ISSN Online: 3006-4708
ISSN Print: 3006-4694
SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com

A Comparative Morphological and Computational Analysis of Saraiki and Urdu Verbs

Luqman Manzoor[1], Noshaba Bano*[2], Muhammad Riaz Khan Dasti[3]
[1] MPhil Scholar, Government College University Faisalabad. luqmanmanzoor875@gmail.com
[2] MPhil Scholar, Government College University Faisalabad.
*Corresponding Author: noshababano10@gmail.com
[3] Department of English, University of Layyah. riazdasti@ul.edu.pk

DOI: https://doi.org/10.70670/sra.v4i1.1885

Abstract
This paper presents a comprehensive comparative analysis of verb morphology in Saraiki and Urdu, two prominent Indo-Aryan languages of Pakistan. The researchers combine two research methods in their study: a corpus-based study of Saraiki and a computational analysis of Urdu based on established linguistic rules. The Saraiki study results show that the grammatical system of Saraiki verbs uses inflectional forms to create approximately 81% of all verb forms. The research identifies 40 morphological patterns that exhibit a Zipfian distribution, with the five most common suffixes accounting for more than 53% of all verb tokens. The study demonstrates that the most common suffix -w-n serves multiple functions, occupying a position between the grammatical categories of inflexion and derivation. The Urdu analysis demonstrates high generative power, producing 49,879 inflected forms from 975 base verbal roots. The two languages share an Indo-Aryan base structure, but their specific morphological features show different patterns of development. Urdu uses a complex system of honorifics with four levels of respect, while Saraiki uses a simpler system that serves its communication needs.

Keywords: Comparative Indo-Aryan Linguistics, Morphological Polyfunctionality, Zipfian Distribution, Finite State Transducers, Split-Ergative Alignment , Sociolinguistic Honorifics

1. Introduction

1.1 Background: A Sociolinguistic and Typological Context

The languages of South Asia create a complex linguistic situation that stems from their common ancestry with the Indo-Aryan languages, which began with ancient Prakrit and Apabhramsha. Among the New Indo-Aryan (NIA) languages, Urdu and Saraiki provide researchers with a valuable opportunity to compare their linguistic features. The national language of Pakistan, Urdu, maintains its official status through educational institutions and government offices, thereby shaping its sociolinguistic and grammatical development. More than 25 million people in Southern Punjab speak Saraiki, a Northwestern Indo-Aryan language. The language has been used by many speakers but excluded from formal settings, limiting its use to informal and literary contexts. The differences in status between Saraiki and Urdu actually determine how their speakers create their different verbal systems. The Saraiki Movement and present-day linguists consider Saraiki to be a separate language, while Grierson (1916) classified it as a dialect of "Lahnda" (Western Punjabi). The phonological inventory of Saraiki distinguishes it through its system of implosive consonants, which Punjabi and Urdu lack. The research paper conducts a detailed comparative analysis of Saraiki and Urdu verb patterns through two studies: a comprehensive corpus analysis of Saraiki verb morphology and a computational study

of Urdu verb rules. The report compares two research methods that yield different results, providing detailed information on their linguistic features and their theoretical and practical effects on language structure.
The study demonstrates how people who speak the same language utilize different rules of word formation to create new words, thereby changing their language. These languages pose difficulties for researchers attempting to build computational models for both major and minor languages. The research method applies the basic principles of Corpus Linguistics, as established by Sinclair in 1991, because it studies actual language use in real-world conditions rather than developing complete theoretical models.

1.2 Statement of the Problem

Research on the evolution of Indo-Aryan languages has produced numerous studies of their individual morphological systems. Existing studies on Saraiki language research have focused on describing the language without establishing an empirical foundation through corpus-based research, thereby restricting both linguistic analysis and language technology development. The assessment of Urdu-language research has mainly focused on generating content through established patterns without tracking how often people actually use those patterns in everyday conversations. The absence of research comparing these two languages, based on actual data, between their common roots and their present state, creates a major obstacle to understanding their linguistic evolution.

