

Challenges of Automated Processing of Arabic Morphology in Light of Artificial Intelligence Technologies: A Study of Root and Affixes Structure

Hager Mohamed Hussein Nasroon

Department of Arabic Language and Literature,

College of Languages and Humanities, Qassim University, Saudi Arabia.

Email ID:- hagardongola@gmail.com

Abstract:

This research aims at identifying and analyzing the fundamental linguistic and technical challenges that encounter Artificial Intelligence (AI) systems in processing the Arabic morphological system. It focuses on bridging the gap existed between traditional morphological rules and modern computational modeling in processing roots and affixes in Arabic words. The study adopts a descriptive-analytical approach to deconstruct or analyze the Arabic word structure, alongside a comparative approach to evaluate the efficiency of rule-based systems versus Deep Learning Models and Large Language Models (LLMs). It concludes that the non-compounding nature of Arabic, represents the most significant challenge to the automated processing, because Arabic basically is not a polysynthetic language and its morphological system primarily depends on the derivation process which uses the system of root and specific patterns in forming words, especially when the rules of substitution (replacement) (ali'lal) and vowel change (ali'lal) take place and intervene, they eventually lead to high morphological ambiguity. These two rules are called in Arabic ali'bdal and ali'lal successively; they are defined in the course of research. Furthermore, the absence of diacritics leads to an increase of morphological possibilities that statistical models are unable to decipher without an advanced semantic context. In conclusion, the research recommends adopting a "hybrid approach" that combines the rigor of traditional morphological logic with the flexibility of neural networks, while emphasizing the need to build "good" linguistic corpora that are deeply morphologically labeled to ensure accurate automated processing.

Keywords: Automated Processing, Arabic Morphology, Artificial Intelligence, Roots and Derivation Patterns, Computational Linguistics, Tokenization, Morphological Ambiguity

INTRODUCTION

The morphological system in the Arabic language is considered the basic pillar upon which the linguistic structure is based; because Arabic is a derivational language par excellence, it depends on the structure of "root and pattern" rules. Therefore, with the dominance of artificial intelligence technologies, "NLP" (Natural Language Processing) has become necessary to be used for modeling this system to enable the computer to understand and generate Arabic text with linguistic awareness.

Statement of the Problem

The research problem lies in the deep structural gap between the Arabic morphological system, based on internal derivation, and computer algorithms designed primarily for agglutinative languages. They were mainly designed in Indo-European languages' environments that rely on an agglutination (the process would be called in Arabic as blending-like process) system, while Arabic is characterized, to some extent, by internal derivation. This structural difference leads to deficiencies in analyzing words that undergo structural changes, such as vowel changes (called ali'lal) or those that lack diacritics, making the automated output susceptible to "linguistic or morphological confusion" or errors in tracing words back to their roots.

This problem generates the following fundamental research questions: How can artificial intelligence recover fused roots that have undergone elision and deletion? How can statistical models resolve morphological ambiguity resulting from the absence of diacritics without falling into the trap of statistical bias? Can large language models (LLMs) move beyond superficial analysis of affixes to understand the contextual awareness that links morphology and syntax with semantics

Objectives of the Study:

This research aims to achieve the following objectives:

1. To identify and analyze the linguistic and technical obstacles that hinder the accurate automated processing of Arabic morphology.
2. To evaluate the efficiency of modern artificial intelligence models in handling complex morphological cases.
3. To reveal the impact of the absence of diacritics and the overlap of affixes on the accuracy of automated morphological analysis.
4. To present a proposed framework to bridge the gap between traditional rules and computer modeling.

Review of Literature

The automated processing of Arabic morphology has received widespread and continuous attention in the field of computational linguistics, given the unique nature of Arabic's derivational structure and its distinctive challenges. Previous scholarly efforts addressing this topic can be categorized into three main, integrated approaches:

The First Approach has focused on the Rule- and Lexicological-Based Systems: These studies focused on the mathematical and logical modeling of traditional morphological rules. The most prominent study of them is the study by Buckwalter (2002), which laid the foundation for lexical morphological analyzers by constructing a massive database of roots and affixes. This study remains a fundamental reference for subsequent analyzers. Also it is noteworthy that the study conducted by Hamada (2016) was specialized in building precise mathematical models for processing weak and hamzated verbs, attempting to encompass all the morphological transformations resulting from vowel changes (ali'lal) and substitutions (ali'bdal). Despite the high linguistic accuracy and clear explanatory power of these models, they faced the challenge of "structural rigidity." They were also unable to handle newly generated words or common spelling errors, in addition to their high maintenance and update costs.

The second approach has studied the Statistical models and deep learning. With the advent of the data revolution, researchers turned to statistical models that learn from linguistic corpora. In this context, Al-Sallab et al. (2017) provided a comprehensive review of the potential of deep neural networks in understanding the structure of Arabic words and overcoming the obstacle of "manual segmentation". Zeman (2018) also discussed the challenges of automated segmentation in languages with complex morphology, emphasizing that statistical models excel in flexibility and generalizability. However, these studies revealed a fundamental limitation: the "lack of linguistic awareness." These models rely on abstract mathematical probabilities, which can lead to statistically correct but linguistically impossible morphological analyses (such as assigning a fictitious root to a weak word).

The researchers of the third approach have conducted research on large language models (LLMs) and generative artificial intelligence. This approach is the most recent one, which has begun to take shape with the emergence of models such as GPT. Recent studies, such as Abdelali et al. (2024) and Sengupta et al. (2023), have examined the efficiency of these models in processing Arabic, demonstrating a superior ability to understand context and resolve morphological and syntactic ambiguities. However, these studies have often focused on the overall performance of the models without delving into the subtle morphological challenges, such as the interference of complex affixes or the effect of vowel changes (ali'lal) in restoring fused roots.

THE SIGNIFICANCE OF THE CURRENT RESEARCH:

This research differs from previous studies in that it does not merely provide a technical description or general evaluation, but rather offers a comparative and critical experimental study of the latest Large Language Models (LLMs) that emerged after 2023 (such as GPT-4o, Gemini 1.5 Pro, and Jais). The qualitative addition of this research lies in proposing innovative strategic solutions represented by the "Morphological

Balancing Engine" and the "Hybrid Model." These frameworks aim to restore "linguistic dominance" to automated analysis by combining the rigor of traditional rules with the flexibility of artificial intelligence, thus ensuring the accuracy and consistency of the output with Arabic morphological patterns and standards.

