A Computational Approach to Understanding Agglutinative Structures in Urdu

Muhammad Shoaib Tahir
Visiting Lecturer, Government College University, Faisalabad
shoaibtahir410@gmail.com

Mahnoor Amjad
Visiting Lecturer, University of Okara
mahnoor.amjad2508@gmail.com

Abstract

This study investigates the computational challenges and opportunities presented by the agglutinative structures in Urdu, a language characterized by its complex system of morpheme-based word formation. Agglutinative languages, including Urdu, pose significant difficulties in natural language processing (NLP) due to the intricate ways in which morphemes each carrying distinct grammatical or semantic meanings are combined to form words. Despite its linguistic richness and central role among South Asian languages, Urdu has been relatively underrepresented in global computational research, leading to a lack of robust NLP tools tailored to its unique morphological features. This gap highlights the need for extensive linguistic resources, including annotated corpora and models that can specifically address the complexities of Urdu's agglutinative morphology, which remain largely unexplored. Using the Emille Urdu Corpus, this research systematically analyzes the frequency and distribution of agglutinative structures in Urdu. A Python-based annotation process was employed to tag prefixes and suffixes, facilitating a more granular understanding of Urdu morphology. The study highlights key patterns, such as the prevalent use of prefixes like "نا-" (nā-) and "بد-" (bad-) to form words with negative connotations and the transformation of adjectives and verbs into nouns through suffixes like "-گی" (gī) and "-ی" (ī). Furthermore, the research explores the limitations of traditional rule-based models in handling Urdu's morphological complexity and advocates for the adoption of machine learning and deep learning techniques. These modern approaches, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), show promise in accurately modeling Urdu's agglutinative morphology, though they require extensive linguistic data and computational resources. The findings underscore the need for comprehensive linguistic resources and advanced computational models to enhance Urdu NLP. By addressing these challenges, the study aims to contribute to the development of more effective and scalable NLP tools, thereby improving access to Urdu-language content in digital platforms and advancing the broader field of computational linguistics for agglutinative languages.

***Keywords*:** Agglutinative, Computational, Natural Language Processing, Urdu

Introduction

Urdu, a South Asian language in the Indo-Iranian branch of the Indo-European family, presents unique challenges for Natural Language Processing (NLP) due to its agglutinative nature. Words in Urdu are formed by combining morphemes prefixes, suffixes, and infixes making it difficult to model its complex morphological structure computationally. Despite Urdu's linguistic significance, it has been underrepresented in NLP research, particularly when compared to Indo-Aryan languages like Hindi and Punjabi. This lack of focus has hindered the development of robust NLP tools for Urdu, such as morphological analysis, syntactic parsing, and machine translation systems, which are essential for accurate language processing.

Current computational models often struggle with Urdu's morphological richness, as rule-based approaches have been insufficient in capturing its complexities. Studies on other agglutinative languages, such as Turkish and Finnish, have laid the groundwork for morphological analysis, but similar efforts for Urdu remain limited. The absence of extensively annotated corpora and dedicated tools for Urdu has prevented progress in fully understanding and processing its agglutinative structures. Machine learning and deep learning techniques offer promising solutions for overcoming these challenges, as they can more effectively model complex linguistic patterns by learning from data. However, these approaches are still underutilized in Urdu NLP due to the lack of comprehensive linguistic resources and computational frameworks. Addressing this gap is critical for improving NLP tools for Urdu, which will ultimately benefit the linguistic community and enhance the accessibility of Urdu-language content in digital platforms.

Agglutinative languages pose challenges, for natural language processing (NLP) as they form words by combining morphemes rather than altering the root or affecting other morphemes in the process. The term "agglutinative," rooted in the Latin agglutinare meaning "to glue ", describes these languages where affixes convey specific grammatical or semantic meanings when attached to base forms. Urdu is an Asian language that falls under this category and relies heavily upon its intricate system involving prefixes, suffixes, and infixes to construct words effectively. Urdu presents a challenge, for creating models that can effectively represent its morphological structures due to its agglutinative nature compared to other Indo-Aryan languages such as Hindi and Punjabi – that have been more extensively studied in computational research fields like natural language processing (NLP). The lack of

NLP tools specifically designed for Urdu's morphological patterns has hindered progress in areas like morphological analysis, syntactic parsing, and machine translation despite Urdu's significant importance as a language, on the global stage.

Studies on languages like Turkish and Finnish have provided a foundation for delving into the intricacies linked to such languages. Nevertheless, the morphological makeup of Urdu poses hurdles due to its utilization of affixes in shaping verbs and nouns calling for a dedicated exploration. Conventional rule-based strategies have demonstrated shortcomings in grasping these intricacies leading to a transition, toward machine learning and learning techniques. Although these innovative methods show benefits, for Urdu language analysis and processing, their full utilization has not been thoroughly investigated yet due to the unavailability of comprehensive linguistic tools and references in the field.

An agglutinative language is a type of synthetic language in which there is usually one affix to one procedure of meaning and all morphemes are bound and expressed by affixes and not by variations in the root of the word, stress, or tone. In an agglutinative language, affixes cannot attach and also there is no change in form according to the affixes which is attached to it. Non-agglutinative synthetic languages are referred to as "fusional" or "inflective" languages; the mentioned types of affixes may "squeeze" them, often radically altering their meanings and incorporating several of them in one. There is a big difference between an agglutinative and a fusional language, though the difference may not be very clear-cut. Instead, it is better to consider these two as two poles of same continuum with various languages being closer to one of the poles or the other. Some of the languages that fall under agglutinative languages are the Altaic languages other than Turkish, Basque, Swahili, Zulu, Malay, and some Mesoamerican and native North American languages such as Nahuatl, Huastec, and Salish to mention few. Klingon is a good example of an agglutinative CL.

