QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

HARNESSING COMPUTATIONAL LINGUISTICS FOR SEMANTIC PATTERN ANALYSIS: A CORPUS-BASED INVESTIGATION OF LANGUAGE COHESION AND CONCORDANCE

Amjad Shaheen
M.Phil Research Scholar, Department of English, University of Okara, Pakistan,
Email: amjadcmk6@gmail.com
Dr.Irfan Mehmood
Lecturer, University of Okara, Pakistan
Email: irfan@uo.edu.pk
Research Associate, University of York, UK
Email: irfan.mehmood@york.ac.uk
Muhammad Iqbal
(Correspondence)
Visiting Lecturer, Department of English, The University of Sahiwal, Pakistan,
Email: m.iqbl@uosahiwal.edu.pk

Abstract
This paper explores the complex association between cohesion and meaning and attempts to work out potentially complex association between cohesion and meaning using computational methods on a wide range of English literature. In going a step further beyond mere theorizing, the study uses both quantitative and qualitative research to track patterns of certain words and phrases across non-academic and academic genres. The results obtain a vivid discrepancy in the use of such devices as conjunctions and pronouns as constructing an argument and cohesive narrative. Not only does such analysis present empirical evidence of functional variance in language use, it also offers practical insight into the importance of using corpus tools as a means of revealing architecture in texts that are symbolically hidden. The acquired results are useful to the linguistic theory and advanced, more natural language processing applications.
Keywords: *computational linguistics, corpus analysis, semantic cohesion, discourse markers, genre analysis*

Introduction

The terrain of linguistic investigation has changed irrevocably and fundamentally, due to the digital revolution. Searches in language were long based on introspection, isolated sentences and the untried sentences of grammarians. Although this method bore fruitful results, it was necessarily constrained because of its nature and risk of subjectivity, as it tended to present a fancy image of language and not as it is, dynamic, at once. The emergence of computational power and an ability to, collect and analyze, large data sets of authentic text, corpora created a new paradigm. This change is the advent of corpus linguistics as a major approach and the development of the field out of a philosophical and gut-driven discipline into an empirical and evidence-driven science. Lying at the centre of this change is the discipline of computational linguistics, an interdisciplinary field that lies

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

at the intersection of computer science, artificial intelligence, and linguistic theory. It not only gives the tools and frameworks to manage such immense databases of text but also creates ways to distill meaningful patterns out the data which in many cases are on the eye. It is within the context of this thrilling tradition that this research finds itself, and it is the intention of the research to utilize tools of computation in order to gain a better understanding of two fundamental pillars of discourse, semantic pattern and linguistic cohesion, and the role that they play in the construction of meaning across genres of text.

The idea of learning language by means of large-scale samples is not absolutely new, and lexicographers used the collections of citations to learn about word meaning and usage long time ago. But the scale and the speed that is possible now with modern computing are different now. The main ideology of corpus linguistics is that language frequency cannot be taken lightly on as a deciding factor in cognition, conventionality, and cultural worth. Naturally, corpus analysis can give linguists a higher level of objectivity and replicability, as McEnery and Hardie (2012) persuasively reason it is indeed the case since the statements made about language will no longer be based on personal intuition but on empirical evidence, extracting a higher degree of objectivity and replicability. This empirical direction enables academics to know what is main and normal in a language as compared to what is marginal and unusual. It allows a study of grammar in terms of probabilities, rather than as a set of abstract rules, one that is shaped by millions of individual acts of use. This textual data-driven methodology is essential to the analysis of texts as artifacts, in their capacity as whole coherent objects, the focus of concern in this work.

At the center of any corpus based analysis is the notion of concordance which is an area that has been mastered as part of computational linguistics. Concordance is an index to all the occurrences of a word or phrase in a text, usually given in context, frequently known as a Key Word In Context (KWIC) display. It is a simple but effective tool turning the otherwise frozen text into dynamic set of data. Looking at lines of concordance of a target word, a researcher will be able to see with whom the word habitually combines with, its collocations, and its common colligations, i.e., the construction types it is used in. The profound implications of this were brought out by John Sinclair (1991) more than anyone scholar. He has offered a principle already called idiom to the effect that a language-user possesses in his repertory a very large host of phrases that are semi-preconstructed, that is, that are single choices, although they may be segmentable. This means that we understand a phrase like "ring a bell" not by analyzing the individual words "ring" and "bell" but as a single unit meaning "to seem familiar." Concordancing is the primary method for uncovering these vast, underlying networks of prefabricated chunks that are essential for fluent language production and comprehension.