Methodology

The researcher has applied five cases of study in this research. Five different linguistic expressions (words or morphemes) have been examined by four leading Large Language Models For assessing the actual capabilities of these models in recognizing and determining the meaning of these expression in texts and contexts. The language models are GPT-4 (OpenAI), Gemini 1.5 Pro (Google), Claude 3 (Anthropic), and the Jais model (Core42) as a specialized Arabic model. The models were tested by using different Arabic expressions without any diacritics. The complex morphological cases were as follows:

The first case study implied the analysis of the hollow and stressed verbs (word: قلنا /qulna/ means 'said'). Words or verbs are called hollow (jwfa'a' when their second sound or letter is a vowel 'ya ي' or 'u و', whereas verbs are called 'doubled verbs' when they have two repeated or doubled letters in its (base) root. The researcher took the root of the hollow verb /qul' قول to apply it on the selected Large Language models.

The second case study used complex prefixes intertwine with the root of the word (/fastajbna/ meaning we replied فاستجبنا). This word presents a significant challenge in "automatic segmentation," as it consists of (prefixes /fa- ف- and /iss اس/, the root /jwaba جواب meaning reply/ and the suffix /na نا meaning 'we'.

The third case study tested context-dependent morphological-grammatical ambiguity in the sentence. The sentence used in this case was "The manager informed the employees of the matter أعلم المدير الموظفين الأمر". Here, the word "informed" أعلم without diacritics can be either the past tense verb "to inform" or the comparative adjective "know more than".

The fourth case study examined doubled verbs in conjugation with a pronoun (the verb "رددت"/radadtu/ meaning response or reply "رددت"/ (with segmenting or resolving the doubled letter /dada/ versus "ردّ" (doubling or gemination by using the shadda [شدّة]). The challenge here lies in resolving the doubling when the doubled verb is conjugated with a subject pronoun, which alters the apparent form of the root.

Finally, the fifth case study applied indeclinable nouns and plural forms for the word "birds" in the context of the sentence; 'I passed by many birds مررت بعصافير كثيرة /marartu be asafeera katheerah/. The challenge here is to identify the specific morphological pattern (mafā'ā امفاعل) and understand its inflectional characteristics. The results of examining the expressions are discussed in more details in the (Findings) section.

Discussion

This chapter seeks to discuss and clarify the basic operational terms and related concepts used in this study. They imply the terms used in the intersection process between Arabic morphology, Arabic linguistics and computational techniques:

Arabic Morphology

The morphological system in Arabic is the fundamental pillar upon which linguistic structure rests. Arabic is a derivational language par excellence, relying on the "root and pattern" structure. With the advent of artificial intelligence, "natural language processing" has necessitated the modeling of this morphological system to enable computers to understand and generate Arabic text. (Ali, 1994; Al-Qahtani, 2021).

Natural Language Processing and the Role of the Morphological Analyzer:

Natural language processing is defined as a branch of artificial intelligence that aims to equip machines with the ability to understand and analyze human languages. In the context of the Arabic language, the morphological analyzer stands out as the most important tool in this process; it is the engine that breaks down the word into its basic elements (root, prefixes, and suffixes) (Hamada, 2016, Al-Juhani, 2022).

The importance of the morphological analyzer lies in the fact that it is the starting point for any other computer application, such as: machine translation, where understanding meaning depends on determining the morphological function of the word; information retrieval, which involves searching for words by tracing them back to their morphological roots; and automatic diacritization, which requires knowing the word's structure (morphology) before determining its inflectional endings.

The Uniqueness of the Arabic Morphological System (Derivation vs. Affixation)

The Arabic language possesses unique characteristics that distinguish it from Indo-European languages (such as English), and this represents the core of the computational challenge. In most Western

languages, processing is achieved by adding affixes to the beginning or end of the word (root) without fundamentally altering the structure of the root of the word. Arabic also relies on "internal derivation," where the root (such as ك ت ب in the verb /kataba/ ك ت ب) is molded into a pattern or form (such as اسْتَفْعَلَ /istafa'la/) to produce اسْتَكْتَبَ (/istakatab/). This molding between the root and the final form makes the process of "automated segmentation" far more complex than simply removing prefixes, as the root's features sometimes change due to the rules of vowel change (called in Arabic ali'lal) and substitution or replacement (called in Arabic ali'bdal). (Mansour, 2019; El-Jaward, 2024)

Levels of Automated Morphological Processing:

The morphological processing is carried out by the computer through integrated levels including: the first level is the *Morphological Analysis*: This level involves analyzing the word into its constituent elements of sounds or letters (e.g., "sayathahaboon/س + ي + هب + ون" (سيذهبون). In the second level, the computer determines the *word's characteristics/features* for example, whether the word is a verb, its tense; present tense, past tense, or noun; masculine, plural and so on. The third level is the *Morphological Generation*: It refers to the system's ability to construct and form a grammatically correct word based on a correct root, pattern, and specific features (it is the opposite of the morphological analysis [level] process). Finally, the last level is the *Morphological labeling* which means giving a "label" to each word that determines its category (noun, verb, particle etc.) based on its context. It is the most affected level by the process of "linguistic or morphological confusion" resulting from the absence of diacritics.

Findings and Results

Problems of Derivational Structure (Roots and Patterns)

The Arabic language is fundamentally based on "discontinuous morphological structure," where words are generated through the interaction of two elements: the "root," which carries the general semantic or conceptual meaning, and the "pattern/form," which gives the word its semantic, morphological and functional specificity. The dilemma for artificial intelligence lies in the fact that this interaction is not merely mechanical juxtaposition, but rather a linguistic fusion that results in structural changes that may obscure the features of the root or disguise one pattern for another.

The Dilemma of "Roots' Fusion" and the Effect of Vowel Change (ali'lal) and Substitution (ali'bdal)

Computer systems face significant difficulty in tracing words back to their roots when a root undergoes a phonetic or morphological change. For example, in words like /yaidu/ يَعِدُ which is the present verb 'promise', the verb /i'ee/ عِي which means to 'be aware', and /qee/ قِي which means to 'protect', the computer struggles to recognize that roots for these words are (/waqa/ وقى meaning 'protected', /wa'aa/ وعى meaning 'was aware', and /wa'ad/ وعد meaning 'promised'); so, the small word size and the omission of the initial or final consonant of these verbs confound the segmentation algorithms. Similarly, in a word like (انْصَلَ/itsaala/ meaning called), the computer might analyze the root as (تصل/ tsala/), while the original form is (وصل/ wasala/), so iy has the vowel sound or letter (/wa. و) into (/ta/ ت). These transformations require the morphological analyst not only to know the rules but also to be able to "predict" the underlying root behind the apparent change, which is called "deep morphology" in computing. (Zeman, 2018; Hamada, 2016).