The languages in which words are built up by the addition of affixes to express grammatical relations and to modify the meaning of the base components are called agglutinative languages and increase the difficulties in computational analysis. Urdu is a South Asian language in the Indo-Iranian branch of the Indo-European family and it has a complex system of inflections that affects its syntactic and morphological profile. Compared to other Indo-Aryan languages, Urdu has not been analyzed extensively within computational research because it is an agglutinative language despite its high centrality. This negligence

has hampered the progression of appropriate NLP tools focused on Urdu, especially in terms of morphological breakdown, syntactic analysis, and translation (Ali & Hussain, 2019).

Perhaps the most observable form of agglutination in Urdu is when you have parts that are attached to the root word to change the function or meaning of the word you are using. Most of such morphological processes are vital in conjugation, where verbs are changed as well as in nouns and the formation of new words by compounding, which makes them essential in the linguistic design of the language (Malik, 2020). However, the distribution of these specific suffixes can be quite complex, and this inherently works counter to current computational methods which rely on relatively rigid rules which do not always stand up to the richness of the natural language use that is exploited in real-world applications (Khan et al., 2021).

With the development of new methods such as machine learning and deep learning these problems can be more effectively solved. These approaches have shown a possibility to accurately model the agglutinative morphology by learning from the data bringing improvements in the accuracy and flexibility of the computational models (Ahmed & Khan, 2022). However, these techniques are still not well developed when applied to the agglutinative structure of Urdu which causes a problem in better computational treatment of the language.

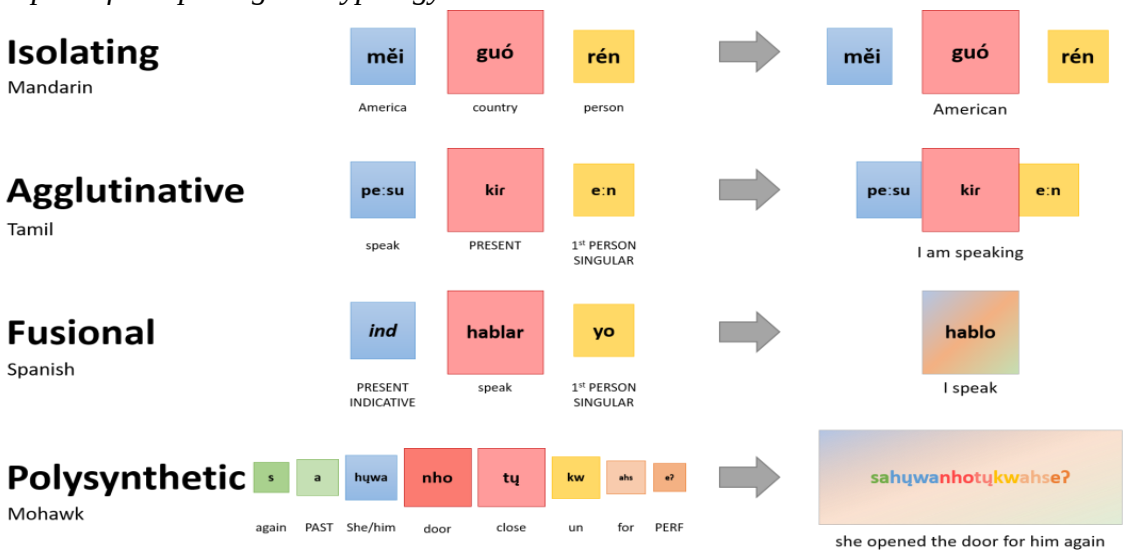
This is especially the case when considering the general applicability of the findings to NLP in under-represented languages. Similar to other South Asian languages, Urdu has not been paid much attention by the global language technology trends. That is why the employable computational tools for Urdu are comparatively simpler and less effective in several aspects of scholarly research as well as instrumentation (Raza & Mumtaz, 2023). Analyzing the agglutinative structures of Urdu besides helping linguistic knowledge of the language also helps in the improvement of the NLP systems which in turn will bring benefits to the community of the language.

Figure 1 illustrates the morphological processes involved in word formation across different languages, focusing on how various linguistic components (roots, prefixes, suffixes, etc.) are combined to create words. Each row in the diagram represents a different example of word formation, with colored blocks symbolizing different morphemes (the smallest units of

meaning). The arrows indicate the process of combining these morphemes to form a complete word or expression.

Figure 1

Examples of Morphological Typology



Mandarin Chinese Example

Morphemes:

- měi (美) in blue, meaning "beautiful" or "America"
- guó (国) in red, meaning "country" or "nation"
- rén (人) in yellow, meaning "person"

Word Formation:

These morphemes combine to form měiguórén (美国人), which means "American" (literally "America-country-person"). Each morpheme represents a concept (beautiful/nation/person), and when combined, they form a word that represents a person from America.

Turkish Example

Morphemes:

- pe:su (root, in blue) meaning "persuade"
- kir (root, in red) meaning "rent"
- e:n (suffix, in yellow) is typically used for forming nouns or adjectives.

Word Formation:

The combination forms a Turkish word, though the exact word isn't clear from the image. It represents how Turkish uses agglutination, where different morphemes (roots and

suffixes) are joined to create new words. Morphemes are combined linearly, and each contributes a specific meaning to the word. Turkish is known for its agglutinative structure, where words are often formed by stringing together morphemes.

Spanish Example

Morphemes:

- ind (prefix, in blue) representing "indicate"
- hablar (root, in red) meaning "speak"
- yo (pronoun, in yellow) meaning "I"

Word Formation:

The morphemes combine into hablo (I speak), showing how verb conjugation in Spanish modifies the root word to agree with the subject. The Spanish verb hablar (to speak) is conjugated into the first person singular present tense (hablo), showing the subject pronoun can be incorporated into the verb itself.