However, understanding individual word patterns is only part of the story. For a text to be perceived as coherent and meaningful, its constituent parts must be connected logically and semantically. This is the domain of cohesion, a concept masterfully elaborated by Halliday and Hasan (1976). Cohesion refers to the linguistic devices that create texture, binding a text together and differentiating it from a random sequence of sentences. These devices include grammatical cohesion, such as using pronouns to refer back to a noun (anaphoric reference) or conjunctions like "however" and "therefore" to signal logical relationships.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

They also include lexical cohesion, achieved through the repetition of words or the use of synonyms and related terms throughout a text. cohesion is not the same as coherence; coherence is the psychological experience of the text making sense, while cohesion is the tangible linguistic machinery that facilitates that experience. A text can be cohesive but not coherent (e.g., a well-connected but nonsensical argument), but it is rarely coherent without being cohesive. Therefore, analyzing cohesive devices provides a crucial window into how speakers and writers guide their audience through a narrative or argument, explicitly signaling how each new sentence links to what has been said before.

The intersection of concordance and cohesion is where this study finds its focus. While concordance often reveals local, phrase-level patterns, the sum of these patterns contributes to the global cohesive structure of a text. For example, the frequent use of certain conjunction-like adverbials ("furthermore," "consequently") in a corpus of academic writing, as revealed through concordance, is a strong indicator of that genre's emphasis on creating overt, logical cohesion. Conversely, a high frequency of pronouns in a narrative genre, as shown by frequency lists, points to a different cohesive strategy—one that relies on a reader's ability to track participants through a story. Different genres, by virtue of their communicative purposes and intended audiences, employ these devices in systematically different ways. Biber, Conrad, and Reppen (1998) have extensively documented these genre variations, showing how linguistic features cluster together to define registers such as academic prose, newspaper reporting, and casual conversation. Academic writing, for instance, is characterized by a high density of nominalizations, attributive adjectives, and precise conjunctive markers, all working to create a dense, authoritative, and logically structured discourse.

This study, therefore, aims at contributing to this existing body of knowledge, by investing the targeted, mixed-methods research of a balanced corpus of the works. It is intended to go beyond generalizations that many cohesion markers simply apply across the board, or develop a closer analysis of the ways in which particular cohesion markers work within and across genres with the application of tools such as AntConc. The research questions on this research are: What are the similarities and differences between the means and frequencies of cohesive devices between academic and non-academic genres? What do such differences suggest about the mean making strategies that each uses? And what does it take to draw effective computational tools to represent these involved patterns of semantic and pragmatic choices? The study herewith seeks to answer these questions, and in so doing, not only proves the stronger shoulder on which theories that have already been developed in the linguist field rest but also how computational linguistic can be applied to light on the very makeup of human communication. The results have great implications not only to the study of applied linguistics and language education, but also the Natural Language Processing (NLP) industry whereby better comprehension of cohesion is needed in enhancing machine capabilities to generate and comprehend human-like text.

Literature Review

The theoretical approach to this study lies in the strong tradition of corpus linguistics, that is, the study of language using actual, real-world evidence as opposed to theoretical, made-up or intuitive. As well documented by McEnery and Hardie (2012), this paradigm shift

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

proposes that language as used and not as described by models can be thoroughly and objectively studied based on large digitized text corpora because in such an approach, language can be quantitatively analyzed and one will not need to rely on ad hoc collections of other people. This shift in direction towards data-driven discovery transformed linguistic study on its own account, enabling scholars to recognize and measure patterns that used to exist only anecdotally or not at all. The given research is conducted in the environment of such methodological activity in which the potential of computational research tool is used to study linguistic phenomena on the scale that has never been possible before.

One of the most important scholars in this field is John Sinclair (1991) who transformed our perceptions of how vocabulary and meaning works through collocation. He was convincing in his exposition that words are not singly selected but in clumps, what he called the idiom principle. This implies that users of the language are endowed with an extensive mental bank of semi-preconstructed units in language, and that these units often mediate meaning instead of the simple addition of word meaning. The semantic value of a word, then, is as clearly shown by its associates of culture as are the linings of the blue-grey mud-waste by the water-lilies. This regulation can be directly applied to our study in identifying semantic patterns since we will be trying to establish various collocational networks that are central to academic and non-academic language.

