical, collocational, phraseological, and discourse-organizational features together. The findings therefore move beyond single-feature descriptions and show that register is produced by clusters of features rather than a single isolated marker.

Media discourse displayed a different profile. It did not have the same level of grammatical abstraction as academic writing, but it showed stronger evaluative keywords and event-framing collocations. This supports the view that public intelligibility, narrative immediacy, and audience orientation shape media language. Media texts need to name events, attribute claims, identify social actors, and make consequences visible. Corpus evidence therefore helps explain why words associated with crisis, public reaction, conflict, and consequence cluster in media discourse. These patterns are not neutral lexical choices; they help frame events for public interpretation.

Professional communication was distinguished by obligation markers, procedural bundles, and institutional collocations. This profile reflects professional discourse's regulatory and action-oriented function. Unlike media discourse, professional communication often avoids dramatic evaluation and instead prioritizes clarity, responsibility, compliance, and procedural sequence. Modal verbs such as must, should, and required to occur not simply as grammatical forms but as institutional tools that allocate responsibility

and guide behavior. The findings therefore show how grammar can encode social organization and institutional authority.

Spoken interaction differed sharply from the written domains. It showed lower lexical density but higher frequencies of discourse markers, first-person forms, questions, and interactional bundles. This does not mean spoken language is less complex. Rather, its complexity lies in real-time negotiation, stance management, turn-taking, repair, and interpersonal alignment. Discourse markers such as well, so, I mean, and you know help speakers organize talk while monitoring shared understanding. This confirms discourse-marker and spoken-corpus research while also showing how interactional features contrast with academic, media, and professional registers [19], [20], [21].

The results also extend previous work on lexical bundles and phraseology. Earlier studies show that bundles are sensitive to genre, discipline, and corpus design [14]- [17]. The present findings support this view but add a stronger comparative interpretation: academic bundles organize argument, professional bundles regulate action, media bundles frame public events, and spoken bundles manage interpersonal negotiation. Thus, bundles are not only repeated strings; they are register resources that perform recognizable discourse work.

The study avoids overgeneralizing from English-only data. Its conclusions apply most directly to the four English-language communicative domains represented in the corpus. It does not claim that all languages, media traditions, professional sectors, or spoken communities display identical patterns. Nevertheless, the framework is transferable because it specifies how to compare domains through transparent corpus design, statistical validation, and concordance interpretation. Future multilingual and mixed-effects studies can test whether the same feature clusters operate across other languages and discourse communities.

XI. THEORETICAL CONTRIBUTIONS

The central theoretical contribution is a multi-domain corpus-based register-variation framework. The framework presents language variation as a multidimensional phenomenon shaped by communicative purpose and discourse community. It extends corpus linguistics beyond method by showing how corpus evidence can be interpreted through six theoretical lenses: corpus linguistics identifies recurrent evidence; usage-based linguistics explains routinization; variationist sociolinguistics explains contextual distribution; register theory explains situation-function relationships; discourse analysis explains stance, framing, cohesion, and interaction; and applied linguistics connects patterns to teaching and professional communication.

The contribution to sociolinguistics lies in showing that variation is not only a matter of regional or demographic difference. It is also produced by the communicative situations in which speakers and writers operate. Academic, media, professional, and spoken domains create different expectations, and these expectations shape language choices. The contribution to discourse studies lies in linking repeated linguistic patterns to stance, framing, obligation, cohesion, and interpersonal organization.

XII. METHODOLOGICAL CONTRIBUTIONS

The study contributes methodologically by offering a reproducible protocol for multi-corpus comparison. The protocol begins with balanced corpus construction and proceeds through normalization, keyword analysis, collocation analysis, lexical-bundle extraction, n-gram analysis, dispersion measurement, concordance interpretation, and reliability checking. Supplementary corpus metadata, search strings, stop-word decisions, keyword lists, collocation outputs, bundle outputs, concordance samples, and reliability-check sheets strengthen reproducibility.