1.3 Purpose and Significance

The main objective of this study is to address a research gap by combining corpus-based documentation of Saraiki with rule-based documentation of Urdu. The approach enables research on linguistic elements, including the distribution patterns of different grammatical structures and their use in everyday communication. The analysis holds multiple important reasons for its existence. The study provides the first empirical research on Saraiki verb morphology, an under-researched language. The research provides a test case that enables linguistics to study how inflexion and derivation interact, and how frequency affects the development of morphological patterns. The research results show direct benefits for computational linguistics, as they support the development of digital tools, including morphological analyzers and POS taggers, for both languages.

1.4 Research Questions

The following research questions have guided this study:
1. What are the core inflectional and derivational patterns of verbs in Saraiki and Urdu, respectively?
2. How do the morphological structures of both languages encode tense, aspect, mood (TAM), and agreement (person, number, gender)?
3. What are the key typological similarities and differences, and what are their implications for linguistic theory and computational modelling?

1.5 Theoretical Framework

The study uses an integrative framework which combines three linguistic fields, Corpus Linguistics and Construction Morphology (CxM) and Distributed Morphology (DM), as its theoretical foundation. CxM describes the multiple roles of the suffix -w-n̲, which it performs in both inflexion and derivation, while Distributed Morphology explains the process of converting abstract grammatical features into phonological forms during the time after syntax creation. DM explains how Saraiki speakers exhibit different pronunciation patterns through the use of suffixes that alter the sound of the root khā (eat) in specific aspectual contexts. The framework enables us to create a model which demonstrates that the base syntactic structure stays the same while the visible form undergoes modifications.

2. Literature Review: The Indo-Aryan Verb System

2.1 Saraiki: A Language with a Rich and Complex Identity

The linguistic status of Saraiki has been a subject of long-standing debate. The Saraiki Movement and modern linguists have established Saraiki as a separate language because Grierson's 1916 study classified it as a dialect of Western Punjabi. The Saraiki language shows its special characteristics through its phonological system, which includes sound elements that Punjabi and Urdu speakers do not use, and through its use of a customized Perso-Arabic writing system to represent these particular sound elements. The movement managed to unite the Multani and Jatki dialects through its efforts, which received official recognition during national censuses. The language faces difficulties in academic institutions and urban areas because of the dominance of Urdu and English. The need for linguistic documentation and digital resources arises directly from these patterns of sociolinguistic exclusion. Recent studies have begun to address this research gap by moving from descriptive methods to data-driven solutions that address real-world problems. The researchers Manzoor et al. (2025) created an educational framework for English verb instruction by using a 2-million-word Saraiki corpus. The study shows that about 50% of English verbs use high-frequency morphological patterns, including the causative suffix -w-n. The framework uses Saraiki-speaking learners' morphological expertise to turn language complexity into a resource for learning the target language.

2.2 An Overview of Urdu Verb Morphology

Urdu is a morphologically rich language, with its morphology primarily based on suffixation. A single verb can convey a wealth of information about person, gender, number, tense, and respect. The single word پڑھیے (parrhain, meaning "read") can be used to indicate a singular second person imperative with a high level of respect, a plural second person imperative, or a subjunctive form for a third person with a second level of respect. The language's verbal system defines its essential polysemy through this specific aspect. The research shows that Urdu verbs have 57 inflected forms, demonstrating their complex morphological patterns. Urdu, which belongs to the Indo-Aryan language group, has a split-ergative system: present-tense verbs require subject agreement, while past-tense transitive verbs require agreement with the direct object.

2.3 Shared Indo-Aryan Typology

The Saraiki and Urdu languages differ in many respects, yet they share a basic typological foundation that defines the Indo-Aryan languages. The two languages share a common linguistic feature, which shows split-ergative morphosyntactic alignment. The two languages use nominative-accusative alignment in the present tense, indicating that the verb agrees with the subject. The two languages use ergative-absolutive alignment in the past tense of transitive verbs, which requires the verb to match the direct object's gender and number. The morphological characteristics of both languages create a linguistic spectrum that lies between agglutinative and fusional language systems. The Saraiki analysis presents the language as having an "agglutinative nature" because speakers use multiple stacked suffixes, whereas the Urdu analysis shows that its suffixes consist of single-syllable forms that carry multiple person-number and respect features. Multi-feature suffixes function as a defining characteristic of fusional morphology. The observation shows that both languages use suffixes as their main structure, but they also mix elements from both language systems, resulting in a greater synthetic nature than English, which operates as an analytic language.