At the phrase level, the variable of cohesion would offer the apparatus through which texts are pieced back together in order to comprise a coherent whole. The classic model put forward by Halliday and Hasan (1976) classifies those grammatical and lexical devices i.e. reference, substitution, ellipsis, conjunction and lexical repetition which can add texture as well as bind a text together. Their work put to rest the assumption that these cohesive relations are purely stylistic and are also needed in order to lead the reader through logical and semantic space of a discourse. Whereas cohesion is the visible stitches of a text, it enables coherence, which is the readers ability to make sense of a text in its entirety as one message.

It is based on this foundation that modern-day studies have looked at the systematic variation of these forms of cohesive devices in various genres and registers. These differences have been well documented by scholars such as Biber, Conrad and Reppen (1998) who show that the forms of language vary together in predictable ways, depending upon the communicative purpose of a text. An example would be their study of the dense feature such as academic prose which is assessed as high in density of nominalization and precise conjunctive markers and elaborated reference, creating dense, authoritative and logically explicit discourse. Alternatively narrative genres would be expected to use pronouns and temporal conjusbcnctions to help sustain a story line. Most recent research, namely that of Baker et al. (2013), has used corpus-based computational methods to trace these discourse markers through large-scale repositories of data; also, with confidence, it can now derive empirical confirmation of these discourse marker processes in relation to genre specificity. This pool of literature bears witness to the fact that cohesion is not a one-dimensional phenomenon but flexible system used in different contexts. We seek to make this contribution by offering a highly detailed study of the roles played by a set of cohesion markers identified on a corpus with the help of modern concordancing tools in a balanced

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

corpus, thus connecting theoretical insights about the workings of cohesion markers of the past and the data-crunching capabilities of modern linguistic tools.

Research Questions

1. How does concordance analysis serve as a fundamental methodology for identifying and interpreting lexical and syntactic patterns within a large linguistic corpus?
2. In what ways do the frequency and functional applications of cohesion markers, such as conjunctions and referential devices, vary between academic and non-academic genres?
3. To what extent can contemporary computational tools be effectively leveraged to automate the detection and analysis of complex semantic patterns across diverse textual data?

Research Objectives

1. To investigate the role of concordance in corpus-based linguistic analysis.
2. To examine how cohesion markers function within different genres.
3. To explore the application of computational tools for semantic pattern detection.

Research Methodology

To answer the research questions, this research had a mixed-methods design and consequently used the two techniques of research in a synergistic manner. This design was chosen to take the best out of the two paradigms: the quantitative aspect offers quantitative, generalizable data on the prevalence and distribution of linguistic characteristics, whereas the qualitative aspect will allow one to interpret how these characteristics are deployed in context in a nuanced and context-sensitive fashion (Creswell & Plano Clark, 2017). This triangulation of approaches contributes to validity and richer level of interpretation of the results that goes beyond mere counting of results but to a logical explanation of the observed patterns.

This analysis uses data based on a purpose built, balanced written contemporary English corpus, created to allow representative genres to be compared. The corpus was separated into two main sub-corpora: one included academic journals in the field of studies affiliated with humanities and social sciences, and the other one non-academic texts of online-based resources (news and magazine articles of high-quality media). The design is suitable to address the controlled solution to test genre-specific differences in the use of cohesion markers. The texts have been thoroughly chosen by taking into consideration a recent period (published in 2018-2023) to represent the current use of language. As a final step before the actual analysis was carried out, the corpus was cleaned and normalized, that is, converted into plain text format and the extraneous metadata removed together with standardizing on a single variant of spelling (in this case, American English) so as to avoid inconsistencies in automatic processing.

Data extraction and initial analysis was conducted with the aid of the AntConc (Version 4.2.0), a freeware concordance software package created by Laurence Anthony (2022). There are some main operations that were carried out using this tool. It first produced the frequency lists of the whole corpus and the single sub-corpora respectively, which allow to take a macroscopic glance at the most frequently used words and gain a general idea of lexical distinctions between genres. Second and most importantly, it also enabled the

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

generation of Key Word In Context (KWIC) concordonce lines of pre-selected search terms. These words were modeled using the theories of cohesion, concentrating on which of the items provided by Halliday and Hasan (1976), additive conjunctions (e.g., furthermore, moreover), contrastive conjunctions (e.g., however, nevertheless), causative conjunctions (e.g. thereafter, consequently), and a set of personal and demonstrative pronouns.