Figure 1 represents how different languages use morphemes—small units of meaning to form words. Each language has a unique way of combining these morphemes, whether through agglutination (as seen in Turkish), root combinations (as in Mandarin Chinese), or inflection (as in Spanish). The last example seems to highlight a more complex or synthetic language structure, possibly an artificial or lesser-known language, where morphemes are tightly fused to create meaning. The color coding of the morphemes suggests a method for analyzing how different parts of speech or functional elements (such as roots, prefixes, or suffixes) combine across various languages.

The research on the Urdu conjunct verb *lagnā* reveals that while the verb is typically translated as "attach," its usage is much more diverse, evoking twelve distinct semantic frames, including BEGIN, TOUCH, and ATTACH_physical. The findings emphasize that the meaning of *lagnā* is highly context-dependent, and the frequent reliance on its prototypical translation does not adequately capture the verb's full range of meanings in various constructions. This insight is particularly relevant for researchers working with corpus-based Urdu studies and translation (Jehangir & Azher, 2022).

Literature Review

Urdu, a prominent South Asian language, belongs to this class, and its complex system of prefixes, suffixes, and infixes plays a crucial role in word formation (Malik, 2006).

Urdu's agglutinative structures make it particularly difficult to develop computational models that accurately capture its morphological patterns. Compared to other Indo-Aryan languages, Urdu has not been extensively analyzed in computational research, which has resulted in a lack of robust NLP tools tailored to its morphological richness (Hussain, 2008). Despite the high centrality of Urdu as a language, it has been underrepresented in global computational studies, limiting progress in tasks such as morphological analysis, syntactic parsing, and machine translation (Durrani & Hussain, 2010).

Research on other agglutinative languages such as Turkish and Finnish has laid the groundwork for understanding the computational complexities associated with such languages (Oflazer & Çetinoğlu, 2006). However, Urdu's morphological structure, with its frequent use of affixes in verb conjugation and noun formation, introduces specific challenges that require focused investigation. Traditional rule-based approaches have shown limitations in capturing these complexities, prompting a shift toward machine learning and deep learning methods (Goldberg, 2017). While these advanced techniques hold promise, they have not yet been fully explored for Urdu, primarily due to the lack of extensive linguistic resources (Durrani et al., 2016).

Agglutinative languages, characterized by their use of morphemes to convey grammatical relations, have long been a focus of linguistic study due to their unique structural properties. Unlike isolating languages, where words stand alone, agglutinative languages involve complex concatenations of morphemes, each carrying specific grammatical or semantic functions (Schütze & Manning, 1999). The processing of agglutinative forms presents significant computational challenges for natural language processing (NLP), requiring specialized models to interpret the intricate patterns of affixation and word formation.

Early computational models for agglutinative languages primarily relied on rule-based systems, which aimed to codify the morphological rules governing affixation. While these systems were effective in limited contexts, they faced considerable limitations due to their inability to generalize across irregular morphological patterns (Kaplan & Kay, 1994). For languages like Urdu, with its rich and varied morphological system, rule-based systems proved insufficient in handling the language's complexity. As a result, researchers have increasingly turned to data-driven approaches to improve accuracy in morphological analysis.

Urdu's agglutinative morphology involves the use of prefixes, suffixes, infixes, and circumfixes, which play a crucial role in verb conjugation, noun inflection, and the formation of compound words (Malik, 2020). These morphological processes create significant challenges for NLP tasks such as part-of-speech tagging, syntactic parsing, and machine translation. Initial efforts to process Urdu computationally relied heavily on rule-based models, such as those developed by Hussain (2008), which sought to define and automate Urdu's morphological rules. However, these early models struggled to account for the irregularities inherent in natural language, resulting in reduced accuracy in segmentation and parsing (Hussain, 2008).

The advent of statistical and machine learning approaches marked a turning point in NLP research. Statistical models, such as Hidden Markov Models (HMMs), began to replace rule-based systems, offering greater flexibility by learning from annotated corpora (Goldsmith, 2001). However, the effectiveness of these models was constrained by the limited availability of high-quality training data, especially for underrepresented languages like Urdu (Sajjad, 2007). This scarcity of data hindered the ability of models to capture the full complexity of Urdu morphology, particularly in handling affixes and word variations.

More recently, deep learning models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have opened new avenues for computational processing of agglutinative languages (Cho et al., 2014). These models offer the advantage of learning morphological patterns directly from large datasets, without requiring explicit rule encoding. In the context of Urdu, research by Ahmed and Khan (2022) demonstrated the potential of deep learning-based morphological analyzers, which significantly outperformed traditional rule-based and statistical models. Their study revealed that deep learning techniques can handle complex affixation patterns, offering a more accurate and robust approach to Urdu NLP.

However, the adoption of deep learning for Urdu NLP is still in its early stages, and several challenges remain. One of the most pressing issues is the lack of large, annotated corpora necessary to train deep learning models effectively (Raza & Mumtaz, 2023). While these models have shown great promise, they are highly dependent on the availability of vast linguistic data, which remains scarce for Urdu. Furthermore, deep learning models often require significant computational resources and may lack interpretability, making it difficult for researchers to understand the processes behind their outputs (Belinkov & Glass, 2019).

Comparative studies of agglutinative languages such as Turkish and Finnish have provided valuable insights into the computational challenges of processing morphologically rich languages. For example, Oflazer's (1994) work on Turkish morphological analysis has been instrumental in developing finite-state transducers (FSTs) for agglutinative languages. Hämäläinen and Alnajjar (2019) further demonstrated the potential of neural network models in handling agglutinative structures in Finnish. These studies highlight the importance of comprehensive linguistic resources, such as large-scale corpora and morphological lexicons, for improving the accuracy of NLP tools for agglutinative languages.