Multiple Functional Meanings for a Single Morphological Pattern

The phenomenon of "morphological polysemy" presents a significant challenge for artificial intelligence; A single form can perform multiple semantic meanings or/and grammatical functions, causing confusion for the machine, as is the case with the form (/mufta'l/ meaning active past participle of do مُفْتَعِل), which could be an active participle (مُجْتَمِع/ mujtami'/ meaning gathered) or (مُجْتَمَع/ mujtama'/ meaning society or a place of gathering). Distinguishing between them automatically requires contextual understanding that goes beyond the limits of the single word. In addition, some morphological forms may be confused with other forms with the absence of diacritics, such as (/a'lam/ meaning flag, /i'lm/ عِلْم meaning science, /u'lim/ اُعْلِم meaning was learnt or known, and /a'llama/ عَلَّمَ meaning taught many times. عَلَّمَ); they all appear to the computer as one single form word (/i'lm/ علم). Herein lies the need for "confusion resolution" algorithms that rely on statistical probabilities or deep neural networks and can interpret the meaning from the context and select the suitable form and meaning the given text and context. (Inoue et al., 2021).

The Challenge of Non-Standard Derivation and Generated Words

Artificial intelligence often relies on standard dictionaries and rules, but language is a living, evolving entity, posing new challenges. This is the case with words that have recently entered Arabic (such as: /tilfaza/ televisionism, /magnata/ magnetism) that may not strictly conform to traditional morphological patterns,

leading the morphological analyzer to classify them as "unknown words."

Similarly, words like (/basmala/ meaning saying [In the name of God (Allah), the most gracious, the most merciful] and /hawqala/ meaning [There is no power nor strength except with God (Allah)]) are derivations that deviate from the usual "triliteral root" system and require special processing that goes beyond simply relying on consistent morphological analysis. This also includes invariable nouns that lack a root, and how the machine can differentiate them from derived nouns to ensure the accuracy of the linguistic retrieval process. (Suleiman & Al-Kabi, 2020; Al-Farsi and Al-Shammari, 2021).

The Dilemma of Affixation and additions (Morphological Challenges)

The Arabic language is characterized by the phenomenon of "affixation" alongside "derivation." Multiple morphological and syntactic elements (prefixes, suffixes, and connectives) may be attached to the Arabic root (word) and they function as conjunctions, definite articles, future tenses, and pronouns etc. These "additions or affixes" make one single word a dense linguistic unit that requires the computer to have a superior ability to "deconstruct or analyze" and "segment" it before proceeding to the semantic comprehension.

The Interplay of Prefixes and Suffixes and the Problem of Automatic Segmentation

Segmentation is the first step in automatic Natural Language Processing (NLP), but in Arabic, it faces structural obstacles such as the phenomenon of cumulative word formation. An Arabic word may begin with multiple prefixes and end with compound suffixes, such as the word (ويكتبونها/ wasayakyubunaha/ meaning "and they will write it". It is one word but it is equivalent or translated to a sentence in English) which consists of the following affixes ; prefixes include (و /wa/ meaning and), (س /sa/ means will), (ي /ya/ meaning present tense), (/kata / كتب it is the root of the word and means 'write'), and the suffixes include (و /wu/ is a connected pronoun refers to the plural pronoun 'they'), (ن /nu/ a letter which is called a protection nun and has a grammatical function more than semantic meaning) and (هـ /ha/ it is a connected pronoun refers to it).

The challenge here is to build an algorithm which is capable of distinguishing these semantic and functional additions or affixes from the letters of the root, especially when the affix (added letter [morpheme]) is similar to a root letter (such as the و /wa/ in the root /wajada/"وجد" meaning 'find' and the (و /wa/ as a conjunction). In addition to the challenge posed by spelling rules, additions sometimes lead to changes in word formation (such as the omission of (the connected [al-was] Hamza /أ a/ the glottal sound), or a change in the spelling of the medial hamza), presenting the machine with a "visual image" that differs from the original lexical form. This necessitates dual morphological and spelling processing (Obeid et al., 2020; Mansour, 2019).

Attached and Hidden/Covert pronouns and Their Impact on Computer Modeling

Pronouns in Arabic present a dual challenge to morphology and syntax. In foreign languages, pronouns are separate words, while in Arabic, there three types of pronouns; the first type includes detached pronouns, the second type implies hidden or covert pronouns and the third type involves attached pronouns. The attached Arabic pronouns are morphological suffixes added to the roots of words. Artificial intelligence must accurately determine the pronoun's position and its grammatical function to reconstruct the "predicative relationships" in the sentence. Similarly, there is the problem of the hidden/covert pronoun, as artificial intelligence processes the "digital presence," while the hidden/covert pronoun is a "linguistic absence" that carries functional value. In the sentence (the Arabic order verb أكتب /uktub/ which is equivalent to 'write' in English), the morphological analyst needs to "generate" a hidden subject, which is understood to be (you), to ensure a correct and logical analysis of the text; it is something that simple statistical models (algorithms) find difficult to clarify. (Abu Shaqra, 2021; Nivre et al., 2020).

Ambiguity Resulting from the Absence of Diacritization and the Identification of Affixes

The absence of diacritization remains the greatest obstacle to the automated processing of Arabic morphology. With absence of diacritization for the attached pronoun "ت /taa/" of the subject which may refer to the masculine 'you' or feminine 'you', the same problem occurs for "نا /naa/" which may refer to the dual subject or the plural object pronoun. For example, (ضربنا /dharbna/ we hit [past tense]) could mean we hit (/nahnu dharbna/ نحن ضربنا) or he hit us (/Huwa dharbna/ هو ضربنا). This morphological ambiguity inevitably leads to failures in syntactic and semantic analysis. Modern artificial intelligence (via neural networks) attempts to predict diacritization based on context, but the error rate remains high in complex or literary texts, highlighting the need for an "auto-diacritizer" prior to the morphological analysis of affixes. (Al-Otaibi, 2025; Antoun et al., 2020).