The analysis was an iterative procedure The quantitative data were computed to determine the significant disproportions in the academic and non-academic sub-corpora in references to raw frequency and normalized frequency (per 10,000 words) of each cohesion marker. Later on a substantial number of the concordance lines of these markers was exported and intensively analysed qualitatively. This required readings through the context of each of them to be able to categorize their discursive deployment, such as to classify whether "however" was deployed to contrast a whole proposition or merely to shading some aspect of one claim. This mixed strategy that goes back and forth between macro and micro trends is the finest example of the practice of a corpus-based linguistic research since empirical evidence is firmly rooted in textual reality (McEnery & Hardie, 2012).

Quantitative Overview: Frequency and Distribution of Cohesive Devices

The first stage of analysis was macroscopic, quantitative study of the corpus providing the basis of understanding of a distribution of cohesive resources in the academic and non-academic genres. This quantitative overview is the necessary preliminary step, shifting between the general tendencies and particular characteristics which will be considered in subsequent sections. The procedure started with the generation of exhaustive lists of frequencies in each sub-corpus, which indicated immediately the existence of a clear lexical profile in each genre. The academic sub-corpus was typified by the presence of more nouns, Latinate words and longer words as it character stereotypical informational density and the formal register of this sub-corpus. The non-academic group of texts, in turn, featured more verb and pronoun usage and shorter, more Anglo-Saxon words, a feature in keeping with its narrative and interactive orientation. This overall departure pre-conditioned a more focused inquiry into the concrete words that form the cohesion markers.

To systematically compare the use of cohesive devices, a predefined list of search terms was established based on the theoretical framework of Halliday and Hasan (1976). This list included items from the categories of conjunctions (e.g., *however, therefore, furthermore, consequently, thus, hence, although, because*) and reference (e.g., *this, that, these, those, it, he, she, they*). The AntConc software was used to calculate the raw frequency of each item in both sub-corpora. However, because the two sub-corpora were of slightly different sizes, these raw counts were then normalized to a rate per 10,000 words to allow for accurate and fair comparison. This normalization process is a standard and crucial practice in corpus linguistics, as it neutralizes the effect of corpus size and allows researchers to make meaningful statements about the propensity of a feature to occur in one genre versus another (McEnery & Hardie, 2012).

The results of this normalized frequency count unveiled a stark and statistically significant contrast between the two genres. As hypothesized, logical conjunctive adverbials—often referred to as linking adverbials—were overwhelmingly a hallmark of the academic texts.

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

For instance, a term like *however* appeared with a normalized frequency of 18.7 occurrences per 10,000 words in the academic sub-corpus, compared to just 6.2 in the non-academic sub-corpus. Similarly, causative markers like *therefore* (12.4 vs. 3.1) and *consequently* (5.8 vs. 1.3) showed a dramatic skew towards academic prose. The additive marker *furthermore* was almost exclusively an academic feature, appearing 9.5 times per 10,000 words in academic texts and a mere 0.8 times in non-academic texts. This pattern strongly supports the findings of scholars like Biber, Conrad, and Reppen (1998), who identify the explicit marking of logical relationships as a defining characteristic of academic register. The high frequency of these markers underscores the academic writer's need to meticulously build a complex argument, constantly guiding the reader through claims, counterclaims, evidence, and conclusions with overt signals. It is a style that prioritizes precision, authority, and logical transparency above all else.

Conversely, the non-academic sub-corpus exhibited its own distinctive cohesive profile. While it used fewer elaborate conjunctions, it displayed a significantly higher frequency of personal pronouns (*he, she, they*) and demonstrative references (*this, that*). The normalized frequency for the pronoun *they*, for example, was 40.3 in non-academic texts compared to 24.1 in academic texts. This discrepancy points to a fundamentally different cohesive strategy. Non-academic writing, particularly journalism and feature articles, is often driven by human actors and specific events. Cohesion is therefore achieved by tracking these participants through a narrative using pronouns, a device that creates fluency and pace while avoiding cumbersome repetition. Furthermore, the demonstrative *this* was used more frequently in the non-academic corpus, but a preliminary glance at the concordance lines suggested a key functional difference—a qualitative aspect to be explored in depth in the next section. Often, in non-academic writing, *this* referred back to a specific, concrete event or person mentioned in the previous sentence (e.g., "The policy was enacted on Tuesday. This decision sparked immediate protest"). In academic writing, *this* often pre-modifies a noun that summarizes an entire abstract concept (e.g., "The data showed a significant correlation between the variables. This finding supports the initial hypothesis").