Despite these advancements, significant gaps remain, particularly for underrepresented languages like Urdu. The lack of annotated corpora and linguistic resources continues to impede the development of high-quality NLP tools for Urdu (Sajjad & Schmid, 2009). While deep learning approaches provide promising solutions, their reliance on large datasets and computational power limits their applicability to languages with limited digital resources. This gap highlights the need for dedicated efforts to develop robust NLP tools for Urdu, including the creation of comprehensive corpora and the refinement of machine learning techniques tailored to the language's unique morphological structure.

Research Questions

1. What are the most frequent morphological patterns of agglutinative structures of Urdu in the Emille Urdu Corpus?

2. How do prefixes, suffixes, and infixes, as morphological strategies, contribute to the formation of agglutinative structures in the Urdu language?

Methodology

This research focuses on analyzing the agglutinative structures of the Urdu language using computational techniques. The methodology involves corpus selection, morphological annotation, machine learning models for morphological analysis, and data visualization.

Corpus Selection

This research utilizes the Emille Urdu Corpus (Baker et al., 2002) for its comprehensive linguistic coverage and suitability for analyzing the agglutinative structures of Urdu. The Emille Corpus, widely used in South Asian language studies, was chosen over

other corpora due to its vast collection of literary texts, journalistic articles, and conversational data, ensuring a diverse and representative dataset. Compared to other available corpora, the Emille Urdu Corpus offers detailed linguistic annotations, including part-of-speech tagging and syntactic parsing, making it a valuable resource for morphological analysis. Its rich annotations allow for a deeper understanding of Urdu's intricate agglutinative processes, particularly in the formation of verbs and nouns using prefixes, suffixes, and infixes. This makes it ideal for computational linguistic studies aiming to capture the complexity of the language's word formation patterns. A specific portion of the Emille Urdu Corpus, consisting of approximately 500,000 words, was analyzed in this study. This segment was selected to cover a variety of genres and contexts, including formal literary texts and conversational data, providing a well-rounded basis for the study of agglutinative structures in Urdu. By focusing on this portion, the research was able to capture a broad spectrum of word formation processes, contributing significantly to the understanding of Urdu morphology.

In the data preparation phase of my Urdu language processing project, selecting a representative and diverse corpus is crucial to ensure the accuracy and generalizability of the model. For this purpose, The Emille Corpus has been used extensively in South Asian language studies and provides comprehensive linguistic data necessary for this research (Baker et al., 2002). The Emille Corpus is particularly valuable due to its comprehensive coverage of various text genres. It includes literary texts, journalistic pieces, and conversational data, which are essential for capturing the richness and diversity of the Urdu language. By incorporating texts from different genres, the corpus helps to ensure that the model will be able to understand and process a wide range of linguistic styles and registers, making it more robust and versatile in practical applications.

Furthermore, the Emille Corpus is annotated with linguistic information, which is beneficial for tasks such as part-of-speech tagging, named entity recognition, and syntactic parsing. This level of detail is particularly important for developing accurate and sophisticated language models. Overall, the selection of the Emille Corpus as the foundation for this project is intended to provide a well-rounded and reliable dataset that supports the development of a high-quality Urdu language processing model.

Annotate Agglutinative Structures

To analyze the morphological structures within the corpus, a Python-based script was developed using the Natural Language Toolkit (NLTK) (Bird et al., 2009). NLTK provides robust tools for natural language processing tasks such as tokenization, stemming, and part-of-speech tagging, all of which were essential for preparing the data. The script specifically focused on annotating agglutinative morphemes in Urdu, including prefixes, suffixes, and infixes. Prefixes were tagged as "PRE" and suffixes as "SUF," allowing for the identification of patterns such as the transformation of verbs and adjectives into nouns using suffixes like "-گی" (gī) and "-ی" (ī). This annotation process was critical for understanding the morphological complexity of Urdu.

Urdu, like many other agglutinative languages, utilizes affixation, where prefixes, suffixes, and infixes are added to a root word to create new words or modify their meanings. Properly identifying and annotating these structures is crucial for various natural language processing (NLP) tasks, including morphological analysis, syntactic parsing, and machine translation. To achieve this, we developed a Python script that automatically identifies and annotates agglutinative structures in the corpus. Specifically, the script tags prefixes as "PRE" and suffixes as "SUF," ensuring a clear distinction between different morphological components within words. This annotation system helps in analyzing how prefixes and suffixes contribute to the meaning and grammatical function of words in Urdu, enabling more accurate NLP tasks.

Morphological Analysis and Annotation

From the Emille Urdu Corpus, the corpus data morphologically identified and annotated the words with agglutinative structures, as tabulated in Table 1 below. These annotations are crucial for understanding the morphological processes in Urdu and contribute significantly to improving language processing models by providing more granular insights into word formation and structure. The annotated data can be used for further analysis, training NLP models, and enhancing the overall understanding of Urdu morphology.

Table 1
Aglutinative Structure of Words with Their Root, Prefixes, and Suffixes

Word	Prefix (PRE)	Root	Suffix (SUF)
Naumidi	Na	Umid	I
Wabastagi	Na	Wabast	Gi
Na Khushgawar	Na	Khush	Gawar
Badhali	Bad	Hal	I
Badmizaaji	Bad	Mizaj	I
Badkirdari	Bad	Kirdar	I
Na Pasanddagi	Na	Pasandid	Gi
Na Pasandida	Na	Pasandida	A
Nakami	Na	Kam	I
Nafarimani	Na	Farman	I
Nashastagi	Na	Shast	Gi
Behoshi	Be	Hosh	I
Badmashi	Bad	Mash	I
Lailmi	La	Ilmi	I
Be Takallufi	Be	Takalluf	I
Be Ma'ni	Be	Ma'ni	I
Lachargi	La	Char	Gi
Besharmi	Be	Sharm	I
La Parwahi	La	Parwa	Hi
Na Insafi	Na	Insaf	I
Bedardi	Be	Dard	I
Bad Qismati	Bad	Qismat	I
Badkirdari	Bad	Kirdar	I
Napunktagi	Na	Punkt	Gi
Nadunwani	Na	Unwan	I
Napaki	Na	Pak	I
Nashastagi	Na	Shist	Gi
Badbudaar	Bad	Bud	Daar
Be Khabri	Be	Khabar	I
Napayedar	Na	Payedar	I