The automated processing of the Arabic language has undergone radical transformations, from rigid systems based on logical rules to the current revolution represented by Large Language Models (LLMs) based on deep learning. This chapter aims at evaluating these different approaches and make critique to their ability in handling the complex morphological system of Arabic through an extensive applied case study, comparing the behavior of a number of the most well-known artificial intelligence models that are currently available, and exploring the future of the Arabic morphological processing in light of these findings.

The Results of the Study

As previously mentioned, for assessing the actual capabilities of modern models, an applied test was conducted on four leading language models: GPT-4 (OpenAI), Gemini 1.5 Pro (Google), Claude 3 (Anthropic), and the Jais model (Core42) as a specialized Arabic model. The models were tested without any diacritics across three progressively more complex morphological cases as follows:

Case 1: Ambiguity in Hollow and stressed Verbs (Word: قلنا /qulna/ means 'said')

Words or verbs are called hollow (jwfa'a' when their second sound or letter is a vowel 'ya ي' or 'u و', whereas verbs are called 'doubled verbs' when they have two repeated or doubled letters in its (base) root. The researcher took the root of the hollow verb /qul' from /qala/ and qulna for the past and conjugated with plural pronoun 'we'. It was presented to the computer as a single word in isolation 'without a text) and no diacritics. The result revealed two main morphological possibilities:

Possibility (A): From the root without diacritics 'qul', the past tense verb qulna conjugated or attached with the first-person plural pronoun (na), resulting in qulna after elision.

Possibility (B): From the doubled root 'qull', the past tense verb 'qullna' conjugated with the first-person plural pronoun (na), resulting in 'qullna'.

All three models (GPT-4, Gemini 1.5 and Claude) almost exclusively favored the first possibility (root 'qul'. However, when the researcher asked the computer about other possibilities, they were able to mention the second 'qull', but considered it "much less common." This behavior confirms the phenomenon of "probabilistic bias", where models favor the most frequently occurring analysis in their training data over the grammatically correct but rare analysis. Furthermore, Jais (Core42) model presented a similar analysis focusing on the root (ق و ل qul'), indicating that the training data of large Arabic models also reflect this pattern.

Case Study 2: Complex Prefixes Intertwine with the Root of the Word (/fastajbna/ meaning we replied فاستجبنا)

This word presents a significant challenge in "automatic segmentation," as it consists of (prefixes /fa ف - / and /iss اس/, the root / jwaba جواب meaning reply/ and the suffix /na نا meaning 'we'/). Accurate analysis requires: separating the prefixes and the suffixes from the root, determining the pattern (الاستفعل istafala) and its alterations, identifying the pronoun (نا na) as the subject, and tracing the word back to its correct root (/jwaba جواب), which underwent a vowel change (called ali'lal) where the /wa و/ was transformed into an/ a / and then deleted.

The GPT-4 and Gemini 1.5 models excelled at breaking down (segmenting) the word into its constituent parts (the conjunction "fa," the "istaf'ala" pattern, and the pronoun "na") and tracing it back to its root (jwaba جواب). This demonstrates their strength in dealing with agglutinative morphology. Claude 3 also provided a sound analysis, but with a greater emphasis on explaining the morphological changes (vowel changes) that occurred in the root. Moreover, Jais model (Core 42) demonstrated a good ability at segmentation, but was less detailed in explaining the morphological rule behind the root change compared to the other models.

Case Study 3: Context-dependent morphological-grammatical ambiguity in the sentence:

"The manager informed the employees of the matter أعلم المدير الموظفين الأمر /a'lama almudeeru almwathafeena ala'mra/. Here, the word "informed أعلم" without diacritics can be either the past tense verb "to inform" or the comparative adjective "know more than" Distinguishing between them is impossible without grammatical context. So there might be two possible sentences; the first sentence is "The manager informed the employees of the matter," and the second sentence is "I know more about this matter than somebody." The three models (GPT-4, Gemini 1.5, and Claude 3) excelled in this task. In the first sentence, "informed" was identified as a quadrilateral past tense verb of the pattern "af'ala" that takes two objects. In the second sentence, it was identified as a comparative adjective. This demonstrates that modern large language models develop a strong "contextual awareness" that allows them to resolve morphological ambiguities through grammatical clues. Jais model (Core 42) also succeeded in distinguishing between the

two cases, emphasizing the importance of context in analysis, which highlights the development in models specializing in Arabic.

Case Study 4: Doubled Verbs in Conjugation with a pronoun (The Word "رددت")

The challenge here lies in resolving the doubling when the doubled verb is conjugated with a subject pronoun, which alters the apparent form of the root. In this case, the researcher examined the word or verb /radadtu/ meaning response or reply "رددت"/ (with segmenting or resolving the doubled letter /dada/ versus "ردّ" (doubling or gemination by using the shadda [شدة]).

The GPT-4o and Gemini 1.5 Pro models demonstrated a remarkable ability to revert the word to its original root (radda ر د د), providing a precise linguistic explanation for the reason behind the resolution of the doubling (conjugation with the subject pronoun). This reflects a "morphological depth" that goes beyond mere superficial letter correspondence. The Claude 3.5 and Jais-30b models successfully identified the root but were less detailed in explaining the morphological rule associated with the resolving of the doubling, limiting themselves to structural analysis without explicitly indicating the cause of the morphological change in some instances.

Case Study 5: Indeclinable Nouns and Plural Forms (the word "birds")

The challenge here is to identify the specific morphological pattern (mafāʿil) and understand its inflectional characteristics that prevent it from taking nunation. In this case, the word under test is "birds" (in the context of the sentence: I passed by many birds مررت بعصافير كثيرة/marartu be asafeera katheerah).