This quantitative overview thus does more than just present numbers; it provides a clear map of the linguistic territory, highlighting the major fault lines between the genres and directing our attention to the specific features whose functional nuances require closer, qualitative inspection to be fully understood.

Qualitative Insights: Functional Analysis of Key Markers

While the quantitative data painted a clear picture of *how often* specific cohesive devices are used, it is the qualitative, context-rich analysis that reveals the crucial *how* and *why* behind their usage. This second phase of analysis moves beyond frequency counts to a functional examination of concordance lines, uncovering the nuanced roles these markers play in shaping meaning and guiding the reader. This step is critical because, as Hunston (2002) emphasizes, words themselves are not inherently cohesive; it is their function within a specific textual environment that creates cohesion. A word like "so" or "this" can serve multiple masters, and its true contribution to a text's coherence can only be determined by inspecting its immediate context and discursive purpose. This functional analysis revealed that the same surface form often concealed profound genre-

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

based differences in application, a finding that pure quantitative data would have entirely missed.

The conjunction "so" provided a paradigmatic example of this functional split. In the academic sub-corpus, its primary role was that of a logical causative, semantically equivalent to "therefore" or "thus." It functioned as a crucial pivot point in an argument, signaling that a conclusion was being drawn from previously established evidence. For instance, a typical concordance line read: "The initial model failed to account for key variables, *so* it was refined through an iterative process." Here, "so" explicitly marks a relationship of cause and effect between two clauses, reinforcing the text's logical structure. Its collocates were often other academic verbs like "conclude," "suggest," and "hypothesize." In stark contrast, its function in the non-academic corpus was far more varied and discourse-oriented. While it was sometimes used causally, it more frequently acted as a discourse marker to initiate explanations, narratives, or new topical segments. Examples included: "The market crashed unexpectedly. *So*, what does this mean for the average investor?" or "We had no idea what to expect. *So* we just started walking." In these instances, "so" does not signal a logical conclusion but rather manages the flow of information for the reader, often implying a spoken, conversational tone that is characteristic of engaging journalism and feature writing (Carter & McCarthy, 2006).

A similarly revealing functional divergence was observed in the use of demonstrative reference, particularly the word "this." The quantitative data showed it was frequent in both genres, but the qualitative analysis uncovered a fundamental difference in what it referred to—a concept known as its referent. In academic writing, "this" was almost exclusively followed by a noun that summarized a complex, often abstract, piece of prior information. This practice, known as "encapsulation" or "nominalization," is a cornerstone of academic discourse. Concordance lines showed patterns like: "...the results indicated a strong positive correlation. *This finding* challenges previous assumptions..." or "...the theory encompasses several competing ideologies. *This complexity* makes it difficult to apply..." Here, "this" does not point to a simple noun but to an entire proposition ("that there is a correlation") that has been packaged into a single noun ("finding"). This creates a dense, authoritative tone and allows the writer to build arguments by treating previous claims as objects to be analyzed.

In the non-academic texts, the function of "this" was typically more immediate and concrete. It often referred directly to a specific person, event, or object mentioned in the previous sentence without the need for an accompanying noun. For example: "The new park opened downtown yesterday. *This* has been a project ten years in the making." or "The mayor announced her resignation. *This* surprised many of her colleagues." The referent ("the park's opening," "the announcement") is clear from the immediate context without requiring abstraction into a nominal form. This use builds a graver and more approachable flow of narrative that draws the reader along a set of events without discursive burden of the academic trappings. Such a differentiation compellingly argues that cohesion is not a simple unitary grammatical phenomenon but rather an elastic rhetoric. The similarity in using the word this to create unity in both genres is perfect with the respective communicative aims of the genres in mind: one aimed at building up a complicated abstract

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

argument and the other to present an intelligible and interesting story. This profound functional examination indicates that the all-but-unavoidable step in making sense of language in action is to go beyond the word as a single entity to see how it plays out in its own unique genre system.

