Pakistan Journal of Society, Education and Language (PJSEL)
Journal Homepage: https://pjsel.jehanf.com/index.php/journal ISSN
2521-8123 (Print)
2523-1227 (Online)

A Comparative Corpus-Based Study: Exploring Pos-Tagged Phraseological Patterns and Lexical Density In Selected Urdu Novels

Ayesha Ali[1], Nimra Iqbal[2], Farhat Abdullah[3]

Original Article

1. **University of Central Punjab, Lahore.**
Email: ayeshaali.01994@gmail.com
2. **University of Central Punjab, Lahore.**
Email: nimradoggar123@gmail.com
3. **University of Central Punjab, Lahore.**
Email: Farhat.abdullah@ucp.edu.pk

Keywords

Corpus Linguistics, POS Tagging, Phraseological Patterns, Urdu Fiction, Lexical Density

Abstract

The current study explores phraseological patterns of selected Urdu novels of contemporary Urdu novelists by using the corpus-based approach. Corpus-based approaches offer regular ways of analyzing the common linguistic patterns and stylistic decisions in literary texts. Previous studies in the Urdu corpus have concentrated on the methods of POS tagging, the development of a corpus, and the accuracy of computation, but little attention has been paid to the utilization of POS-tagged phraseological analysis in the study of stylistic variations in Urdu fiction. Hence, the present study is based on the phraseological patterns of the selected Urdu novels using corpus-based research methodology. A qualitative and quantitative approach was used in the study. Some novels of contemporary writers, namely Meher-un-Nisa Shahmeer and Noor Rajpoot, were selected, and Phraseological units were extracted manually, then it was POS tagged manually. The tagged data were then analysed with the aim of investigating POS frequency distributions of both authors, the occurrence of recurring phrases, and the stylistic differences between the two authors in AntConc.In addition, lexical density was computed to compare the ratio of content words of both corpora. It was found that the dominant category in the combined corpus included nouns. The analysis was carried out to reveal the common phraseological structures and stylistic differences between the author's works. A comparison between the two corpora revealed that Mehrunisa Shahmeer's corpus tended to be more descriptive and lexically dense, whereas Noor Rajpoot's corpus had more dynamic and action-oriented constructions with a higher rate of verbs. The study revealed that corpus-based methods are effective in revealing authors' stylistic differences in Urdu fiction. Future research can enhance the corpus by incorporating additional Urdu novels and authors, and may employ more advanced corpus tools or automated tagging methods for broader linguistic investigation.

1. Introduction

Language is an integral part of human interaction and one that plays a central part in expressing human thoughts, emotions, experiences, and cultural values. Systematic research into language facilitates an examination of language from various perspectives. Corpus Linguistics is a relatively new branch of linguistics that provides empirical tools for researching language using authentic data. As opposed to other approaches to language research, which tend to be based only on linguistic intuition, Corpus Linguistics makes use of systematically gathered data to find regularities in language structure. Corpus linguistics is widely recognized as a methodological approach that enables the systematic analysis of authentic language data (McEnery & Wilson, 2001).

Corpus Linguistics is a systematic approach to the study of language regularities in actual texts. When used in the context of literary texts, it offers scholars the opportunity to study recurrent lexical and phraseological features instrumental in the shaping of textual meaning and stylistic expression. Phraseology is an important area of linguistic research that is concerned with the study of recurrent word combinations and conventionalized patterns of language use. It is based on the principle that language is made up not only of individual words but also of multiword units that often co-occur and act as meaningful expressions.

Urdu novels are a rich source of such linguistic patterns as they reflect diverse narrative techniques and culturally embedded forms of language use. The study of phrases taken from Urdu novels is of special importance as phrases are important units of meaning and frequently indicate the grammatical and stylistic inclinations of literary discourse. Hence, a corpus-based analysis of phrases from Urdu novels can provide meaningful insights into the linguistic structures underlying them. Moreover, a selection of novels by two different authors provides an opportunity to compare phraseological patterns and therefore highlights common and unique linguistic features in contemporary Urdu fiction.

One of the most common methods utilized in corpus linguistics is known as Part-of-Speech (POS) tagging. The assignment of grammatical labels according to syntactic functions performed by words such as nouns, verbs, adjectives, adverbs, and determiners is called POS tagging. This process allows for a systematic investigation of grammatical constructions and for researchers to observe patterns of language use. POS tagging allows the distribution and frequency of different grammatical categories in a corpus to be studied and their contribution to textual construction and making meaning to be understood.

Linguistic patterns within textual data are examined using specialised software tools for corpus-based analysis. One popular tool for corpus research is AntConc, which uses many analytical aspects to enable the methodical investigation of language. The frequency of POS-tagged categories and recurrent phraseological patterns in the chosen Urdu novels is examined in this study using AntConc. By offering quantitative insights into the incidence and distribution of grammatical categories throughout the corpus, the tool facilitates the investigation of language structures.

While great research has already been done in the area of Urdu corpus linguistics and POS tagging, the bulk of the work seems to concentrate on the development and improvement of tagging systems instead of analyzing the tagged POS information to discover the linguistic patterns. More specifically, not enough research has been undertaken in exploring the possibilities of using POS-tagged information to study phraseological patterns and recurrent linguistic combinations. Phraseological analysis in corpus linguistics emphasizes the study of recurrent multi-word units and their role in meaning construction and discourse patterns

(Sinclair, 1991). In this way, few studies have dealt with analyzing Urdu literary texts, like novels, from the perspective of corpus-based grammar. This creates a gap that needs to be filled by analyzing phraseological linguistic patterns and structures in literary Urdu texts.