The Gemini 1.5 Pro and GPT-4o models excelled in accurately identifying the morphological pattern (mafāʿil) and indicated that the word is indeclinable for a single reason (its plural form). This demonstrates that the models are beginning to integrate morphological analysis with advanced grammar. In contrast, the Claude 3.5 and Jais-30b models occasionally made minor errors in identifying the precise pattern (such as faʿāʿil), highlighting their need for an additional verification layer, such as the proposed "pattern engine," to refine the standard patterns. Table (1) highlights the cognitive gap between the specialized linguistic analyst and the computational model:

Table (1): Comparison of the Performance of Large Language Models in Analyzing Morphological Ambiguity

Comparison Aspect / Case	GPT-4 (Open AI)	Gemini 1.5 Pro (Google)	Claude 3 (Anthropic)	Jais (Core42)
Case Study 1: (Qulna - قلنا)	Strong statistical bias towards (q-w-l); mentions the other possibility upon request.	Similar behavior to GPT-4; favors the most common usage.	Mentions both possibilities with a clear preference for (q-w-l) based on frequency.	Focuses on (q-w-l) as the primary analysis.
Case Study 2: (Fastajabna - فاستجبنا)	Excellent analysis and segmentation with precise identification of the root and affixes.	Very strong performance in decomposition and returning to the root with an explanation of the weak-letter transformation (l'al).	Accurate analysis with a detailed morphological explanation of the transformation rules.	Good segmentation, but less detailed in explaining morphological rules.
Case Study 3: (A'lam - أعلم)	Excellent. Distinguishes between the verb and the comparative adjective based on the grammatical context.	Excellent. Effectively uses context to resolve morphological-syntactic ambiguity.	Excellent. Shows a deep understanding of the relationship between morphology and syntax.	Very good performance; succeeds in distinguishing using context.

Challenges of Automated Processing of Arabic Morphology in Light of Artificial Intelligence Technologies: A Study of Root and Affixes Structure

Case Study 4: (Radadatu - رددت)	High ability to return the verb to the root (r-d-d) with an explicit explanation of the rule for de-gemination (فك (التضعيف) when assigned to the subject's "Taa".	Excellent performance in identifying the root and understanding the reason for the morphological change (assignment to the nominative moving pronoun).	Correct analysis of the root (r-d-d), but the explanation is less detailed regarding the reason for de-gemination, focusing more on structure than the rule.	Correct identification of the root, but the explanation for the cause of de-gemination was superficial or absent in some attempts.
Case Study 5: (Asafeer - عصافير)	Precise identification of the pattern (Mafa'eel) and noting the prevention of declension (Man' min al-Sarf) for a single cause (ultimate plural form).	Very strong performance in recognizing the morphological pattern and the grammatical property (lack of nunation and being in the genitive case with a Fatha).	Occasionally hesitated between patterns (such as Fa'aleel), with a general awareness of the prevention of declension without strict scale control.	Minor errors in identifying the exact pattern, with general knowledge that the word is among the ultimate plural forms, but it is not governed by a precise scale rule.

Existing Morphological Gaps and a Future Perspective

The extensive case study confirms that artificial intelligence models, despite their significant advancements, still suffer from structural morphological gaps when dealing with Arabic. This was evident in the case (1) of the verb (قلنا /qulna), where the models tend to generalize based on frequency, tracing words back to more common roots, thus losing accuracy in rare or ambiguous cases. They lack a rigorous logic that invokes the rules of vowel change (ali'lal) and substitution (ali'bdal) as a mandatory step in the analysis. In the case (2) of (فاستجابنا /fa'stajabna), the models became adept at handling affixes (prefixes and suffixes), a challenge similar to agglutinative languages. However, the deeper challenge (as in the case (1) of /qulna/ قلنا) remains in understanding the changes that occur within the root structure itself. The case (3) of (أعلم /a'lam/) demonstrated that the future of accurate processing lies in contextual morphology. Shifting from analyzing isolated words to analyzing them within a complete sentence (text) is what allows the machine to resolve even the most complex morphological ambiguities.

Challenges of Morphological Processing in Arabic Dialects

The transition from Classical Arabic to colloquial dialects represents a qualitative leap in the level of morphological complexity. Dialects are not merely "distortions" of Classical Arabic, but rather linguistic systems with their own morphological rules, they often deviate from traditional morphological norms. Arabic dialects (Eastern, Maghrebi, Gulf, Sudanese, etc.) differ in the way they form verbs and pronouns. For example, in the Sudanese dialect, one can find the use of the future tense prefix "ba" (ب) (as in "bamshi") instead of "s" (س) or "sawfa" (/ swfa =will/ سوف) which means I will go. However, in Maghrebi dialects, one may find a more complex morphological fusion that leads to the deletion of short vowels, completely altering the word structure. A machine trained in Modern Standard Arabic is unable to trace these words back to their roots because it cannot recognize dialectal affixes.

Furthermore, dialects are written orally, lacking standardized spelling conventions. A single word may be written in multiple forms (e.g. / a'shan/, /ashaan/, /a'la shan/ or /a'shan khater meaning for my sake "عشان", "عشان", "عشان", "عشان خاطر", which confuses the morphological analyzer and makes it unable to accurately identify prefixes and suffixes. This multiplicity of spelling forms leads to an inflated dictionary size without a real increase in efficiency. Additionally, dialects use grammatical patterns that are not found in Modern Standard Arabic or modify existing patterns. For example, the pattern "/itfa'l/ التفعل" in some dialects is used to denote responsiveness differently than in Modern Standard Arabic. Thus, artificial intelligence requires vast, morphologically labeled dialect corpora, which is currently lacking in digital

Arabic content. (Boucetti, 2025; Alayba, 2025).

Auto-Diacritization and Resolving Morphological Ambiguity

Diacritics in Arabic constitute a "parallel morphological system." Without them, a word remains a mere skeleton, open to numerous interpretations. A machine cannot distinguish between (/a'lam/ meaning flag, /i'lm/ علم meaning science, /u'lim/ علم meaning was learnt or known, and /a'llama/ علم meaning taught many times. علم) without having diacritics on the letters. Studies have shown that adding a precise auto-diacritics system increases the efficiency of a morphological analyzer by up to 40%. Modern language processing models such as Farasa and Shakkelha rely on recurrent neural networks (RNNs) to predict vowel changes based on grammatical context, but they still struggle when dealing with rare words or foreign nouns.

Finally, there is a "chicken and egg" relationship between morphology and diacritics; to correctly diacritize a word, one must know its morphological analysis, and to analyze it morphologically, one needs to diacritize it. Thus, the proposed computational solution is a "parallel language processing" model where the system generates all possible formation possibilities, then passes them through a morphological filter to exclude linguistically illegal forms. (Al-Otaibi, 2025)

Philosophical and Linguistic Dimensions of Language Modeling

Morphological analysis is not limited to the technical aspect; it extends to raise philosophical questions about the nature of language and intelligence. The human mind processes morphology through the "linguistic faculty" that instinctively connects sound, meaning, and structure. A machine, however, processes morphology as a "series of symbols" subject to mathematical probabilities. This difference explains why machines succeed with common words but fail in "morphological innovation" (generating new words based on standard patterns). When raising the question: "Can a machine analyze morphology without understanding meaning?" the pure structuralism school believes this is possible. However, technological reality has proven that "morpho-semantics" is the future. The machine must understand that adding an ("a' /alif/" in (qatil) is not merely adding a letter, but adding the meaning of "participation or the doer of the verb." This requires linking the morphological system to a broad "semantic web." (Zainhom, 2025; Sabreen Ali, 2025).