The study also shows that statistical validation and discourse interpretation are not competing approaches. Log-likelihood, normalized frequency, dispersion, association measures, and effect-size logic help identify patterns worth attention, while concordance analysis and register interpretation explain their meaning. This combination prevents overreliance on frequency, under-documentation of corpus design, and insufficient interpretation of quantitative findings.

XIII. PRACTICAL IMPLICATIONS

For language education and applied linguistics, the findings show that teaching materials should be informed by authentic register-specific patterns. Academic writing instruction should address evidence-framing bundles, cautious stance, nominalization, and logical connectors. Professional communication training should include procedural bundles, obligation markers, and risk-related collocations. Media literacy work can use keyword and collocation analysis to show how public issues are framed. Speaking instruction should treat discourse markers as functional resources for interaction rather than as errors or fillers.

Table 12. Concise practical implications linked to corpus findings.

Area	Practical implication	Direct link to findings
Academic writing pedagogy	Teach bundles, stance markers, and nominalization through authentic corpus evidence.	Academic writing showed strong lexical density, nominalization, and argument-organizing bundles.
Media literacy	Analyze keywords and collocations that frame public issues.	Media discourse showed stronger evaluative keywords and event-framing collocations.
Professional communication	Evaluate obligation markers, procedural bundles, and clarity patterns.	Professional communication foregrounded compliance, responsibility, and procedure.
Spoken-language teaching	Teach discourse markers as interactional resources.	Spoken interaction showed the highest interpersonal markers and first-person forms.

XIV. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, the corpus represents four broad communicative domains, but each domain contains internal diversity that deserves more specialized study. Academic writing differs across disciplines, media discourse varies across political and editorial traditions, professional communication varies across sectors, and spoken interaction varies across activity types and communities. Second, the analysis focuses on English-language data, so cross-linguistic comparison remains a future priority. Third, some discourse interpretations rely on researcher judgment, although reliability checks and concordance analysis were used to strengthen transparency.

Fourth, topic effects may remain within each subcorpus. A word or bundle may be frequent because of register conventions, but it may also reflect recurring topics within a domain. Dispersion checks reduce this risk but do not eliminate it entirely. Fifth, the study does not use mixed-effects modeling or multidimensional analysis, which would allow stronger testing of whether register, genre, topic, source, and author variables independently predict linguistic variation.

Future research should extend the design to multilingual corpora, learner corpora, digital communication corpora, and diachronic datasets. It should also combine corpus analysis with interviews, ethnographic observation, classroom intervention, or reader-response methods. More advanced statistical modeling, including mixed-effects regression and multidimensional analysis, would deepen understanding of how language use changes across contexts and communities.

XV. CONCLUSION

This study examined the role of corpus linguistics in investigating language use and variation through a multi-corpus comparative design. It showed that corpus linguistics is not simply a method for counting words; it is an empirical and interpretive approach for explaining how language patterns vary across communicative domains. By comparing academic writing, media discourse, professional communication, and spoken interaction, the study demonstrated that variation operates across lexical density, grammatical

structure, collocational behavior, lexical bundles, discourse markers, stance expression, and rhetorical organization.

The findings indicate that each domain has a distinct linguistic profile. Academic writing is characterized by lexical compression, nominalization, evidential collocation, and argument-organizing bundles. Media discourse relies on evaluation, event framing, and public-facing collocations. Professional communication foregrounds obligation, procedure, compliance, and institutional responsibility. Spoken interaction relies on discourse markers, pronouns, questions, and interpersonal bundles that support real-time negotiation and social alignment. These patterns show that purpose, audience, genre, and discourse community shape language variation rather than isolated linguistic choice.