The current study aims to manually extract and POS-tag data to investigate phraseological patterns in a selected number of Urdu novels using a corpus-based methodology. The study employs AntConc to analyze the frequency of POS categories and their patterns of recurrence in order to determine the recurrent grammatical structures inside phraseological units. Additionally, a comparison analysis is carried out to investigate the parallels and discrepancies in the chosen writers' phraseological usage. The study intends to advance knowledge of phraseological elements and grammatical trends in modern Urdu fiction through this methodical corpus-based investigation.

The importance of this research lies in that it adds value to the field of Urdu corpus linguistics and linguistic annotation. In applying methods of manual POS tagging along with techniques from corpus analysis, this research provides an understanding of how Urdu literary phrases have grammatical structures and highlights how the methodologies applied in this field can be employed to analyze Urdu literary texts. This research may also provide future value for other research within various related fields, like computational linguistics and linguistic annotation practices in various languages.

Research Questions

1. What grammatical patterns emerge from the POS tagging of phrases extracted from selected Urdu novels?

2: How do the selected authors differ in terms of POS distribution within phraseological units and lexical density based on the frequency analysis of POS-tagged data using AntConc?

2. Literature Review

Evaluation of different computational methods has been one of the major research areas for Urdu POS tagging. Jaskani et al. (2020) and Ali et al. (2024) compared different approaches of Urdu POS tagging using machine learning and deep learning techniques such as CRF, RNN, SVM, Naïve Bayes, KNN, and CART to achieve improved POS tagging efficiency for Urdu texts. As found by the authors, CRF showed the best results (97%-98% accuracy), thus revealing the efficiency of the CRF model in dealing with Urdu morphology. An additional conclusion of the paper is related to the need for annotated data sets and the importance of linguistic parameters to achieve effective POS tagging. At the same time, it should be noted that the main emphasis of the paper was put on assessing the efficiency of tagging models, which creates a gap since the aspect of phraseology should be considered in detail to understand literary structures better. Hence, the current study extends this line of research by investigating phraseological units in selected Urdu novels through manual POS tagging and corpus-based analysis. The study examines the distribution of POS categories, recurring phraseological structures, and differences in linguistic patterns between the selected authors using AntConc.

Research on Urdu POS tagging has also been extended to the processing of social media-based textual data. Baig et al. (2020) created a corpus of Urdu tweets where the POS tagging was done, and the Urdu Noisy Text POS (UNTPOS) tagset was suggested to cope with problems in tagging the language in the social media domain. The research created a corpus consisting of 5,000 Urdu tweets, provided manual annotations for some of them, and used bootstrapping with the Stanford POS Tagger to enhance tagging effectiveness. The research results have shown that the suggested tagging model functioned well and proved the efficiency

of the annotation of social media data in Urdu. However, the main focus of the research lies in tagging and corpus creation for social media data and does not cover phraseological patterns in literary texts. This gap is significant for the present study as the present research study enhances this area of research by applying POS tagging in exploring phrases from contemporary literary fiction.

Another research direction has been the adaptation of existing linguistic resources and tools for Urdu language processing. Naseem et al. (2017) attempted to investigate the effectiveness of English part-of-speech (POS) taggers in processing Urdu. The authors utilized Urdu tweets, their translations in English, POS-tagging through English POS taggers, and statistical methods for assessment. It was concluded that POS taggers could be used effectively for Urdu text processing. This indicates the possibility of reusing existing POS-taggers to work with resource-scarce languages. Nevertheless, since the study is based largely on machine translation and deals with social media texts only, the generalizability of its findings to other types of texts can be questioned. Moreover, the main interest of the paper lies in POS tagging itself, not in the investigation of specific linguistic/phraseological aspects. This paper is important in terms of the current one, as it stresses the importance of POS tagging in working with the Urdu corpus. However, unlike the above paper, the current research seeks to find and analyze phraseological patterns in Urdu novels. The analysis involves manual POS tagging, as well as the usage of AntConc to discover lexical repetitions and frequency of parts of speech in the selective writings of two authors.

Systematic analysis of stylistic patterns and textual features has been another application of corpus-based approaches in literary studies. McIntyre (2015) analyzed the contribution of corpus stylistics to literary stylistics and suggested an approach to corpus stylistics that would combine corpus stylistic methods along with other forms of stylistic analyses, especially cognitive stylistics. To prove this claim, McIntyre conducted an analysis of patterns and style in the texts from Mark Haddon's novel The Curious Incident of the Dog in the Night-time and TV series Deadwood using corpus annotation, corpus statistics, and stylistic interpretations. Results obtained show that corpus methods can be efficient in discovering meaningful patterns, but for a complete understanding of their stylistic meaning, it is required to integrate them with qualitative methods of stylistic analysis. This study has done a lot towards showing how important it is to combine quantitative and qualitative studies, yet it did not consider phraseology in literary texts. Therefore, there was a certain area left to work on, which is what is done by the current study.

Corpus-based methods in literary studies have offered new ways of approaching the analysis of patterns, structures, and meanings in literary texts. Awanaj and Nofal (2023) performed a corpus-based study of Isabella Hammad’s novel The Parisian to examine the East/West dichotomies from the postcolonial point of view. Taking The Parisian as a corpus consisting of nearly 196,000 words, the researchers carried out a frequency and concordance analysis via AntConc software and extracted repeated linguistic patterns concerning the features, geography, and characters of the novel. The findings reveal that there are various binary oppositions constructed in the novel, which indicate the postcolonial discourses about the East and the West. Even though this study incorporates corpus linguistic techniques to investigate literary works, it concentrates primarily on ideological oppositions rather than linguistic phraseology in the text. Accordingly, the current research contributes to the existing literature by using part-of-speech tagging to explore the phraseological features of Urdu novels.