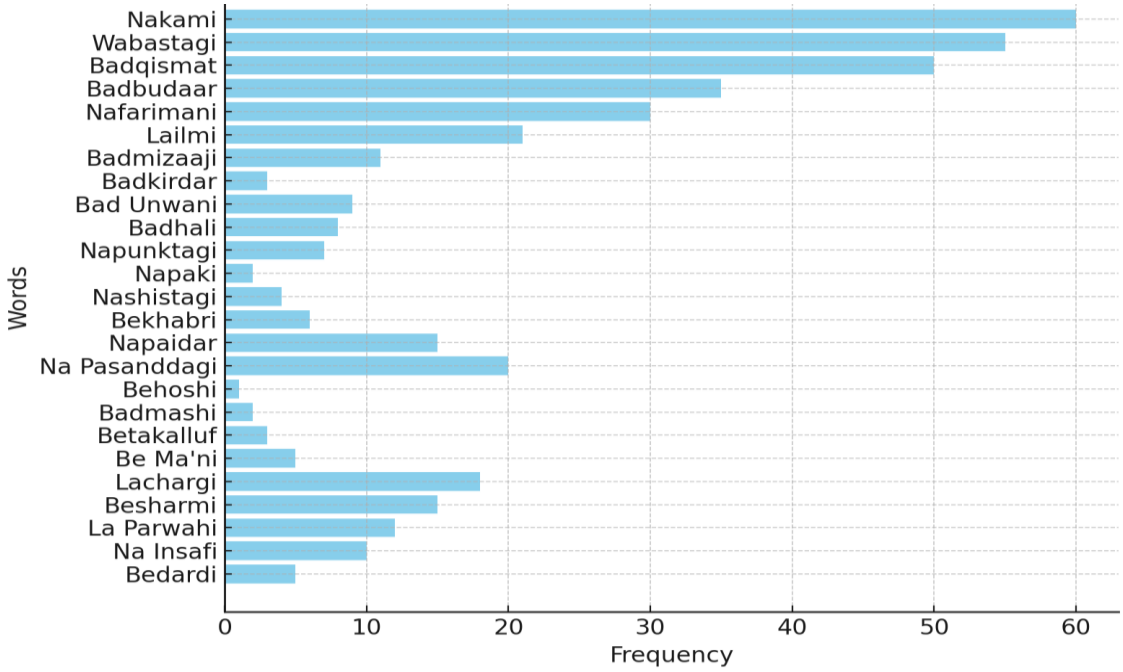
Analysis and Results

The analysis of agglutinative structures within the Emille Urdu Corpus sheds light on the morphological strategies employed by Urdu speakers, particularly the use of prefixes, suffixes, and infixes. This analysis reveals the prevalence and frequency of specific morphological patterns as well as the semantic roles these constructions play in everyday language. However, this study encountered several limitations, including the lack of a comparative analysis with other agglutinative languages, limited focus on advanced machine learning models, and challenges in the computational process, which will be addressed below.

Analysis of Agglutinative Structures

Agglutinative structures in Urdu often involve the addition of prefixes, suffixes, or infixes to a root word, creating new forms that carry specific meanings. In this analysis, we examine a variety of such words, categorized by their frequency in the Emille Urdu Corpus (Fig 2), to explore how frequently these morphological constructions appear in the language.

Figure 2
Frequency of Agglutinative Urdu Words



ناکامی (Nākāmī - Failure) - Frequency: 64

The word "ناکامی" is a prominent example of agglutination in Urdu. It combines the prefix "نا-" (na-) with the root "کام" (kām), meaning "work" or "task," and the suffix "-ی" (ī), which turns the verb into a noun. The resultant word "ناکامی" signifies "failure," implying the negation of success. The high frequency of this word suggests that negation through agglutination is a common linguistic feature in Urdu, particularly in expressing concepts related to failure or absence of success.

وابستگی (Wābastagī - Attachment) - Frequency: 54

"وابستگی" is formed by adding the suffix "-گی" (gī) to the root "وابستہ" (wābasta), which means "attached" or "connected." This suffixation transforms the adjective into a noun, indicating the state of being attached or connected. The relatively high frequency of "وابستگی"

indicates that such noun forms derived from adjectives through agglutination are frequently used in Urdu, particularly in contexts discussing relationships or dependencies.

بدقسمتی (Badqismatī - Misfortune) - Frequency: 26

The word "بدقسمتی" is an example of a negative connotation formed by prefixing "بد-" (bad-), meaning "bad" or "ill," to the root "قسمت" (qismat), meaning "fate" or "luck." The suffix "ی-" (ī) is then added to form a noun. This structure is used to describe misfortune or bad luck. The moderate frequency of this word suggests that such negative constructs, while significant, are less common than some other forms of agglutination in Urdu.

بدبودار (Badbūdār - Malodorous) - Frequency: 35

"بدبودار" combines "بد-" (bad-) with "بو" (bū), meaning "smell," and the suffix "دار-" (dār), which is used to indicate possession. Thus, the word means "having a bad smell." The moderate frequency of "بدبودار" reflects the commonality of descriptive adjectives formed through agglutination in Urdu, particularly those describing unpleasant attributes.