The study's strongest contribution is its multi-domain, corpus-based register-variation framework, which links corpus construction, feature extraction, statistical comparison, concordance interpretation, register explanation, and applied implications. The conclusions depend on transparent corpus documentation, reproducible search procedures, validated statistical thresholds, and accessible supplementary evidence. With these conditions in place, corpus linguistics offers a rigorous way to make empirically grounded claims about language use while preserving the interpretive depth needed to explain social, rhetorical, and communicative meaning.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

[1]. T. McEnery and A. Hardie, Corpus Linguistics: Method, Theory and Practice. Cambridge, UK: Cambridge University Press, 2012, doi: 10.1017/CBO9780511981395.

[2]. D. Biber and S. Conrad, Register, Genre, and Style, 2nd ed. Cambridge, UK: Cambridge University Press, 2019, doi: 10.1017/9781108686136.

[3]. J. Egbert, T. Larsson, and D. Biber, Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User. Cambridge, UK: Cambridge University Press, 2020, doi: 10.1017/9781108888790.

[4]. P. Baker and T. McEnery, Eds., Corpora and Discourse Studies: Integrating Discourse and Corpora. London, UK: Palgrave Macmillan, 2015, doi: 10.1057/9781137431738.

[5]. J. Egbert, D. Biber, and B. Gray, Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness. Cambridge, UK: Cambridge University Press, 2022, doi: 10.1017/9781316584880.

[6]. V. Brezina, Statistics in Corpus Linguistics: A Practical Guide. Cambridge, UK: Cambridge University Press, 2018, doi: 10.1017/9781316410899.

[7]. M. Gillings, G. Mautner, and P. Baker, Corpus-Assisted Discourse Studies. Cambridge, UK: Cambridge University Press, 2023, doi: 10.1017/9781009168144.

[8]. R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery, "The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations," International Journal of Corpus Linguistics, vol. 22, no. 3, pp. 319-344, 2017, doi: 10.1075/ijcl.22.3.02lov.

[9]. J. Laws and C. Ryder, "Register variation in spoken British English: The case of verb-forming suffixation," International Journal of Corpus Linguistics, vol. 23, no. 1, pp. 1-27, 2018, doi: 10.1075/ijcl.16036.law.

[10]. J. Egbert, "Fiction - one register or two? Speech and narration in novels," Register Studies, vol. 2, no. 1, pp. 72-101, 2020, doi: 10.1075/rs.19006.egb.