Corpus-based research has also been used to study morphology and language patterns in Urdu by examining large-scale data. The productivity of Urdu affixes in newspaper writing was examined using a corpus-based method in Shafique et al. (2019). The authors constructed a corpus of over ten million words from Urdu newspapers and analyzed the corpus using corpus analytical techniques to find the most frequent prefixes and suffixes. The results indicated that بے was the most productive prefix, ی was the most productive suffix, and suffixes had greater productivity in Urdu compared to prefixes. This study made an important contribution to Urdu corpus linguistics by highlighting the potential of corpus studies to discover productive linguistic phenomena and affixes. Nonetheless, this study only looked at the morphological aspects of Urdu, particularly affixation and word formation, without analyzing the phraseological patterns and POS tagging. Thus, this paper attempts to advance this research line further by applying POS tagging for analyzing the phraseological patterns of Urdu novels.

Phraseological patterns in learner academic English were explored in Malá's (2020) study employing a corpus-driven framework. The research examined the phraseological properties of academic essays written by Czech university learners vis-à-vis academic writings produced by native-speaking students and academics via corpus-driven techniques, such as frequency lists, keywords, and lexical bundles. It was found that learners overused only a few lexical bundles and discourse-organizing expressions, as opposed to experts who used complicated phraseological patterns. Thus, the research proved the efficiency of corpus-driven approaches in exploring phraseology in academic writing and distinguishing novice from expert writers in terms of their phraseological proficiency. Nonetheless, it considered the problem of linguistic expertise in English academic writing as opposed to literary works. The research thus fails to address the topic of Urdu novels, which is what the present study focuses on.

To increase the precision and effectiveness of autonomous language annotation, a number of computational techniques have been developed. In particular, Onyenwe et al. (2021) and Ishaq et al. (2025) analyzed a number of POS tagging techniques employed in NLP, dividing them into four groups: rule-based, probabilistic, corpus-based, and hybrid. The authors considered the ways of functioning of different POS tagging techniques, such as Constraint Grammar, Hidden Markov Models, Brill Tagger, among others, and assessed their effectiveness in overcoming the problem of ambiguity. As a result, it was revealed that the most frequently used POS tagging techniques are corpus-based and probabilistic as they benefit from contextual analysis and statistical patterns. Nevertheless, even though the article presents valuable insights concerning different POS tagging techniques and their functioning, the main concern of the study is to consider the technique itself rather than its possible application in linguistic studies. Namely, the question of utilizing POS-tagged corpora for phraseology studies remains open.

Recent work in Roman Urdu natural language processing has focused on the development of annotated corpora and the evaluation of advanced POS-tagging techniques. Faheem et al. (2025) constructed a large corpus in Roman Urdu and tested various models for POS tagging that included Hidden Markov Models, neural networks, transformer architecture, and also large language models. With the use of benchmark Urdu corpora converted to Roman Urdu, the authors examined the influence of lexical variations on POS tagging accuracy and found that among different models, the transformer architecture model, such as XLM-RoBERTa, scored the best in terms of accuracy. Likewise, Amin et al. (2025) highlighted that creating high-quality annotated Urdu corpora substantially improves semantic parsing and other NLP applications. The paper made a significant contribution to Roman Urdu NLP by creating POS-tagged

datasets, which can serve as benchmarks for future research. However, the main purpose of the paper is to improve accuracy in POS tagging, and it leaves a gap that the current paper examines the dominant phraseological patterns and uses POS tagging in order to do the comparative analysis of the frequency of parts of speech in Urdu novels of the selected two writers.

3. Methodology

3.1. Research Design

The study used a corpus-based research design to examine phrase patterns in Urdu novels through Part-of-Speech (POS) tagging and AntConc analysis. This approach was chosen because it enables a thorough and objective review of real language data from literary texts. The study used both quantitative and qualitative methods. The quantitative part analyzed frequency distributions of POS categories and words. The qualitative part looked into the interpretation of phrase structures and their linguistic behavior in the corpus. Corpus-based studies allow researchers to identify recurrent patterns of language use across large datasets, providing both quantitative and qualitative insights into linguistic structure (Biber et al., 1998).
The study used manual POS tagging to investigate phraseological tendencies in a few chosen Urdu books. AntConc was used to analyze the tagged data in order to look into recurrent phraseological structures and POS frequency.

3.2. Corpus Development

The set of data created for the purpose of this research attempted to develop a specialized literary corpus. The corpus was designed for contemporary Urdu fiction, and it was built to investigate the phrase level. The texts of Urdu novels written by two contemporary authors made up the existing corpus.
Noor Rajpoot was a popular contemporary Urdu fiction writer who had become popular among online readers of Urdu novels. Her novels dealt with the subjects of romance, suspense, psychological angst, spiritual and social values, and explored emotional experiences, family relationships, and moral insights. Two of her novels, entitled Sulphite (2023) and Hinave (2024), were chosen for this study. While Sulphite looked into topics of love, faith, logic, and spirituality, Hinave looked at love, loyalty, betrayal, resilience, and societal pressures. In modern times, like many other Urdu novelists, Noor Rajpoot had nurtured her readership on digital platforms and episodic publications.
Meher-un-Nisa Shahmeer was an Urdu novelist of his times who was in vogue among Urdu fiction readers in Pakistan and India. She wove themes of romance, suspense, mystery, social issues, character development, loyalty, justice, and emotional experiences into her novels. Two of her novels were chosen for the present study, namely, Baab-e-Dehar (2024) and Bogi Number 12 (2024); both of these novels dealt with issues of friendship, power, betrayal, and mental conflict. Her fiction was chosen because it portrays contemporary Urdu narrative styles and their usefulness in the analysis of the phraseological patterns of the literary texts.
The dataset provided a balanced representation of literary texts and offered scope for comparison of POS distribution within phraseological units in the writings of the selected writers. The corpus of study for the present research comprised 200 phraseological units, which were manually extracted from the selected Urdu novels. The corpus was made up of a total of 1,675 tokens that represent the total number of individual lexical items in the extracted phrases. It includes 734 tokens from Meher-un-Nisa Shahmeer and 941 tokens from Noor Rajpoot's writings. The present study represented the phrase-level corpus, rather than a full-