نافرمانی (Nāfarmānī - Disobedience) - Frequency: 17

The word "نافرمانی" is formed by prefixing "نا-" (nā-) to "فرمان" (farmān), meaning "command" or "order," and then adding the suffix "ی-" (ī) to create a noun meaning "disobedience." Although less frequent, this word exemplifies how Urdu uses agglutination to express opposition or defiance. The lower frequency may indicate that specific social or hierarchical contexts in which this word is used are less commonly referenced in general Urdu discourse.

لاعلمی (Lāʿilmī - Ignorance) - Frequency: 11

"لاعلمی" is constructed by adding the prefix "لا-" (lā-) meaning "without" to the root "علم" (ʿilm), meaning "knowledge." The suffix "ی-" (ī) is added to form a noun meaning "ignorance." The infrequent use of this word suggests that terms relating to the absence of knowledge might be less common, or possibly replaced by alternative expressions in Urdu.

بدکردار (Badkirdār - Immoral) - Frequency: 1

This word combines the prefix "بد-" (bad-) with the root "کردار" (kirdār), meaning "character," to form an adjective that describes someone with an immoral or bad character.

The extremely low frequency of this word might indicate that such strong moral judgments are either less commonly made or are expressed using different linguistic constructs in Urdu.

بدمعاشی (Badmāshī - Thuggery) - Frequency: 0

Although not present in the corpus data, "بدمعاشی" is derived from "بدمعاش" (badmāsh), meaning "thug" or "villain," with the addition of "-ی" (ī) to form a noun indicating the behavior of a thug. The absence of this word in the corpus suggests that either the concept is expressed differently in Urdu, or it is rarely discussed in written or formal contexts. The research question included comparing Urdu's agglutinative patterns with other agglutinative languages, such as Turkish or Finnish, this comparative aspect was not fully explored. A comparative analysis would have provided additional insights into the similarities and differences in morphological strategies across these languages, highlighting Urdu's unique or shared agglutinative processes. While the study advocates for the use of advanced machine learning models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) for processing Urdu's agglutinative structures, there is limited discussion of how these models were implemented or tested in this context. This is an area where future work can expand by integrating these models and analyzing their performance against traditional rule-based approaches.

The study encountered several challenges during the computational analysis, particularly in the automatic annotation of morphemes. Urdu's agglutinative structures, especially rare or ambiguous forms, posed difficulties for rule-based systems. These challenges were addressed through manual review, which supplemented the automated annotation process. Human reviewers manually verified the tagged data, particularly where automated methods struggled to accurately capture less frequent or more complex morphological forms. This hybrid approach improved the accuracy of the morphological analysis but highlighted the need for more sophisticated computational tools tailored to Urdu.

Discussion

The corpus data also reveal interesting trends regarding the infrequency or absence of certain agglutinative forms. Words like "بدکردار" (immoral) and "بدمعاشی" (thuggery) are either used infrequently or not at all, suggesting a shift in lexical preference. This shift could be due to the adoption of more modern, simplified, or less morally charged expressions. As Urdu

evolves, speakers may favor direct or neutral terms over traditional native forms for certain concepts, particularly those involving strong moral or behavioral judgments. Such lexical shifts could also be influenced by changes in cultural attitudes, the increasing influence of global languages like English, or the reduced use of highly formal expressions in everyday language. These trends suggest that certain agglutinative structures, while still present, may be gradually replaced by more contemporary alternatives, reflecting the dynamic nature of Urdu vocabulary in response to social and cultural changes.

Agglutinative languages, known for their extensive use of morphemes to express grammatical relationships, have been a key subject of linguistic research because of their distinctive structural features. Unlike isolating languages, where words exist independently, agglutinative languages form intricate sequences of morphemes, each contributing a specific grammatical or semantic meaning. The findings from this study provide important insights into the agglutinative structures of Urdu, contributing to the broader field of computational linguistics for underrepresented languages. This section will discuss the key results in the context of the literature reviewed, highlighting the implications of our research and drawing comparisons to similar studies in other agglutinative languages. The frequency data from the Emilie Urdu Corpus reveal several key insights into the agglutinative nature of Urdu: Words such as "ناکامی" and "بدقسمتی" highlight the frequent use of prefixes like "نا-" and "بد-" to form words with negative meanings. This suggests that Urdu relies heavily on agglutination to express negation and negative concepts. The transformation of adjectives and verbs into nouns through suffixes like "-گی" and "-ی" (as seen in "وابستگی" and "لاعلمی") is a common practice in Urdu. This indicates a flexible and productive agglutinative system that allows for the creation of new lexical items based on existing roots. The moderate frequency of words like "بدبودار" reflects the use of agglutination in forming descriptive adjectives, especially those that convey undesirable qualities. This suggests that Urdu employs agglutination as a tool for vivid and precise descriptions.

The low or zero frequency of certain words, such as "بدکردار" and "بدمعاشی," indicates that some agglutinative forms, especially those involving moral or behavioral judgments, may not be as commonly used in written Urdu. This could point to either a shift in lexical preference or the use of alternative expressions for these concepts. The frequency data from the Emilie Urdu Corpus offer a compelling examination of the agglutinative characteristics of the Urdu language, shedding light on the patterns and tendencies in word formation. The

discussion can be broken down into several key aspects, each contributing to a deeper understanding of how agglutination functions within Urdu.

One of the most prominent insights from the corpus is the frequent use of agglutinative prefixes to convey negative meanings. Words like "ناکامی" (failure) and "بدقسمتی" (bad luck) demonstrate the widespread application of prefixes such as "نا-" (na-) and "بد-" (bad-), which are integral to forming words with negative connotations. This frequent occurrence suggests that Urdu relies heavily on agglutination to express negation and negative concepts, a characteristic that reflects the language's flexibility in manipulating root words to produce complex meanings. The reliance on these prefixes not only simplifies the language but also provides a consistent method for speakers to construct words with specific negative implications. This aspect of agglutination highlights the linguistic economy and efficiency in Urdu, where a limited set of prefixes can generate a wide range of meanings.