[11]. T. Scheffler, L.-A. Kern, and H. Seemann, "The medium is not the message: Individual-level register variation in blogs vs. tweets," Register Studies, vol. 4, no. 2, pp. 171-201, 2022, doi: 10.1075/rs.22009.sch.
[12]. J. Grieve, "Register variation explains stylometric authorship analysis," Corpus Linguistics and Linguistic Theory, vol. 19, no. 1, pp. 1-34, 2023, doi: 10.1515/cllt-2022-0040.
[13]. H. Li, J. Dunn, and A. Nini, "Register variation remains stable across 60 languages," Corpus Linguistics and Linguistic Theory, vol. 19, no. 3, pp. 397-426, 2023, doi: 10.1515/cllt-2021-0090.
[14]. K. Hyland and F. K. Jiang, "Academic lexical bundles: How are they changing?" International Journal of Corpus Linguistics, vol. 23, no. 4, pp. 383-407, 2019, doi: 10.1075/ijcl.17080.hyl.
[15]. Y. Wang, "Lexical bundles in spoken academic ELF: Genre and disciplinary variation," International Journal of Corpus Linguistics, vol. 22, no. 2, pp. 187-211, 2017, doi: 10.1075/ijcl.22.2.02wan.
[16]. F. Pan, R. Reppen, and D. Biber, "Methodological issues in contrastive lexical bundle research," International Journal of Corpus Linguistics, vol. 25, no. 2, pp. 216-230, 2020, doi: 10.1075/ijcl.19063.pan.
[17]. V. Cortes, "LBiaP: A lexical bundle identification and analysis program," International Journal of Corpus Linguistics, vol. 28, no. 2, pp. 237-263, 2023, doi: 10.1075/ijcl.21100.cor.
[18]. D. Gablasova, V. Brezina, and T. McEnery, "Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence," Language Learning, vol. 67, no. S1, pp. 155-179, 2017, doi: 10.1111/lang.12225.
[19]. C. Ruehlemann, "Turn structure and inserts," International Journal of Corpus Linguistics, vol. 25, no. 2, pp. 186-215, 2020, doi: 10.1075/ijcl.19098.ruh.
[20]. K. Dobrovoljc, "Identifying dictionary-relevant formulaic sequences in written and spoken corpora," International Journal of Lexicography, vol. 33, no. 4, pp. 417-442, 2020, doi: 10.1093/ijl/ecaa008.
[21]. V. Werner, "A diachronic perspective on telecinematic language," International Journal of Corpus Linguistics, vol. 26, no. 1, pp. 38-70, 2021, doi: 10.1075/ijcl.00036.wer.
[22]. P. Rodriguez-Puente, "New methods for analyzing diachronic suffix competition across registers," International Journal of Corpus Linguistics, vol. 27, no. 4, pp. 517-546, 2022, doi: 10.1075/ijcl.22014.rod.
[23]. L. L. Aull, "Linguistic markers of stance and genre in upper-level student writing," Written Communication, vol. 36, no. 2, pp. 267-295, 2019, doi: 10.1177/0741088318819472.
[24]. M. Paquot and L. Plonsky, "Quantitative research methods and study quality in learner corpus research," International Journal of Learner Corpus Research, vol. 3, no. 1, pp. 61-94, 2017, doi: 10.1075/ijlcr.3.1.03paq.
[25]. T. McEnery, V. Brezina, D. Gablasova, and J. V. Banerjee, "Corpus linguistics, learner corpora and SLA: Employing technology to analyze language use," Annual Review of Applied Linguistics, vol. 39, pp. 74-92, 2019, doi: 10.1017/S0267190519000096.
[26]. N. Tracy-Ventura and M. Paquot, Eds., The Routledge Handbook of Second Language Acquisition and Corpora. London, UK: Routledge, 2021, doi: 10.4324/9781351137904.
[27]. A. O'Keeffe and M. McCarthy, Eds., The Routledge Handbook of Corpus Linguistics, 2nd ed. London, UK: Routledge, 2022, doi: 10.4324/9780367076399.
[28]. E. Friginal and J. A. Hardy, Eds., The Routledge Handbook of Corpus Approaches to Discourse Analysis. London, UK: Routledge, 2021, doi: 10.4324/9780429259982.
[29]. M. Gerlach and F. Font-Clos, "A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics," Entropy, vol. 22, no. 1, art. 126, 2020, doi: 10.3390/e22010126.
[30]. A. Kilgarriff, V. Baisa, J. Busta, M. Jakubicek, V. Kovar, J. Michelfeit, P. Rychly, and V. Suchomel, "The Sketch Engine: Ten years on," Lexicography, vol. 1, no. 1, pp. 7-36, 2014, doi: 10.1007/s40607-014-0009-9.
[31]. A. Hardie, "CQPweb - combining power, flexibility and usability in a corpus analysis tool," International Journal of Corpus Linguistics, vol. 17, no. 3, pp. 380-409, 2012, doi: 10.1075/ijcl.17.3.04har.
[32]. D. Gablasova, V. Brezina, and T. McEnery, "Exploring learner language through corpora: Comparing and interpreting corpus frequency information," Language Learning, vol. 67, no. S1, pp. 130-154, 2017, doi: 10.1111/lang.12226.

https://doi.org/10.48161/ajst.v1n4a87

Appendices

1. APPENDIX A. CORPUS METADATA

Appendix A summarizes the metadata fields used to document corpus construction, source status, and comparability across the four communicative domains.