text corpus, as the study aimed to analyze recurrent phraseological patterns. The selection of phraseological units was purposeful, as these expressions constitute the primary focus of the study and serve as the basis for POS tagging and corpus-based analysis. The POS-Tagged corpus of phrases from both writers is provided in Appendix B.

3.3. POS Tagging Procedure

The first step was to make a corpus by extracting phrases from certain Urdu novels manually. All those phrases that had linguistic and grammatical significance were then arranged in an organized format to form the corpus. Furthermore, manual POS tagging was carried out according to the grammatical category possessed by each word. Then the dataset was converted into a text file to ensure compatibility in order to enable the dataset to be run by AntConc software. The tagsets were used, which included NOUN, VERB, ADJ, ADP, and DET, which were consistently applied throughout the corpus. The abbreviations used for POS Tagging are presented in Appendix A. Ishaq et al. (2025) created a customized Urdu POS corpus capable of identifying nouns, pronouns, verbs, and adjectives and applied machine learning classifiers.

3.4. Data Analysis

The reason why AntConc was selected is that it is efficient in processing textual data, and it can help identify the frequent language patterns that appear in a corpus. The tagged data were then imported into AntConc for analysis of the distribution of various POS categories. It was aimed at revealing the dominant categories and the phraseological structures that are repeated in the text, which include nouns, verbs, adjectives, and other grammatical features. In addition, lexical density was determined as a measure to compare the percentage of content words between the two authors' corpora.

The comparative analysis was carried out to analyse differences and similarities in the phraseological patterns and POS distribution of the selected authors. The overall result of all selected novels was also displayed, which included the number of POS categories, and the examples of the phraseological structure that were found in the corpus were also displayed. The results were presented in a comparative way to get a language picture of the Urdu fiction in modern times.

3.5. Research Reliability and Validity

A systematic and transparent research procedure was used to ensure the reliability and validity of the study. The selected Urdu novels were subject to manual collection of phraseological units based on predetermined criteria, with the help of which all the data were manually POS tagged for consistency of tagging. This tagging system and analysis method were used for both corpora, which increased the reliability of the results. Moreover, the popular corpus analysis tool AntConc helped to ensure the accuracy of frequency analysis and pattern identification. The validity of the study was enhanced by the process of data collection and analysis in accordance with the research objectives, so that the phraseological units that were obtained and the POS categories that were obtained reflect the linguistic pattern studied.

4. Results

4.1. Corpus Description

Table 4.1 Corpus annotation

Number of Novels	4
GENRE	Contemporary Urdu Fiction
Number of Phrases	200
Corpus Size	1,675 tokens
Corpus Format	POS-tagged text corpus

Table 4.1 presents the general characteristics of the corpus used in this study. The corpus consisted of 200 phraseological units extracted from four Urdu novels. After POS tagging and corpus compilation, the dataset contained 1,675 tokens and was converted to plain-text format for analysis in AntConc.

Table: 4.2. Frequency of words

Type	Rank	Freq	Range
کا	6	30	1
کی	6	30	1
میں	8	27	1
کے	9	26	1
سے	10	18	1
تھی	12	12	1
تھا	13	11	1
پر	15	10	1
جانا	17	9	1
کر	17	9	1
کو	17	9	1
آنکھوں	21	6	1
دل	21	6	1
ہو	21	6	1
اور	24	5	1
بے	24	5	1

The most common lexical items from the corpus, such as those that occurred at least five times, are shown in Table 4.2 The words "کا" and "کی" appeared the maximum number of times, that is, 30 times, "میں" (27 times), "کے" (26 times), and "سے" (18 times). Other common words were also both "تھی" (12 times) and "تھا" (11 times), "پر" (10 times), and "کر", "جانا" and "کو" (every

9 times). Words such as 'آنکھوں' and 'دل' were used six times, 'اور' and 'بے' were used five times in the lexical words.

Table 4.3. Part of Speech Frequency

	TYPE	RANK	FREQ	RANG
1.	NOUN	1	329	1
2.	ADP	2	167	1
3.	ADJ	3	122	1
4.	VERB	4	108	1
5.	AUX	5	64	1
6.	PART	11	13	1
7.	DET	13	11	1
8.	CCONJ	15	10	1
9.	ADV	20	7	1
10.	PRON	29	4	1
11.	NUM	63	2	1

The frequency distribution of Part of Speech tags (POS) within the corpus is displayed in Table 4.3. Table 4.3 shows the frequency distribution of the part-of-speech tags (POS) within the corpus. The largest number at NOUN (329) was followed by ADP (167) and ADJ (122). The number of countries experiencing VERB was 108, and for AUX it was 64. The remaining categories had fewer than 13 occurrences each – PART (13 occurrences), DET (11 occurrences), and CCONJ (10 occurrences). The rarest POS tags were ADV (7 times), PRON (4 times), and NUM (2 times).