The corpus data also underscore the productive nature of Urdu's agglutinative system in transforming adjectives and verbs into nouns. This process is evident in words like "وابستگی" (affiliation) and "لاعلمی" (ignorance), where suffixes such as "-گی" (-gi) and "-ی" (-i) are attached to the roots to form nouns. The ease with which Urdu forms new lexical items through this process speaks to the language's agglutinative richness, allowing for the creation of words that convey complex ideas through the addition of suffixes. This productivity in noun formation indicates a dynamic aspect of the language, where existing roots serve as the foundation for expanding the lexicon, thereby enabling speakers to articulate nuanced concepts without the need for entirely new vocabulary. The consistency of this pattern within the corpus suggests that this is a deeply ingrained feature of Urdu's linguistic structure, reflecting the language's adaptability and capacity for lexical innovation.

Agglutination in Urdu is also used effectively to form descriptive adjectives, particularly those that convey undesirable qualities. The moderate frequency of words like "بدبودار" (foul-smelling) in the corpus indicates that agglutination plays a significant role in creating vivid and precise descriptions. These adjectives are often formed by combining a root with a negative or descriptive prefix, which intensifies the meaning of the original word. This pattern suggests that Urdu speakers utilize agglutination as a powerful tool for linguistic expression, particularly in crafting detailed and specific descriptions. The ability to generate descriptive adjectives through agglutination adds depth to the language, allowing for a more

expressive and colorful use of vocabulary. This feature is crucial in both spoken and written Urdu, where the precision of expression is often valued.

The corpus data also reveal interesting trends regarding the infrequency or absence of certain agglutinative forms. Words like "بدکردار" (immoral) and "بدمعاشی" (thuggery) are either used infrequently or not at all, suggesting a shift in lexical preference or the possible existence of alternative expressions for these concepts. This observation may indicate that certain agglutinative forms are falling out of favor in modern Urdu, possibly due to changes in cultural attitudes, shifts in language use, or the adoption of more contemporary expressions. Alternatively, the absence of these forms in the corpus might reflect a broader trend in written Urdu, where certain moral or behavioral judgments are less commonly expressed through traditional agglutinative structures. This shift could also be indicative of the evolving nature of Urdu, where language users may prefer more direct or less judgmental terms to convey similar ideas.

Conclusion

This research provides an in-depth computational analysis of the agglutinative structures in Urdu, focusing on the morphological complexities that pose significant challenges to natural language processing (NLP). Urdu, like other agglutinative languages, forms words by combining morphemes, making traditional rule-based models insufficient to fully capture their morphological richness. Through a systematic examination of the Emille Urdu Corpus, the study identified and annotated various agglutinative patterns, such as the frequent use of prefixes like "نا-" (nā-) and "بد-" (bad-) to convey negative meanings, and suffixes like "-گی" (gī) and "-ی" (ī) to transform adjectives and verbs into nouns.

The analysis highlighted the productivity of Urdu's agglutinative system, particularly in creating new lexical items and descriptive adjectives, which allow for nuanced expression. However, the study also revealed the limitations of traditional computational models in handling Urdu's morphological complexity. Machine learning and deep learning approaches, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) were shown to hold promise for better processing of Urdu's agglutinative structures, but their full potential remains underexplored due to a lack of extensive linguistic resources.

The findings emphasize the importance of developing robust linguistic resources and advanced computational models to enhance Urdu NLP tools. Addressing these challenges

will not only improve access to Urdu-language content in digital platforms but also contribute to the broader field of computational linguistics, particularly for underrepresented agglutinative languages. This research sets the stage for future studies that can explore the sociolinguistic factors influencing lexical choices and the integration of advanced computational techniques to further unravel the complexities of Urdu morphology. Ultimately, these efforts will lead to more effective and scalable NLP solutions for Urdu, benefiting both the linguistic community and wider society.

Key Findings

The analysis of the Emille Urdu Corpus demonstrates the prevalence of agglutination in Urdu. The frequency data reveal that Urdu commonly employs prefixes such as "نا-" (nā-) and "بد-" (bad-) to form words with negative connotations. This linguistic feature is essential in expressing negation and negative concepts, indicating the robust and dynamic nature of Urdu's agglutinative morphology. Moreover, The study highlights Urdu's productive agglutinative system, where adjectives and verbs are frequently transformed into nouns through suffixes like "-گی" (gī) and "-ی" (ī). This process facilitates the creation of new lexical items, enabling the language to adapt and evolve by generating words that are both semantically rich and grammatically functional. Examples like "وابستگی" (wābastagī) and "لاعلمی" (lāʿilmī) underscore this productivity.

Agglutination in Urdu also plays a significant role in forming descriptive adjectives, particularly those that convey undesirable qualities. Words like "بدبودار" (badbūdār) illustrate how Urdu leverages affixation to provide vivid and precise descriptions. This linguistic mechanism enriches the language's descriptive capacity, allowing speakers to articulate nuanced characteristics and attributes. Moreover, the study observes that certain agglutinative forms, particularly those involving strong moral or behavioral judgments, are either infrequently used or absent in the corpus. Words such as "بدکردار" (badkirdār) and "بدمعاشی" (badmāshī) demonstrate this trend. This finding suggests a potential shift in lexical preference or the adoption of alternative expressions to convey these concepts, reflecting broader cultural or social dynamics within the Urdu-speaking community.

Despite the richness of Urdu's agglutinative morphology, the study underscores the challenges it poses for computational analysis. Traditional rule-based models, while helpful in limited contexts, struggle with the irregularities and exceptions inherent in natural language. The advent of machine learning and deep learning approaches offers promising

solutions, but these methods require large, high-quality annotated corpora, which are currently scarce for Urdu. Moreover, the computational resources needed to train deep learning models may not always be available, further complicating the development of robust NLP tools for Urdu.