3. Methodology: A Synthesis of Approaches

3.1 Saraiki: The Corpus-Based Approach

The Saraiki research used an empirical methodology which followed the Corpus Linguistics principles to conduct its bottom-up study. The research team established a new corpus of 2 million words by combining literary works with journalistic writing.

Saraiki and Urdu exhibit distinct linguistic traits, yet they share a common Indo-Aryan language family, which

serves as their primary structural basis. The languages share a common linguistic feature that divides their morphosyntactic systems into two parts via split-ergative alignment. Both languages use nominative-accusative alignment, which requires verb agreement with the subject. The languages use ergative-absolutive alignment in the past tense of transitive verbs because the verb must match the gender and number of the direct object.

The two languages display a morphological relationship which extends from agglutinative through fusional typological frameworks. The Saraiki analysis of the language states that its "agglutinative nature" stems from the use of stacked suffixes, whereas the Urdu analysis explains that its suffixes include single-syllable elements that exhibit multiple features, including person, number, and respect. The presence of multi-feature suffixes in a language demonstrates that the language follows fusional morphological patterns.

The researchers developed a comprehensive language study based on actual language data, which they used to create a study framework. The study aimed to develop an evidence-based description of Saraiki's verbal system, presenting actual language patterns rather than theoretical possibilities. The research team studied 552 verb tokens from the corpus, sorting them by morphological function before conducting quantitative and frequency analyses.

3.2 Urdu: The Rule-Based Computational Approach

Computational research on the Urdu language required a top-down approach, applying predefined rules within its computational system. The research process aimed to create every possible inflected version which could result from using the complete set of verbal roots. The study used a list of 975 verbal roots to generate a comprehensive list of 49,879 inflected forms, averaging 51.16 forms per verb. The research used a two-layer morphological analyzer that operated via Finite State Transducers to evaluate both the proposed morphological rules and the newly developed classification system. The research's main objective was to establish a robust classification system that included complete morphological rules to generate all verb forms while maintaining efficient computational resources and strong production capabilities.

3.3 The Comparative Framework

This paper's methodology involves a synthesis of these two divergent approaches. The Saraiki data provides a usage-based picture of a system's actual communicative function, while the Urdu data provides a generative picture of a system's full potential. The analysis of these two languages allows for drawing conclusions which extend beyond a basic comparison of their verb forms. The research investigates how different morphological processes create functional language components for assessment through sociolinguistic factors, which include prestige and honorifics, and how these factors create distinct modelling challenges and advantages for each language.

4. Results: Comparative Data and Analysis

4.1 The Morphological Balance: Inflexion vs Derivation

The corpus-based analysis of Saraiki provides a clear quantitative measure of its morphological balance. The manual annotation of 552 verb tokens shows that inflectional morphology dominates the data because it constitutes 81 per cent of all verb forms. The research demonstrates that Saraiki is a highly inflectional language, as evidenced by its established qualitative characteristics. The corpus data consists of 19% that involves derivational morphology to create new lexical items and modify verb valency. The Urdu analysis generates 47 distinct inflectional forms by examining 975 verbal roots, although it does not provide comparable percentage data.

4.2 Verb Classification and Morphological Patterns

The two studies demonstrate their methodological differences through their distinct approaches to verb

classification. The Urdu analysis establishes its rule-based objective through a complete classification system which uses phonological endings and irregular behaviour patterns. The Saraiki study identified 40 morphological patterns from the corpus. The quantitative analysis of these patterns demonstrates a sharp Zipfian distribution because a few high-frequency suffixes create most verb usage, while 35 patterns produce less common specialized grammatical functions. The analysis of Table 1 shows that the first five suffix patterns together produce more than 53% of all examined tokens.