Table 4.4. Phraseological Patterns frequency

POS PATTERNS	FREQUANCY
ADJ + NOUN	86
NOUN + ADP + NOUN	75
NOUN + NOUN	70
VERB + VERB	19
ADV + VERB	1
ADJ + ADJ + NOUN	10
NOUN + ADJ	50

The phraseological patterns in the corpus, divided by their frequency, are shown in Table 4.4. The most common pattern was ADJ + NOUN (86 times), after which came NOUN + ADP + NOUN (75 times) and NOUN + NOUN (70 times). Another common pattern that was found was the NOUN + ADJ pattern, which occurred 50 times. VERB + VERB appeared 19 times, and ADJ + ADJ + NOUN appeared 10 times, in contrast. ADV + VERB was the rarest pattern with its single occurrence in the corpus. The screenshots of Phraseological Patterns frequencies are provided in Appendix E.

Table 4.5. Phraseological Patterns Examples

POS PATTERNS	EXAMPLES
ADJ + NOUN	زندہ جسموں
NOUN + ADP + NOUN	وقت کی آنکھوں
NOUN + NOUN	سرزمینِ لاہور
VERB + VERB	آنے والے
ADV + VERB	آگے بڑھ
ADJ + ADJ + NOUN	زندہ دلان لوگوں
NOUN + ADJ	خواب ختم

We can present representative examples of the phraseological patterns that are identified from the corpus in Table 4.5. The examples are available, which depict different structural patterns like ADJ + NOUN (زندہ جسموں), NOUN + ADP + NOUN (وقت کی آنکھوں), NOUN + NOUN (سرزمینِ لاہور), VERB + VERB (آنے والے). Other examples involve ADJ + ADJ + NOUN (زندہ دلان لوگوں), and NOUN + ADJ (خواب ختم). The examples provide a picture of the variability of phraseological structures in the corpus.

Table 4.6 Common N-grams in the corpus

Type	Rank	Freq	Range
noun کا adp	1	30	1
noun کی adp	2	28	1
noun میں adp	3	26	1
noun کے adp	3	26	1
noun سے adp	5	18	1

The most common multi-word or N-grams found in the corpus are shown in Table 4.6 NOUN+کا was the most common N-gram, used 30 times. NOUN + کی + ADP came next with a frequency of 28. In total, the N-grams NOUN + میں + ADP and NOUN + کے + ADP were each present 26 times, and NOUN + سے + ADP was present 18 times. The results show that the most frequent patterns in the corpus were noun-based constructions followed by postpositions.

Table 4.7 POS Frequency of writer 1 corpus (Meher-un-Nisa Shahmeer)

Type	Rank	Freq	Range
noun	1	152	1
adp	2	68	1
adj	3	68	1
verb	4	33	1
adv	10	6	1
det	14	4	1

The POS tag distribution in Meher-un-Nisa Shahmeer's Corpus is given in Table 4.7 The most common was the noun (152), followed by the adposition and adjective (68 each). Verbs were used 33 times, the adverbs used were comparatively less, that is, 6 times, and determiners were used 4 times. The results show that a large proportion of the corpus is comprised of noun-based structures.

Table 4.8. Phraseological patterns frequency of writer 1 corpus

POS PATTERNS	FREQUANCY
ADJ + NOUN	47
NOUN + ADP + NOUN	31
NOUN + NOUN	36
VERB + VERB	3
ADV + VERB	1
ADJ + ADJ + NOUN	7
NOUN + ADJ	31

The frequency distribution of phraseological patterns found in the corpus is shown in Table 4.8 47 cases used the ADJ + NOUN structure, and 36 cases used the NOUN + NOUN structure. The number of the NOUN + ADP + NOUN and NOUN + ADJ patterns was 31, respectively. Other less common patterns were ADJ + ADJ + NOUN (7), VERB + VERB (3), and ADV + VERB (1). The results of this study show that there are many kinds of phraseological structures in the selected corpus.

The major categories of POS content words found in the corpus are summarized in Appendix C with some representative examples. The examples show the use of nouns, adjectives, verbs, adverbs, and determiners that have been taken from the selected phraseological units. These lexical items illustrate the extent of the grammatical categories in the whole corpus and are examples of the linguistic material used in the analysis.

Lexical density:

Total Content Words:

NOUN+ADJ+VERB+ADV

152 + 68 + 33 + 6 = 259

Total Tokens: 734

Lexical Density Formula

Lexical Density = (Content Words ÷ Total Tokens) × 100

= (259 ÷ 734) × 100

Lexical Density = 35.29%

Lexical density of the corpus of Meher-un-Nisa Shahmeer was measured by dividing the number of content words (nouns, verbs, adjectives, and adverbs) by the total number of tokens and multiplying the result by 100. Content words in the corpus comprised 259 (35.29%) of the 734 tokens of words. The result shows that the lexical items found in the Corpus are moderate in terms of content, which means that the lexical and grammatical items are used in the author's phraseological expressions in a balanced manner.

Table 4.9. POS Pattern frequency of writer 2 corpus (Noor Rajput)

Type	Rank	Freq	Range
noun	1	180	1
adp	2	99	1
verb	3	75	1
adj	4	50	1
adv	0	0	0
det	13	6	1

The frequency distribution of POS categories in Noor Rajput's corpus is shown in Table 4.9. The most frequent grammatical category was the noun with a total frequency of 180, which

was followed by adposition (99) and verb (75). Adjectives were used 50 times, while determiners were used only 6 times. There were no adverbs found in the corpus. The high rate of noun substantives is in support of the fact that there are fewer phraseological expressions in the author's text in noun forms, and a relatively high rate of verbs indicates that the predominant role is played by action-oriented phraseological expressions in the corpus.