Collocational Patterns and Semantic Prosody

The last stratum in the analysis is the analysis of collocational patterns and the semantic prosodies they produce. This extends beyond the idiosyncrasy of words to make the'aura of meaning' that the company a word always keeps. Words, as Sinclair (1991) has definitively laid to rest, are not chosen at random; they are chosen in constructs, and these constructs are heavily dependant on the choice of words with which the words habitually occur. It is important to note that this principle is crucial to the realization of the entire pragmatic value of cohesive devices. The word however does not merely mean a contrast, its conceptual role (its rhetorical significance) may manifest itself via variants of such surrounding words as polite academic understatement or a more forceful sense of opposition. Using the collocate and cluster features in AntConc, this paper has been able to map these tendencies and has revealed the fine but potent semantic prosodies that define cohesive devices across genres and are critical to true mastery of languages.

In the scholarly sub-corpus, however, the collocational topography generated by the wordwhatever divines up a very exclusive type of discourse politeness and specificity. It often and consistently occurred in quantity at the beginning of a sentence, most commonly after a period or a semi-colon and was most often immediately preceded by words that recognized or recapitulated some preceding argument. Its good lefts were hedging expressions such as, "It is worth noting that," "Scholars have insisted," and "Although this may be so." On its right, it was often preceded by a comma and followed by a subject pronoun as it is in this very case, namely, a comma, subject pronoun it, this or the results etc. which helped to swing the writer back to his/her own argument. The case that [previous author] gives in favor of this interpretation is strong. The analysis presented here however makes a different conclusion. The semantic prosody that is produced here is not of mere contradiction, but of courteous interaction as well as subtle argument. The term however acts as a discursive softener to permit the writer to contradict the current scholarship in a manner that does not appear confrontational, thus fulfilling the community norms as explored earlier in this paper of being polite, yet provocative in scholarship. This rule of prosody is a basic, but less noticeable rule of academic writing.

The semantic prosody of thereafter as used in the academic texts was also unique detailing a clear picture of deductive reasoning in this word. The strongest collocates that match its definition of concluding were the verbs that imply the very process of concluding: "conclude," "suggest," "argue," "infer," and "hypothesize." It was often in the form of a cluster as in we can therefore conclude or it is therefore argued that. This gives really a good prosody of logical inevitableness and academic dominance. The word indicates that what will be coming further is not a new theory but an unavoidable evidence to already provided evidence that have been laid down exceedingly. It is the collocational adhesive that attaches evidence to claim and its collocational pattern is the fingerprint of this particular rhetorical move. This pattern was virtually non-existent in the non-academic

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

corpus, where logical copulas are frequently suggested implicitly; they are not explicitly marked with such formal features as logical copulas can take on.

In the non-academic texts, cues to collocation varying around cohesive devices also produced a different narrative, one of immediacy and engagement. It was observed that the demonstrative this helped with the creation of a momentum to the narrative. It was often joined by active verbs referring to occurrences or responses: This aroused demonstrations, This provoked, This caused a dilemma. One of the tangible consequences and direct effects is the prosody. The pronoun it frequently accompanied cleft sentences, which are a mainstay of journalistic writing to get important information up front to the reader: "It was this decision that eventually transformed everything." Moreover, conjunction because was more prominent in non-academic samples, and its collocates were many times less complex and abstract descriptions of causality concerning human behavior or feeling, in contrast to the referential explanation of cause that there-fore intends.

The applications of charting such collocational networks are, however, well beyond theoretical linguistics. On language pedagogy, these data suggest the teaching of cohesive devices not as discrete word items, but as they occur within their typical chunks, and with the attendant specifics of their semantic prosody across genres made explicit. A student needs to be taught that however in an essay carries a different set of companions and a different rhetorical weight to but in a story. More importantly, to the field of Natural Language Processing (NLP), this highlights a major problem. Statistical models trained on large corpora should have to learn to differentiate these nuances of prosodies in order to be able to generate or interpret text. A system fail to identify polite, hedging prosody of academic whatever would produce unsuitable, harsh or uncivil text. What this research shows is that it is not counting word-hybrids that can expose us to the rooted, patterned behavior, which is, itself, the language using behavior, something that beyond much human nuance is needed to find the essential data needed to fill the gap between machine and human understanding.