Table A1. Corpus metadata register and source-documentation summary.

Subcorpus code	Domain and source type	Texts / tokens / date range	Metadata recorded	Inclusion and exclusion control
ACAD	Academic writing: open-access journal articles, repository texts, and scholarly sections	820 texts; 1,199,660 tokens; 2016-2023	document ID, discipline/section, year, token count, source status, cleaning note	Included full research-oriented prose; removed references, navigation text, duplicates, and non-article material.
MEDIA	Media discourse: open news archives, public news pages, editorials, and public media repositories	1,450 texts; 1,199,150 tokens; 2016-2023	document ID, outlet/repository type, genre, date, topic tag, token count	Included news/editorial texts; excluded duplicated syndicated copies, comments, advertisements, and page furniture.
PROF	Professional communication: public reports, guidelines, policy documents, and organizational materials	760 texts; 1,200,040 tokens; 2016-2023	document ID, organization type, document type, date, sector, token count	Included public professional documents; excluded promotional texts without clear genre status and incomplete documents.
SPOK	Spoken interaction: public interview, seminar, service-encounter, and conversation transcripts	950 texts; 1,200,800 tokens; 2016-2023	document ID, speech event type, transcript source, date/status, token count, anonymization note	Included ethically accessible transcripts; excluded private, unverifiable, or personally identifying material.
TOTAL	Four domain-balanced subcorpora	3,980 texts; 4,799,650 tokens; 2016-2023	shared metadata fields plus domain-specific notes	Supports normalized comparison across communicative domains and audit-based interpretation.

2. APPENDIX B. FEATURE LISTS AND SEARCH STRINGS

Appendix B records the main feature lists, search strings, and software settings used for frequency, keyword, grammatical, collocational, n-gram, and discourse-marker analysis.

Table B1. Feature lists, search strings, and analysis settings.

Analysis component	Feature list / target forms	Search strings or settings	Output retained
Lexical variation	content words, technical nouns, evaluative adjectives, first-person forms	normalized per 10,000 words; type-token ratio x10; dispersion checked by subcorpus	frequency tables, lexical-density measures, subcorpus summaries
Keyword analysis	domain-salient lexical items and grammar-sensitive keywords	log-likelihood against combined reference corpus; LL >= 10.83; log ratio >= 1.0 where relevant	keyword lists, LL values, log-ratio values, interpretive labels
Collocation analysis	evidence, crisis, compliance, know, analysis and domain-specific node words	span 5L/5R; MI >= 3; t-score >= 2; minimum frequency = 10	node-collocate tables with frequency, MI, t-score, LL, semantic preference
Lexical bundles / n-grams	3- to 5-word recurrent bundles	minimum 20 per million; present in at least 5 texts; DP <= .80; manual filtering of fragments	bundle list, frequency per million, texts represented, dispersion, function
Grammatical variation	pronouns, modals, passives, nominalizations, questions	POS-assisted search, manual validation, normalized distribution, effect-size comparison	feature-count table, concordance notes, functional interpretation
Discourse markers	well, so, like, you know, I mean, however, therefore, in contrast, where applicable	search strings plus concordance coding; 15% reviewer check	marker frequencies, concordance samples, function codes, reliability notes

3. APPENDIX C. NORMALIZED FREQUENCY AND KEYWORD TABLES

Appendix C presents a compact summary of normalized rates and keyword/keyness patterns used to support the main statistical and corpus interpretation. Full outputs may be supplied as spreadsheets when required by the journal.

Table C1. Normalized frequency and keyness summary for selected features.