Table 4.10 Phraseological patterns frequency writer 2 corpus

POS PATTERNS	FREQUANCY
ADJ + NOUN	39
NOUN + ADP + NOUN	44
NOUN + NOUN	34
VERB + VERB	16
ADV + VERB	0
ADJ + ADJ + NOUN	3
NOUN + ADJ	19

The following is the frequency distribution of the phraseological patterns that have been identified in Noor Rajput's corpus (see Table 4.10). The most common pattern was NOUN + ADP + NOUN (44) and ADJ + NOUN (39), followed by NOUN + NOUN (34). The NOUN + ADJ pattern was used 19 times, VERB + VERB, 16 times, and ADJ + ADJ + NOUN, 3 times. No example of the ADV + VERB pattern was found. These results suggest that the phraseological structures in the corpus selected are dominated by noun-based and postpositional constructions.

Appendix D lists some representative examples of the major categories of POS extracted from Noor Rajput's corpus. The examples show how the selected phraseological units with nouns, adjectives, verbs, and determiners are used. These words illustrate the variety of the grammatical categories used by the writer and also yield authentic illustrations for the material studied in this research.

Lexical density:

Total Content Words:

NOUN+ADJ+VERB+ADV

180 + 50 + 75 + 0 = 305

Total Tokens: 941

= (305 ÷ 941) × 100

Lexical Density = 32.41%

Lexical density of Noor Rajput's corpus was calculated as the total number of content words (nouns, verbs, adjectives, and adverbs) in his corpus divided by the total number of tokens in those words and multiplied by 100. A total of 305 out of 941 tokens were content words in the corpus, giving it a lexical density score of 32.41%. The results show that there is a moderate level of content-bearing lexical items in the corpus, which means that the author uses lexical and grammatical elements in his/her phraseological expressions in a balanced way.

5. Discussion:

To understand the grammatical phenomena resulting from phraseological units of selected Urdu novels in Urdu, a manual method of POS tagging and a corpus method based on AntConc were adopted in the present study. The study also examined the differences among the selected authors in their POS distribution and lexical density. The results prove the applicability of the corpus linguistic approach to capturing the recurrent grammatical attributes and also illustrate stylistic differences and variation of writing styles among writers. The correlative analysis of the two writers gives more details on the particular trend of phraseology, while the comparative analysis is needed to understand the mannerisms of these two writers in a comprehensive way.

Results of the parallel corpus show that nouns form the overwhelming majority, with the second most frequent group of adpositions (ADP) and adjectives. Since nouns are relatively more numerous, it can be concluded that phraseological expressions in modern Urdu fiction are mostly based on nominal structures. Nouns are often used with detail in literary texts where the author is dealing with the emotions or experiences of the characters or the culture, values, or concepts being introduced, or symbolic thinking, so they tend to be higher-frequency nouns. Likewise, the use of adjectives in Urdu fiction is also quite prevalent, thereby underlining their significance in the descriptions. It shows that the corpus stylistic approaches can reveal meaningful linguistic patterns in literary texts, which is similar to McIntyre's (2015) claim. The abundance of noun derivatives and descriptions shows that the quantitative corpus analysis is able to uncover stylistic tendencies that have not been observed in traditional literary criticism. The phraseological analysis of the combined corpus also indicated that the configuration of ADJ + NOUN, NOUN + ADP + NOUN, and NOUN + NOUN was the most common configuration. These constructions show that many Urdu language fictional texts are highly dependent on descriptive and relational ones in order to construct a phraseological unit. The same was noted by Malá (2020), who affirmed that the use of the corpus-based method is very appropriate to reveal the occurrence of recurrent phraseological structures and lexical combinations. Although Malá's research was on academic discourse, the present results reveal that also in the present discourse, there are recurring phraseological patterns which influence the organisation of the text.

The second research question was regarding the difference among the selected authors in terms of POS distribution as well as lexical density. Through the comparative analysis, both similarities and differences were picked up between the two writers, Meher-un-Nisa Shahmeer and Noor Rajput, where Meher-un-Nisa Shahmeer seems to prefer distinctive stylistic preferences in his phraseological constructions, while Noor Rajput has done the contrary.

Nouns in both corpora appeared the most often, as shown when comparing the frequencies with the non-tabular method. We find 152 nouns in the corpus of Meher-un-Nisa Shahmeer, while 180 nouns are in the corpus of Noor Rajput. A comparison of the two datasets reveals the higher frequency of nouns in both sets of texts, indicating that a nominal construction is an important characteristic of literature in the contemporary Urdu language. Both authors make frequent use of conceptual and descriptive representations, which is likely because nouns are frequent, but they also have symbolic and social uses, as well as uses to denote experiences and emotions. That is a finding that supports the view of McIntyre (2015) that corpus stylistics can be used to identify stylistic patterns of discourse in literature.

Even though somewhat similar, significant differences appeared in other aspects of the grammar. Meher-un-Nisa Shahmeer used significantly more adjectives (68) than Noor Rajput

(50). There is an increase in adjective words, indicating that Shahmeer's writing style is descriptive and imagistic. Adjectives play a very big role in literary works of characterization, emotional effects, and aesthetic representation. Therefore, the percentage of adjectives in Shahmeer's corpus reflects a greater descriptive elaboration and stylistic embellishment.

This is in line with Malá's (2020) observations regarding richer building blocks of phrases and their lexical variability and descriptive complexity. The analysis of academic texts was the field of Mala's work, but it is in the present study that the importance of using descriptive lexical items in order to achieve a stylistic increase and phraseological sophistication is proven.