Conclusion

This study aimed to examine the complexity of cinquinas typology and discursive performance in terms of the application of powerful linguistic methods of Computational linguistics to a new corpus consisting of texts. The progression through the levels of quantitative frequency counts to qualitative and functional analysis and then defining the patterns and patterns of collocation has given a layered conception of how cohesion and concordance together are the designers of coherence in terms of meaning of a text. The results reliably support the claim that the cohesive devices are more than decorative elements of the language but rather the necessities, structured and genre-specific tools used by writers to accomplish coherence, control reader and serve the intended communicative purposes. The paper has convincingly argued that mixed-methods approach, where one would utilize data-driven discoveries of computational tools combined with close rhetorical analysis, would be incredibly effective at putting the surface findings aside to get to the underlying rules that dictate language in use.

The investigation confirmed the core hypothesis that genre exerts a powerful influence on linguistic choice. The quantitative disparity in the use of logical conjunctive adverbials

QUALITATIVE RESEARCH JOURNAL FOR SOCIAL STUDIES
ISSN Online: 3006-4686, ISSN Print: 3006-4678
Volume No: 02 Issue No: 03 (2025)

(e.g., *however*, *therefore*) versus narrative-driven devices (e.g., pronouns) between academic and non-academic texts is not a random variation but a direct reflection of their divergent rhetorical goals. Academic writing, with its high frequency of explicit logical markers, confirms Biber et al.'s (1998) characterization of the genre as informational, abstract, and meticulously structured to build complex arguments. Non-academic writing, with its preference for pronouns and demonstratives, prioritizes narrative flow, participant tracking, and accessibility. However, the most significant insights emerged from moving past these counts. The qualitative functional analysis revealed that even shared words like "so" and "this" perform radically different roles—logical causation versus discourse management, abstract encapsulation versus concrete reference—highlighting that true meaning is inextricable from context and genre convention. This aligns with Sinclair's (1991) idiom principle, showing that meaning resides not in words alone but in their habitual and functional patterns of use.

Perhaps the most nuanced contribution of this study lies in its exploration of collocational patterns and semantic prosody. By mapping the company that words keep, we uncovered the unspoken rhetorical nuances that define genres. The polite, hedging environment of "however" in academic prose, surrounded by phrases of acknowledgment, constructs a community-specific ethos of respectful scholarly debate. In contrast, the narrative-driven collocates of "this" in journalism ("this sparked," "this led to") create a prosody of immediacy and consequence. These patterns are the hidden grammar of discourse, the subtle cues that native speakers and proficient writers internalize and that this research has made visible through corpus analysis.

The implications of this work are twofold. For applied linguistics and pedagogy, the findings argue strongly for a genre-based approach to teaching writing. Students should be taught cohesive devices not as a simple vocabulary list but as part of their common lexical chunks and with a clear understanding of their semantic prosody and appropriate generic contexts. For the field of Natural Language Processing (NLP), this research underscores a critical challenge and opportunity. As noted by scholars like Gries (2009), the success of advanced NLP applications, from sentiment analysis to coherent text generation, depends on machines understanding these very nuances. A language model must distinguish the prosody of academic "however" from its more direct usage to avoid generating tonally inappropriate text. Therefore, the detailed, human-analyzed collocational profiles provided by studies like this one are invaluable for training and refining more sophisticated and accurate machine learning models.

In conclusion, this study has illuminated the powerful synergy between computational methods and linguistic inquiry. It has shown that tools like AntConc are not just for counting words but are windows into the discursive soul of a genre. By marrying quantitative scale with qualitative depth, we can move closer to a complete understanding of the complex, patterned, and elegant system that is human language. Future research should continue on this path, integrating machine learning for even larger-scale pattern recognition and expanding the corpus to include spoken and digital genres to further test and refine these models of linguistic cohesion.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software
Baker, P., Gabrielatos, C., & McEnery, T. (2013). Discourse analysis and media attitudes: The representation of Islam in the British press. Cambridge University Press.
Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.
Carter, R., & McCarthy, M. (2006). Cambridge grammar of English: A comprehensive guide. Cambridge University Press.
Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.
Gries, S. T. (2009). Quantitative corpus linguistics with R: A practical introduction. Routledge.
Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Longman.
Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.
McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.
Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.