Thus, this research suggests that the future of automated Arabic morphology processing lies in a "hybrid approach" that combines the power of statistical models in understanding context with the rigor of rule-based systems in applying traditional morphological logic, as will be detailed in the following section.

Limitations of the study and Future Prospects

This research approves that the proposed framework, which will be discussed in the next sections, despite its ambition, faces some limitations and opens the door to future research prospects. This research primarily focused on the morphological challenges of Modern Standard Arabic (stemmed up from classical Arabic) and did not delve deeply into the challenges of the diverse Arabic dialects, which are characterized by different morphological features. Additionally, the efficiency of the hybrid model heavily depends on the availability of deeply labeled linguistic corpora, which remains limited in Arabic compared to other languages. Implementing the morphological balancing engine and the hybrid model required significant computing resources, especially when dealing with massive amounts of texts.

Conclusion and Recommendations

Based on the profound challenges revealed in previous sections, particularly the shortcomings of modern AI models in handling the complexities of Arabic morphology, this research proposes a comprehensive framework for AI-Enhanced Arabic Morphological Processing aimed at bridging the gap between the rigor of traditional linguistic rules and the flexibility of computing power. This framework goes beyond mere recommendations, offering a practical and applicable vision for developing a more efficient and accurate Arabic morphological analyzer.

The Morphological Balancing Engine

The "Morphological Balancing Engine" represents the core component of the proposed framework. It is an intelligent system designed to overcome the statistical limitations of large-scale linguistic models. Its primary function is to "balance" the morphological probabilities presented by statistical models against the established rules of Arabic morphology, especially in cases of complex vowel changes (ali'lal), substitutions (ali'bdal), and derivations. This engine acts as a meta-layer, receiving the outputs of deep learning-based

morphological analyzers and then subjecting them to rigorous validation based on a knowledge base of Arabic morphological rules.

The mechanism of the morphological balancing engine can be visualized through the following steps:

1. Receiving the Input Word: The engine receives the Arabic word to be analyzed (e.g., /fastajbna/ "فاستجبنا") from the Initial Morphological Analyzer (LLM).

2. Generating Initial Analyses: The LLM provides a possible morphological analysis (root, pattern, affixes) with a statistical confidence level for each possibility.

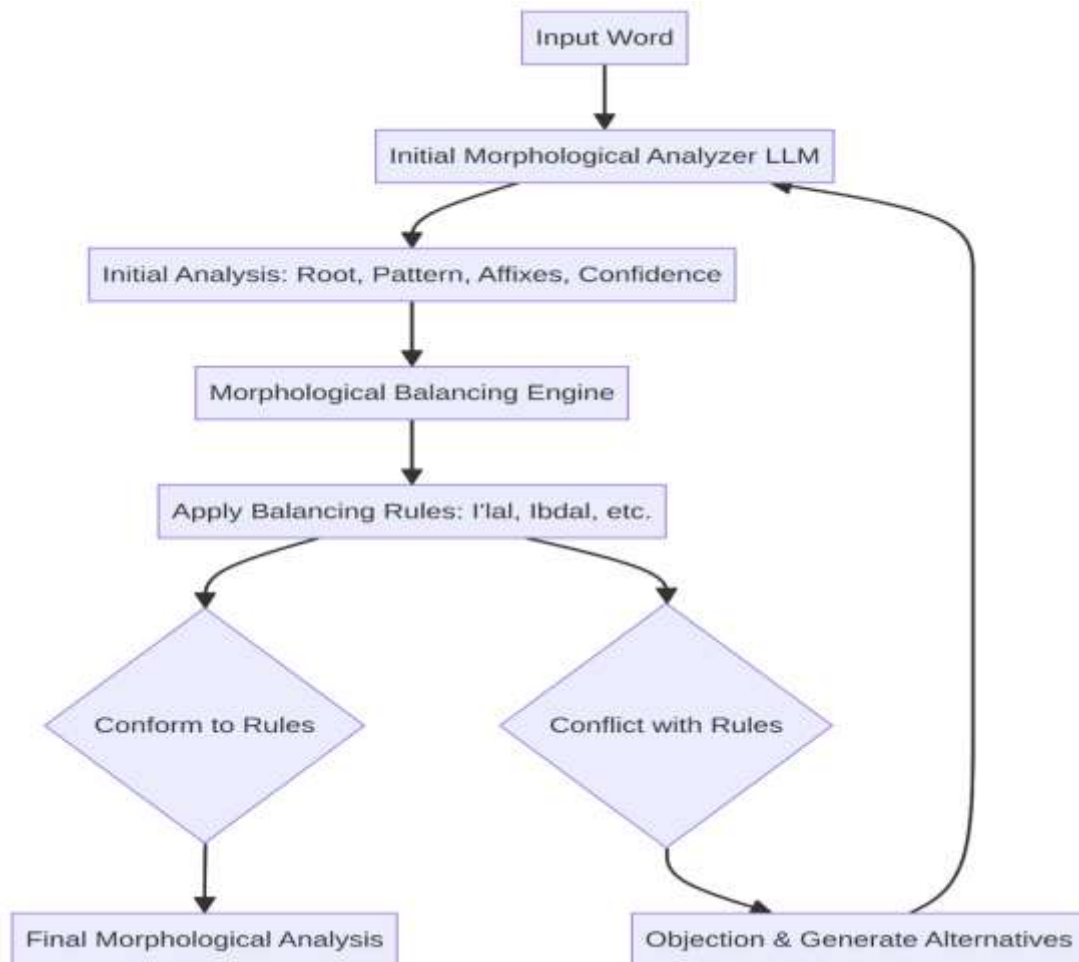
3. Applying Balance Rules: The balancing engine intervenes to apply a set of pre-programmed morphological rules (rules of vowel change (ali'lal), substitution (ali'bdal), deletion, assimilation, and irregular derivation) to the initial analyses.

4. Rejection and Correction: If the initial analyses contradict the balance rules (e.g., the deleted waw in "/fastajbna/ فاستجبنا" is not correctly restored), the engine rejects the result.