However, the comparatively higher percentage of verbs was observed in Noor Rajput's corpus. In the corpus, there are 75 verbs, while in the corpus of Meher-un-Nisa Shahmeer, there are only 33 verbs. The big gap indicates that the expression of phrases in Noor Rajput has more action and dynamism. Verbs are essential words for describing procedures, movement, and change. The increased use of verbs, then, suggests that Rajput's style of writing focuses on what he does in his stories and what happens to him, rather than just on descriptive representations.

The larger number of verbs in Noor Rajput's corpus indicates that he tends to create meaning based on activities and processes, which results in a relatively more narrative and event-oriented style. This is observed to confirm McIntyre's (2015) view that sometimes the corpus method can detect stylistic differences among individual authors. The distinction between the two writers, according to the differences discovered, is evidence of the objective distinction of authors' styles for the purpose of a corpus-based approach.

The analysis of the patterns of the phrases showed that there was also a stylistic difference between the authors selected. The most common structure in the corpus of Meher-un-Nisa Shahmeer was ADJ + NOUN, which was repeated 47 times. On the other hand, Noor Rajput used this pattern 39 times. Considering the high percentage of constructions of ADJ + NOUN in Shahmeer's corpus, it can be inferred that the writer's style is very descriptive. With such build-ups, writers are able to build up meanings, create images, and express feeling nuances. Adjective-based constructions are used very frequently in the corpus, clues important to understand, which means that Shahmeer's phraseological style is very closely linked with the work of literary description and symbolic expression.

On the other hand, Noor Rajput showed a greater leaning towards the NOUN + ADP + NOUN structures, 44 times, whilst Shahmeer's corpus showed 31 such structures. This structure is used more and more in the enrolled text, suggesting that Rajput often conveys semantic connections via postpositions. The grammar of Urdu depends a lot on postpositions, which are used to denote the possessive, locative, and associative meanings. The high incidence of NOUN + ADP + NOUN constructions is therefore an indication that Rajput's phraseological expressions highlight relational and contextual meanings.

Besides, Noor Rajput used VERB + VERB constructions many more times (16 times) than did Meher-un-Nisa Shahmeer (3 times). It can also be noticed here that almost all the phrases are verbal combinations giving vent to the dynamism of Rajput's compositional style. On the other hand, the multiplicity of multi-verbal constructions suggests that the phraseological expressions used by Rajput are relatively more process-oriented, as they can show complex action, process, and continuity. Unlike that, Shahmeer's employment of verbal patterns seems to be minimal, further highlighting his ability to use descriptive and static imagery.

Comparing the phraseological structure of both authors, one can conclude that the author has used the noun-based structures predominantly, but has adopted different, and consequently,

differing performances of them to create the different stylistic effects. Meher-un-Nisa Shahmeer seems to feature phrases characterised by their descriptive and/or imagery-oriented content, and Noor Rajput shows a leaning towards dynamic and/or relational phrases. The results found are comparable to what McIntyre (2015) suggests, which is the use of systematic linguistic analysis for the identification of individual stylistic preferences of corpus stylistics.

Another finding of significance is on lexical density. Lexical density is generally considered to be the index of informational content as well as the lexical richness of a text. Lexical density analysis revealed that Meher-un-Nisa Shahmeer's corpus resulted in a score of 35.29% while Noor Rajput's corpus resulted in a score of 32.41%. The level of lexical density for both scores is in the middle range, while Shahmeer's corpus, with its higher score, is certainly more lexically rich.

In fact, the lexical density of Shahmeer's corpus is higher than that of the others, which we can attribute mainly to the use of adjectives and descriptive lexical items. There is a stronger correlation between the percentage of content words and semantic complexity and informational richness. Hence, this suggests that the phraseological expressions used by Shahmeer have more density of meaning-bearing lexical items in the discourse that produce a denser, more descriptive style.

However, the relative lack of lexical density can be attributed to Noor Rajput in the case of the other samples, as his content is more grammatical and relational, using mostly postpositional verbs and clauses. The variations in the difference between the two authors are not significant, but rather do reflect differences in phraseological resources in the way they use them. The same was said by Malá (2020), who claimed that the contrast in phrase complexity and in word complexity is frequently a stylistic preference of different writers. Recently, the same discursive model can be extended to the Urdu literary field with the notion of lexical density as an author's style discriminative quantitative measure, as is one of the present findings.

The results obtained from this study also help other Urdu POS tagging research. Previous research done by Jaskani et al. (2020), Baig et al. (2020), and Faheem et al. (2025) was mainly concerned with the validation and the creation of annotated corpora. These studies greatly contribute to the field of Urdu NLP, but did not use POS tagged literary corpora to explore the stylistic and phraseological potential. This work continues the study line, contributing to new findings and showing that POS tagging has turned into a useful tool for gaining insights into the phraseological structure, authorial variation, and stylistic features in Urdu fiction literature. Likewise, Muaz et al (2009) highlighted the need for building Urdu POS tagged corpora for computational linguistic analysis. Their work was dedicated to enhancing tagging resources in Urdu, and it was observed that tagged corpora serve as the basis for systematic analysis and study of Urdu grammar.

In general, the obtained results indicate that corpus-based methods are useful in gaining insights into Urdu PEs, and from these, it can be inferred that the Urdu literary corpus has been properly organized according to the grammatical level of the PEs in it. The study shows that nouns are much more common and preponderant constructions, description is common in Urdu fiction, and postposition is a commonly used construction in contemporary Urdu fiction. Comparative analysis, on the other hand, shows that authors have various and different ways of using these units of grammar to create individual-style grammars, as well as to present unique styles. The results, thus, reinforce the validity and potential of corpus linguistic approaches in stylistic investigation of literary texts and are added to the ever-expanding area

of corpus-based Urdu literary research. The result is consistent with the findings of Rajper et al. (2021), who emphasized the fact that POS tagging and corpus-based methods are effective techniques for detecting linguistic patterns and grammatical categories of Urdu and other Pakistani languages.