Feature / keyword set	Academic writing	Media discourse	Professional communication	Spoken interaction	Statistical note
Lexical density (%)	59.4	51.8	47.2	38.6	LL = 412.6; register effect visible across subcorpora
Type-token ratio x10	7.8	6.4	5.7	4.9	dispersion stable across subcorpora
Technical / disciplinary nouns per 10k	214	82	176	31	LL = 568.2; log ratio = 1.94
Evaluation adjectives per 10k	71	126	58	83	LL = 221.4; log ratio = 1.20
First-person forms per 10k	18	11	26	61	LL = 487.9; log ratio = 1.76

Nominalizations per 10k	182	119	143	74	LL = 336.5; log ratio = 1.30
Modals of obligation per 10k	22	17	81	29	LL = 502.1; log ratio = 2.05
Questions per 10k	8	14	21	88	LL = 643.4; log ratio = 2.46

4. *APPENDIX D. COLLOCATION AND BUNDLE OUTPUTS*

Appendix D consolidates the main collocation and lexical-bundle outputs used to link quantitative recurrence with register-sensitive discourse interpretation.

Table D1. Collocation and lexical-bundle outputs with interpretive functions.

Node / bundle	Subcorpus	Main collocates or bundle form	Association / frequency evidence	Interpretive function
evidence	Academic writing	suggests; indicates; supports; provides; robust	MI = 4.21; t = 9.84; LL = 142.7	cautious inference and evidential support
crisis	Media discourse	global; growing; public; economic; political	MI = 4.87; t = 11.26; LL = 188.4	public framing of urgency and collective consequence
compliance	Professional communication	requirement; policy; ensure; risk; procedure	MI = 5.11; t = 10.93; LL = 204.6	institutional obligation and accountability
know	Spoken interaction	I; you; mean; just; do not	MI = 3.76; t = 12.48; LL = 231.2	stance, alignment, and shared understanding
the results of the	Academic writing	3- to 5-word bundle group	88 per million; DP = 0.61	evidence organization and abstraction
according to officials	Media discourse	attribution and event-framing bundle	39 per million; DP = 0.66	source authority and narrative coherence
in accordance with the	Professional communication	procedural bundle group	56 per million; DP = 0.59	procedure, obligation, and institutional clarity
I do not know / you know what	Spoken interaction	interactional bundle group	122 per million; DP = 0.55	turn-taking, repair, and interpersonal alignment

5. APPENDIX E. CONCORDANCE SAMPLES AND RELIABILITY-CHECK SHEETS

Appendix E provides shortened concordance-style samples and a reliability-check summary. The samples are brief to avoid reproducing extended copyrighted source material.

Table E1. Concordance samples supporting the corpus interpretations.

Feature	Subcorpus	Short concordance-style context	Function code	Interpretive note
evidence suggests	Academic writing	the evidence suggests that the association is context dependent	cautious inference	academic prose presents knowledge through hedged evidential reasoning
global crisis	Media discourse	officials warned that the global crisis would affect local communities	public framing	media discourse scales events toward collective consequence
must comply	Professional communication	all applicants must comply with the reporting procedure	obligation	professional texts allocate responsibility through modalized procedure
you know	Spoken interaction	I was thinking, you know, maybe we should start again	alignment	spoken discourse uses markers to invite shared understanding
in accordance with	Professional communication	documents should be submitted in accordance with the guidance	procedural sequence	bundles reduce ambiguity and standardize institutional action

Table E2. Reliability-check sheet and resolution log.

Checked item	Sample size	Reviewer focus	Resolution procedure	Outcome
Concordance interpretations	15% stratified sample	function labels, local context, semantic prosody	reviewed disagreements against local context and memo notes	category wording refined; no major interpretive shift
Lexical-bundle functions	15% stratified sample	bundle boundaries and discourse function	checked against original concordance lines and dispersion sheets	fragmentary bundles removed or recoded
Discourse-marker classifications	15% stratified sample	cohesion, stance, turn management, alignment	resolved through codebook clarification and examples	final categories retained with clearer definitions
Corpus metadata and inclusion notes	full source register check	domain assignment and exclusion justification	duplicated or incomplete items removed before final counts	final corpus kept balanced at approximately 1.2 million tokens per subcorpus