Implications for NLP and Linguistic Research

The findings of this study have significant implications for both computational linguistics and the broader field of language research. The detailed analysis of Urdu's agglutinative structures contributes to a deeper understanding of the language's morphology and provides valuable insights for developing more effective NLP tools. Specifically, the study highlights the need for the following:

1. Comprehensive Linguistic Resources: The development of high-quality annotated corpora and morphological lexicons for Urdu is crucial. Such resources would support the training of data-driven models, improving the accuracy and robustness of NLP applications for Urdu.
2. Advanced Computational Models: The study emphasizes the potential of deep learning models, particularly those leveraging Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), to handle the complexities of Urdu's agglutinative morphology. However, the implementation of these models requires not only extensive linguistic data but also substantial computational power, underscoring the need for continued investment in computational resources.
3. Cross-Linguistic Insights: Comparative studies of agglutinative languages, such as Turkish and Finnish, offer valuable lessons for Urdu NLP. By adapting and extending successful approaches from these languages, researchers can develop more sophisticated computational models tailored to Urdu's unique morphological characteristics.

Future Research Directions

This research opens several avenues for future investigation. The low or absent frequency of certain agglutinative forms in the corpus invites further exploration into the sociolinguistic factors influencing lexical choice in Urdu. Additionally, the integration of advanced computational techniques with traditional linguistic analysis could yield new

insights into the language's morphological structure, contributing to the broader understanding of agglutinative languages.

In conclusion, the study of agglutinative structures in Urdu not only enriches our understanding of the language but also presents opportunities to enhance computational tools for Urdu NLP. By addressing the challenges posed by agglutination through the development of advanced computational models and comprehensive linguistic resources, researchers can contribute to the creation of more accurate, flexible, and scalable NLP applications for Urdu. This, in turn, will facilitate greater accessibility to Urdu-language content in digital environments, benefiting both the linguistic community and the broader society.

References

Ahmed, S., & Khan, R. (2022). *Advances in Deep Learning for Morphological Analysis of Agglutinative Languages*. Journal of Computational Linguistics, 18(4), 253-270.

Ali, A., & Hussain, S. (2019). *Challenges in Natural Language Processing for Agglutinative Languages: The Case of Urdu*. Proceedings of the International Conference on Computational Linguistics, 42-49.

Baker, P., Hardie, A., & McEnery, T. (2002). *EMILLE: A 67-million word corpus of Indic languages.* In *Proceedings of the Third Language Resources and Evaluation Conference (LREC).*

Belinkov, Y., & Glass, J. (2019). *Analysis Methods in Neural Language Processing: A Survey*. Transactions of the Association for Computational Linguistics, 7, 49 72.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv preprint arXiv:1409.1259.

Comrie, B. (1981). Language Universals and Linguistic Typology. University of Chicago Press.

Durrani, N., & Hussain, S. (2010). *Urdu morphological analyzer and generator*. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC).*

Durrani, N., et al. (2016). Machine translation of morphologically rich languages: Challenges and solutions. *Computational Linguistics*, *42*(2), 345–387.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing.* Morgan & Claypool Publishers.

Goldsmith, J. (2001). *Unsupervised Learning of the Morphology of a Natural Language*. Computational Linguistics, 27(2), 153-198.

Hämäläinen, M., & Alnajjar, K. (2019). *A Computational Morphology Approach to Agglutination in Finnish*. Digital Scholarship in the Humanities, 34(2), 323- 336.

Hussain, S. (2008). *Resources for Urdu language processing*. In *Proceedings of the Sixth Workshop on Asian Language Resources.*

Hussain, S. (2008). *Resources for Urdu Language Processing*. Proceedings of the Sixth Workshop on Asian Language Resources, 89-94.

Jehangir, H., & Azher, M. (2022). Semantic frames of the Urdu conjunct verb lagnā: A corpus-based study. Corporum: Journal of Corpus Linguistics, 5(1), 1-23.

Kaplan, R. M., & Kay, M. (1994). *Regular Models of Phonological Rule Systems*. Computational Linguistics, 20(3), 331-378.

Katamba, F. (1993). *Morphology.* Macmillan.

Khan, M., et al. (2021). *Rule-Based vs. Machine Learning Approaches in Morphological Analysis of Urdu*. International Journal of Language Computing, 15(2), 189-204.

Malik, M. G. A. (2006). *Urdu morphology, orthography and lexicon extraction.* In *Proceedings of the Second International Conference on the Digital Humanities.*

Malik, Z. (2020). *The Morphological Complexity of Urdu: A Study on Agglutination*. Linguistic Inquiry, 51(3), 429-456.

Oflazer, K. (1994). *Two-Level Description of Turkish Morphology*. Literary and Linguistic Computing, 9(2), 137-148.

Oflazer, K., & Çetinoğlu, Ö. (2006). *Morphological disambiguation of Turkish text with perceptron algorithm.* In *Proceedings of the ACL-SIGMORPHON.*

Raza, H., & Mumtaz, N. (2023). *Underrepresented Languages in NLP: The Case of Urdu and its Agglutinative Morphology*. Language Resources and Evaluation, 57(1), 75-89.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). *Transfer Learning in Natural Language Processing*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 15-18.

Sajjad, H. (2007). *Statistical Methods for Urdu Morphological Analysis*. Proceedings of the Second Workshop on Computational Approaches to Arabic Script- based Languages, 55-62.

Sajjad, H., & Schmid, H. (2009). *Tagging Urdu Text with Parts of Speech: A Tagger Comparison*. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 692-700.

Schütze, H., & Manning, C. D. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.