5. Generating Morphological Alternatives: The engine generates possible morphological alternatives that conform to the linguistic rules and presents them to the LLM for re-evaluation or to determine the most likely possibility.

6. Final Analysis: The final morphological analysis is presented, combining the flexibility of statistical models with the rigor of linguistic rules. Figure (1) illustrates the flowchart of the operating mechanism of the morphological pattern-balancing engine.

Figure (1): Flowchart of the operating mechanism of the morphological pattern-balancing engine



The Proposed Hybrid Model

For the purpose of achieving the maximum accuracy and efficiency, this research proposes adopting a "hybrid model" that seamlessly integrates a rule-based morphological analyzer (a smart dictionary

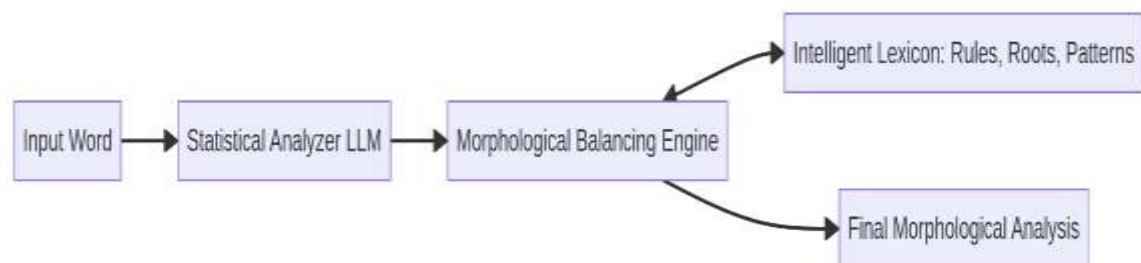
[lexicon]) and a deep learning-based statistical analyzer. This integration aims to leverage the strengths of each approach while mitigating its weaknesses. The hybrid model consists of the following main components:

The Smart dictionary or Lexicon (Rule-based Lexicon): This is a knowledge database containing Arabic roots, morphological patterns, rules of vowel changes (ali'lal), substitutions (ali'bdal), and deletions, as well as lists of affixes, prefixes, and suffixes. This component serves as a rigorous linguistic reference.

The Statistical Analyzer: This is a deep learning model (such as LLM) trained on a vast Arabic linguistic corpus, capable of extracting morphological patterns and predicting analyses based on frequency and context.

Morphological Balancing Engine: This component, detailed in the previous section, acts as a bridge and coordinator between the intelligent lexicon and the statistical analyzer. Figure (2) illustrates the architectural layout of the proposed hybrid model.

Figure (2): Architectural diagram of the proposed hybrid model.



Building Deeply Annotated Corpora

No computational morphological framework can achieve its objectives without high-quality linguistic corpora. This research recommends the necessity of building deeply annotated Arabic corpora that go beyond mere part-of-speech tagging to include tagging of root letters, even in cases of vowel changes (ali'lal) and substitutions (ali'bdal); precise morphological patterns, along with the identification of their variations; affixes, their syntactic and morphological functions, and tagging their semantic contexts in order to resolve ambiguities between morphologically similar words. These corpora will serve as invaluable resources for training statistical models and providing the necessary knowledge for the balancing engine.

Strategic Recommendations and Future Applications

The proposed framework offers a strategic vision for developing Arabic morphological processing and can open up promising future applications. Translation systems will be able to understand the deep structure of Arabic words, leading to more accurate and clearer translations. The proposed framework will also facilitate information retrieval based on roots and patterns, overcoming the challenges of derivation and multiple forms. Furthermore, it will enable researchers to analyze ancient Arabic texts with high morphological accuracy, opening new horizons in linguistic studies. This allows for the use of a precise morphological analyzer in developing interactive educational tools that help learners understand the structure of Arabic words.

This study represents a serious attempt to identify and analyze the structural and technical challenges facing artificial intelligence systems in processing Arabic morphology, focusing on the gap between the traditional grammatical rules and modern computer modeling. Through descriptive and comparative experimental analysis, the research arrived at a set of results and recommendations that outline a roadmap for the future of automated processing of Arabic morphology.

To sum up, this research concludes the followings:

Inadequacies of Purely Statistical Models: The study demonstrated that relying solely on Large Language Models (LLMs) based on statistical probabilities leads to significant morphological gaps, especially when dealing with defective roots (vowel changes [ali'lal] and substitutions [Ali'bdal]) and words with accumulated affixes, as these models tend towards statistical generalization at the expense of linguistic accuracy.

The Problem of Morphological Ambiguity and the Absence of diacritization: The absence of

diacritization remains the biggest obstacle to accurate processing; this leads to an inflation of morphological possibilities that current models are unable to resolve without a comprehensive semantic and grammatical context, as demonstrated in the distinction between verb forms and comparative adjective (أعلم/أعلم) [discussed in the previous sections].

The Challenge of Non-Attachment Nature: The research confirmed that the "root-and-stem"

structure in Arabic presents a unique challenge, fundamentally different from Indo-European languages. This renders traditional tokenization processes insufficient to accommodate the structural fusion between the root and the pattern.

Variable Performance of Modern Models: The experimental study revealed a significant variation in the performance of models. Models such as Gemini 1.5 Pro and GPT-4o demonstrated a greater ability to recover defective roots and resolve contextual ambiguities compared to other models. Meanwhile, the Arabic Jais model (Core 42) was distinguished by its relative adherence to traditional rules, although it requires further development in deciphering complex affixes.

Recommendations

Accordingly, the researcher recommends the followings:

Adopt a hybrid approach: The research recommends moving from purely statistical models to "hybrid models" that combine the flexibility of neural networks with the rigor of intelligent dictionaries based on traditional morphological rules.

Activate the morphological balancing engine: The research suggests adopting the "balancing engine" as a top-level verification layer to ensure that automated outputs conform to the rules of Arabic morphology (the morphological balance), thus reducing the phenomenon of "linguistic confusion."

Invest in deeply labeled corpora: There is an urgent need to build large Arabic linguistic corpora that are deeply labeled morphologically (including roots, patterns, and affixes) to serve as an accurate reference for training and evaluating future models.