6. Conclusion

This study examined the phraseological patterns of selected Urdu novels by doing manual POS tagging and AntConc analysis using a corpus-based approach. The objective of the study was to find out the grammatical structures used in the phraseological units and to explore the difference between the works of Meher-un-Nisa Shahmeer and Noor Rajpoot in terms of POS distribution within the phraseological units. The study used a combination of quantitative corpus analysis and qualitative interpretation to identify the linguistic patterns and how they can shed light on the stylistic features and authorial preferences in Urdu literary discourse.

The results revealed that the most dominant category in the combined corpus was nouns, followed by adpositions and adjectives. The analysis was also carried out to reveal the common phraseological structures and stylistic differences between the author's works. The comparative analysis focused on comparing and contrasting the two writers. Both writers displayed a marked tendency to use noun-based constructions, but there were some differences in their use of other grammatical categories. The differences in the phraseological organization were highlighted in the corpus of Meher-un-Nisa Shahmeer, as it had a higher occurrence of adjectives and lexical density, while the occurrence of verbs and verbal structures was higher in the Noor Rajpoot corpus. These results suggest that it is possible to obtain meaningful information from the use of POS distribution and lexical density of individual writing styles.

The study is significant in Urdu Corpus Linguistics as it took the research from the accuracy of POS Tagging and Corpus development to literary phraseology and stylistic variations. It also highlights the value of AntConc in studying the frequency patterns and regular patterns of words in Urdu fiction. Future studies may involve the inclusion of a wider range of literature texts, authors, and genres, and the use of advanced computational techniques for more comprehensive linguistic analysis in future research.

References

Ali, M., Khan, M., & Alharbi, Y. (2024). *A conditional random field based approach for high-accuracy part-of-speech tagging using language-independent features*. *PeerJ Computer Science, 10*, e2577. https://doi.org/10.7717/peerj-cs.2577

Amin, M. S., Zhang, X., Mazzei, A., Ahmad, A., Khan, W., & Bos, J. (2025). *Semantic processing for Urdu: Corpus creation, parsing, and generation*. *Language Resources and Evaluation*. Advance online publication. https://doi.org/10.1007/s10579-025-09819-2

Awajan, A., & Nofal, M. Y. (2023). Corpus-assisted analysis of East/West binary oppositions in Isabella Hammad's *The Parisian*. *Journal of Educational and Social Research, 13*(4), 153–170. https://doi.org/10.36941/jesr-2023-0098

Baig, A., Rahman, M. U., Baloch, A., & Khan, H. (2020). A part-of-speech tagged dataset for Urdu noisy text. *Computers, 9*(4), 90. https://doi.org/10.3390/computers9040090

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Faheem, A., Ullah, F., Azam, U., Ayub, M. S., & Karim, A. (2025). Part of speech (POS) tagging in Roman Urdu: Datasets and models. *Language Resources and Evaluation, 59*, 4285–4312. https://doi.org/10.1007/s10579-025-09865

Ishaq, M., Asim, M., Ahadullah, Cengiz, K., Konecki, M., & Ivković, N. (2025). Development of novel Urdu language part of speech tagging system for improved Urdu text classification. *Social Network Analysis and Mining, 15*, Article 58. https://doi.org/10.1007/s13278-025-01473-4

Jaskani, F. H., Amin, M. T., Shaikh, H., Azam, M. A., Khuhawar, F., & Aqeel, M. (2020). Comparative analysis of Urdu parts of speech taggers using machine learning techniques. In *2020 IEEE 23rd International Multitopic Conference (INMIC)* (pp. 1–6). IEEE. https://doi.org/10.1109/INMIC50486.2020.9318205

Malá, M. (2020). Phraseology in learner academic English: Corpus-driven approaches. *Discourse and Interaction, 13*(2), 75–88. https://doi.org/10.5817/DI2020-2-75

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction* (2nd ed.). Edinburgh University Press.

McIntyre, D. (2015). Integrating corpus stylistics and cognitive stylistics: Studying the opening of *The Curious Incident of the Dog in the Night-Time*. *Language and Literature, 24*(2), 152–175. https://doi.org/10.1177/0963947014557980

Muaz, A., Ali, A., & Hussain, S. (2009). Analysis and development of Urdu POS tagged corpus. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7), ACL-IJCNLP 2009* (pp. 24–31). Association for Computational Linguistics and Asian Federation of Natural Language Processing.

Naseem, A., Anwar, M., Ahmed, S., Satti, Q. A., Hashmi, F. R., & Malik, T. (2017). Tagging Urdu sentences from English POS taggers. *International Journal of Advanced Computer Science and Applications, 8*(10), 231–238. https://doi.org/10.14569/IJACSA.2017.081030

Onyenwe, I., Ikechukwu-Onyenwe, O., & Ebele, O. (2021). A review on part-of-speech technologies. *International Journal of Computer Science, Engineering and Applications, 11*(5–6), 1–7. https://doi.org/10.5121/ijcsea.2021.11601

Rajper, R. A., Rajper, S., Maitlo, A., & Nabi, G. (2021). Analysis and comparative study of POS tagging techniques for national (Urdu) language and other regional languages of Pakistan. *Sindh University Research Journal (Science Series), 53*(4), 44–53.

Shafique, H., Shahbaz, M., & Ahmad, I. (2019). The productivity of Urdu affixes in newspapers: A corpus-driven research. *Pakistan Social Sciences Review, 3*(1), 112–129.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.