Table 1: The Five Most Frequently Occurring Saraiki Verb Patterns

RANK	PATTERN (SUFFIX)	MORPHOLOGICAL FUNCTION	FREQUENCY (% OF TOTAL)
1	*-w-n̲*	Causative / Purposive Participle	17.0%
2	*-ī-sī*	3rd Person Present (Singular/Plural)	16.3%
3	*-ī-s-n*	3rd Person Present Plural Agreement	8.1%
4	*-w-ī-sī*	3rd Person Present	6.7%
5	*-th-ī-sī*	3rd Person Present	5.1%

The data show that Saraiki speakers rely on a single main group of suffixes for essential communication, whereas Urdu speakers need multiple rules to handle their full range of verb forms.

4.3 The Causative: A Deep Dive

The two languages show their core connection through the causative, which serves as an essential derivational process. The Saraiki analysis shows that the suffix -w-n̲ functions as a crucial morpheme, performing dual functions: it serves both as a standard derivational suffix for causative formation and as an inflectional marker for purposive participles. The usage-based evidence establishes a new relationship between inflexion and derivation which applies to the Saraiki language system. Both languages use suffixation to create causatives, but Saraiki uses its primary causative morpheme (-w-n̲), which functions as a purposive participle marker, creating extreme polyfunctionality that does not exist in Urdu, which has distinct morphological patterns for its causative markers. Urdu maintains its causative system through a strict system that categorizes "irregular consonant verbs" as the only valid causative forms. The stem forms create these words through suffix addition, which transforms the base form into new meanings: marnā 'to die' → mārnā 'to kill' → marwānā 'to have killed'

Table 2: Derivational Causative Formation in Saraiki

VERB ROOT	GLOSS	DERIVATIONAL SUFFIX	CAUSATIVE FORM
بھاگ (BHĀG)	To flee	**-وٹ** /-w-n̲/	**بھاگوٹ** (bhāg-w-n̲)
بول (BOL)	To speak	**-وٹ** /-w-n̲/	**بولوٹ** (bol-w-n̲)

پڑھ (PAṚH)	To read	-وٹ /-w-n̲/	پڑھوٹ (paṛh-w-n̲)

Table 2 demonstrates how the multifunctional -w-n̲ suffix serves as an essential derivational mechanism for forming causative verbs from intransitive or simple transitive roots. The Saraiki language system uses this morphological change to demonstrate how one morpheme can produce multiple different meaning transformations.

4.4 Tense, Aspect, Mood (TAM) and Agreement

The Saraiki data provide a complete numerical demonstration of its agreement system, which shows a strong preference for third-person agreement. The analysis discovered that 88 per cent of finite verb forms in the corpus display third-person arguments because the source texts contain both narrative and descriptive content. The Saraiki language uses morphological agreement patterns that mark gender differences in its singular forms, including the masculine -dā and the feminine -dī. Indo-Aryan languages commonly exhibit this pattern of gender distinction, evident in their plural forms. Urdu's agreement system employs the same grammatical rules but uses a comprehensive honorific system to distinguish different levels of respect. The two languages exhibit this distinction because their sociolinguistic functions require different systems: Urdu, as an administrative language, demands multiple social status indicators alongside polite speech. The word paṛhaiñ serves as a single basic form that allows users to express three grammatical structures: a second-person imperative with high respect, a plural second-person imperative, and a third-person subjunctive. The following table shows all the ways a single verb can be used, demonstrating complex verb usage.

Table 3: Urdu Honorific and Agreement Levels

(Example: *bolnā* — to speak)

LEVEL OF RESPECT	PRONOUN	VERB FORM	SOCIAL CONTEXT / MEANING
INTIMATE	*tū*	*Bol*	Close friends, children, or addressing a deity
FAMILIAR	*tum*	*bolo*	Close family, friends, or younger peers
FORMAL	*āp*	*boliye*	Elders, strangers, or professional settings
RESPECTFUL	*āp*	*boliye gā*	High-status individuals or formal requests

The Urdu system of pronunciation has four distinct levels of respect, expressed through a single form. In contrast, the Saraiki system uses documented direct agreement methods. Saraiki research shows that the present habitual/progressive is the most common tense-aspect form. Urdu uses a hybrid TAM system, combining suffixes with aspectual auxiliaries to express temporal aspects. The phrase paṛh rahā hai means "is reading."