Integrate linguistic levels: Morphological processing should not be separated from syntactic and semantic processing; many morphological ambiguities can only be resolved through understanding syntactic relationships and contextual meanings. In conclusion, the automated processing of Arabic morphology is not just a technical challenge, but a journey to explore the genius of the Arabic language and model it computationally in a way that preserves its identity and keeps pace with the era of artificial intelligence.

A collaboration between researchers and models' developers is badly needed to build up a practical application of the proposed hybrid model as an open-source tool, which will contribute to accelerating research and development in this field.

Finally, this research recommends exploring ways to integrate semantic and contextual analysis more deeply within the syllabic engine, enabling it to resolve morphological entanglements in cases where structural analysis alone is insufficient. Designing graphical user interfaces (GUIs) could help researchers and developers interact with the hybrid morphological analyzer, correct its errors, and feed the system with new rules.

REFERENCES

1. Abdelali, A., et al. (2023). Pre-training Data Quality and its Impact on Arabic LLMs. *Journal of Artificial Intelligence Research*, 76, 112–140.
2. Abdelali, A., et al. (2024). LAraBench: Benchmarking Arabic AI with large language models. *EACL Proceedings*.
3. Abu Shaqra, Jamal. (2021). *Artificial Intelligence and the Arabic Language: A Computational Linguistic Approach to Morphological Structure*. University Book House.
4. Alayba, A. M. (2025). *Arabic Natural Language Processing (NLP): A Comprehensive Review*. MDPI Computing.
5. Al-Enzi, Badr. (2024). Applications of Deep Learning in Automatic Correction of Morphological Errors. *Journal of Language Technologies*, 5(1), 22–40.
6. Al-Farsi, Maryam, & Al-Shammari, Khalid. (2021). Automating Arabic Morphological Rules: A Future Vision in Light of Digital Transformation. *King Saud University Journal of Languages and*

Translation, 33(1), 12–34.

7. Al-Juhani, Fawaz. (2022). Computational Linguistics and Its Applications in Arabic Language Processing Using Deep Learning Techniques. *Journal of Linguistic Studies*, 24(3), 88–115.
8. Al-Khaled, M. (2025). Integrating Traditional Grammar into Modern AI Architectures. *Computational Linguistics Quarterly*, 18(3).
9. Almatham, R. N., Darwish, K. M., & Al-Rasheed, R. (2025). BALSAM: A Platform for Benchmarking Arabic Large Language Models. *Proceedings of the Third Arabic NLP Conference*, 21–35.
10. Al-Otaibi, Sarah. (2025). The Impact of Automatic Diacritization on Improving the Accuracy of Smart Morphological Analyzers. *Arab Journal of Computational Linguistics*, 12(2), 101–128.
11. Al-Qahtani, Saleh. (2021). Automating the Arabic Morphological System in Light of the Fourth Industrial Revolution. *Arab Center for Research and Policy Studies*.
12. Al-Sallab, A., Rashad, M., & Raafat, H. (2017). Deep Learning Methods for Arabic Morphological Analysis. *Proceedings of the ACL Conference on Natural Language Processing*.
13. Al-Shammari, Fahd. (2020). Building Annotated Linguistic Corpora: An Applied Study on Arabic Morphology. *Dar Al-Manhaj*.
14. Al-Thubaity, A. (2023). Evaluating Modern LLMs on Arabic Morphological Disambiguation. *International Journal of Computational Linguistics (IJCL)*, 14(1), 50–72.
15. Al-Zahiri, Abdulrahman. (2023). Challenges of Large Language Models in Analyzing Arabic Morphological Structure. *International Journal of Linguistic Research*, 15(2), 45–67.
16. Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based Model for Arabic Language Understanding. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.
17. Boucetti, R. (2025). Arabic Language Characteristics that Make its Automatic Processing Challenging. *ResearchGate*.
18. Bounhas, M., & Elayeb, B. (2019). A New Hybrid Approach for Arabic Morphological Analysis. *International Journal of Speech Technology*, 22(3), 633–655.
19. El-Jaward, M. (2024). Advanced Tokenization for Non-Concatenative Languages. *Journal of Arabic NLP Systems*, 9(1), 15–38.
20. Hamada, Salwa. (2016). Arabic Computational Lexicography: A Study on Building Morphological Analyzers. *Dar Al-Uloom for Publishing and Distribution*.
21. Inoue, G., Khalifa, S., & Habash, N. (2021). The Interplay of Morphology and Syntax in Disambiguating Arabic. *ACL-IJCNLP 2021*.
22. Khalifa, S., Habash, N., & Rambow, O. (2020). Morphological Analysis and Disambiguation for Dialectal and Standard Arabic. *Journal of Computational Linguistics*, 46(2), 231–255.
23. Mansour, Ahmed. (2019). Mathematical Modeling of Root and Pattern in Contemporary Computational Linguistics. *Dar Al-Fikr Al-Arabi*.
24. Nagoudi, E. M. B., et al. (2022). Turjuman: A Context-Aware Neural Machine Translation System for Arabic. *Proceedings of EMNLP 2022*.
25. Nassar, Mohammed. (2021). Automatic Processing of the Arabic Language: Challenges and Prospects. *Dar Al-Ma'rifa*.
26. Nassar, Mohammed. (2021). Toward an Expert System for Arabic Morphological Processing: A Hybrid Technical Vision. *Journal of Applied Linguistics*, 8(1), 112–135.
27. Nivre, J., de Marneffe, M. C., & Ginter, F. (2020). Universal Dependencies: A Cross-Linguistic Framework for Morphosyntactic Annotation. *Computational Linguistics*, 46(4), 845–862.
28. Obeid, O., Nasser, M., & Khalifa, S. (2020). CAMEL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. *Proceedings of the LREC 2020*.
29. Sabrin Ali Hussein. (2025). Processing the Complexities of the Arabic Language Using Artificial Intelligence Techniques. *Journal of Scientific Research*.
30. Sengupta, N., Sahu, S. K., & Jia, B. (2023). Jais model (Core 42) and Jais model (Core 42) -chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
31. Suleiman, D., & Al-Kabi, M. N. (2020). Deep Learning for Arabic Natural Language Processing: A Survey. *IEEE Access*, 8, 122–145.
32. Zeinham, Haitham. (2025). Applied Linguistic Perspectives in Artificial Intelligence and the Arabic Language. *Applied Linguistics Journal*.
33. Zeman, D. (2018). Reusable Morphological Analyzers for Arabic: A Survey of Recent Tools. *Journal of Language Resources and Evaluation*, 52(4), 915–940.