4.5 Quantitative Summary

The two research methods produced separate sets of results, presented in the following tables. The Saraiki study tests a particular corpus through empirical research, while the Urdu study applies a generative rule-based approach, yielding different research results on Indo-Aryan verb morphology.

Table 4: Saraiki Corpus-Based Findings

METRIC	VALUE
CORPUS SIZE	2 million words
TOTAL VERB TOKENS ANALYZED	552
UNIQUE MORPHOLOGICAL PATTERNS	40
INFLECTIONAL VS. DERIVATIONAL	81% vs. 19%
MOST FREQUENT PATTERN	*-w-n̲* (17% of all tokens)

Table 5: Urdu Rule-Based Generative System

METRIC	VALUE
VERBAL ROOTS ANALYZED	975
TOTAL INFLECTED FORMS GENERATED	49,879
DISTINCT INFLECTIONAL TYPES	47
AVERAGE FORMS PER VERB	51.16

5. Discussion: Interpretation and Implications

5.1 Bridging the Methodological Divide

The study of corpus-based Saraiki research, together with rule-based Urdu research, demonstrates that people need both empirical, usage-based studies and generative, rule-based studies to fully grasp how a language operates its morphological systems. The Saraiki study provides a practical guide for developing natural language processing systems in low-resource languages, emphasizing the importance of building essential vocabulary patterns to achieve maximum system performance. The Urdu data demonstrate that researchers need to implement a comprehensive rule-based system that can handle all its morphological forms, including

the rare yet essential honorific forms.

5.2 Theoretical Implications

The comparative study results generate three contributions which enhance our understanding of Indo-Aryan verb systems. The data from both languages collectively support a model in which the Indo-Aryan verbal system lies on a hybrid continuum between agglutinative and fusional typologies. The Saraiki analysis provides strong, data-based evidence supporting Construction Morphology's claim that one morpheme or construction can perform multiple functions across different inflectional and derivational processes. The suffix -وݨ /-w-n̲/ demonstrates this particular linguistic phenomenon through its ability to perform multiple functions. The morphophonemic variations in the Saraiki data demonstrate how Distributed Morphology principles explain the morphological process by which abstract syntactic features reach their final surface form through post-syntactic stages. The Saraiki data serve as an effective means of evaluating formal morphological theories.

5.3 Sociolinguistic Implications

The different social functions of Saraiki and Urdu create distinct morphological patterns which shape their respective linguistic systems. The complex honorifics system of Urdu establishes social rank distinctions through its various levels of politeness. The relationship between grammar and social role functions is co-evolutionary: complex honorifics developed from formal language requirements, which established grammatical rules that enabled administrative and prestigious language functions.

5.4 Methodological Grappling

The two studies display different research methods; which researchers need to recognize as separate. The comparison between Saraiki and Urdu shows their different capacities to create new content: the former shows only one tree through its usage, while the latter displays all forest pathways through its entire grammar system. The Saraiki data reveals common patterns used in under-resourced language technologies, showing which patterns occur most often. The Urdu data shows all morphological forms needed to establish formal communication.

5.5 Practical Implications for Pedagogy and Technology

Pedagogical Efficiency: The Zipfian distribution presents immediate benefits for Saraiki language teaching after its discovery. The top five suffix patterns (including -w-n̲, -ī-sī, and -ī-s-n) account for more than 50% of verbal usage patterns, making these "high-yield" morphemes essential for teaching. Students who learn this limited vocabulary base will achieve complete language understanding because it includes all essential grammar rules.

The research results yield specific outcomes that advance the development of language technology. The current rule-based classification system in Urdu establishes basic elements which our study used to create a reliable morphological analyzer that produced approximately 50,000 word forms. The corpus-based findings on Saraiki provide an efficient approach to developing its digital resources, given the language's limited online content. Developers achieve high NLP application coverage by focusing on frequent verb patterns using a finite-state transducer. Developing digital infrastructure is essential for this language because it lacks a complete rule-based system. The research on low-frequency patterns also provides insights for creating more specialized applications.

6. Conclusion

The comparative analysis of this research study establishes a comprehensive system that explains all Saraiki verb patterns, together with their corresponding rule-based systems in Urdu. The quantitative analysis of the

Saraiki corpus revealed a pronounced third-person bias, with nearly 88% of finite verb forms indexing third-person arguments. The corpus shows this finding because it contains literary and journalistic texts which have a narrative and descriptive style, but it does not represent an inherent grammatical rule of the Saraiki language. The two languages share an Indo-Aryan linguistic base, leading to split-ergative alignment and a combination of fusional and agglutinative features, but they develop different functional usage. Saraiki uses a distributional system that yields a Zipfian pattern, revealing all its effective suffixes, whereas Urdu requires multiple generative systems to handle its complex system of honorifics. The distinct data points serve as a guide for creating customized NLP technologies that acknowledge the specific ways each language communicates and generates content.

6.1 Limitations of the Study

The study's findings face multiple limitations that limit their validity. The Saraiki data originate from a corpus that consists mainly of literary and journalistic materials and does not fully capture the full range of spoken language variation. The study examines the Central Derawali dialect, but its results do not apply to all regional Saraiki dialects. The rule-based Urdu analysis fails to reflect the actual frequency of verb use in everyday speech, which would have enhanced the analysis.

6.2 Future Research

The current study establishes a basic framework that identifies key research gaps and provides specific guidelines for subsequent studies. The main research objective requires cross-dialectal field studies to document and compare verb usage across various Saraiki dialects. The Saraiki corpus needs expansion to include a wider range of spoken language registers, which are currently underrepresented. The creation of a standardized morphological tag set for Saraiki is a fundamental requirement for developing a comprehensive supervised machine learning system that supports NLP tasks such as morphological analysis, part-of-speech tagging, and syntactic parsing. The report deepens our understanding of the Indo-Aryan verb system while presenting a research roadmap that includes upcoming interdisciplinary studies needed to sustain these valuable linguistic systems.

Author Contributions

L. Manzoor: Conceptualization and development of the research framework.
N. Bano: Manuscript preparation, original draft writing, and data formatting.
M. R. K.Dasti: Supervision, critical review, instructional guidance, and APA 7th edition compliance.
All authors have reviewed the final manuscript, contributed to the interpretation of results, and approved the version for submission.

References

Atta, F., & Rasheed, S. (2019). Morphophonemic variations in the Saraiki language. *International Journal of Linguistics, Literature and Translation*, *2*(3), 42–53. https://doi.org/10.32996/ijllt.2019.2.3.6

Grierson, G. A. (1916). *Linguistic survey of India* (Vol. IX, Part I). Office of the Superintendent of Government Printing, India.

Lowe, J. J., & Birahimani, A. H. (2019). The argument structure of Siraiki causatives. In M. Butt & T. H. King (Eds.), *Proceedings of the LFG'19 Conference* (pp. 191–211). CSLI Publications.

Manzoor, L. (2025). *Morphological investigation of verb patterns in Saraiki: A corpus-based study* [Unpublished master's thesis]. Government College University, Faisalabad.

Manzoor, L., Bano, N., Majeed, Z., & Naeem, R. (2025). A corpus-driven pedagogical framework for teaching English verbs to Saraiki speakers: Leveraging morphological patterns from a 2-million word corpus analysis. *Qualitative Research Journal for Social Studies*, *2*(4), 20–39.

Niazi, A. (2018). Morphological analysis of Urdu verbs. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science* (Vol. 9624, pp. 284–293). Springer. https://doi.org/10.1007/978-3-319-75477-2_19

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

University of Minnesota. (n.d.). *Grammar notes 7.3: The Hindi causative construction*. Hindi-Urdu. https://open.lib.umn.edu/hindiurdu/chapter/grammar-notes-7-3-the-hindi-causative-construction/

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.