MASAQ Parser: A Fine-grained MorphoSyntactic Analysis for the Quran

Majdi Sawalha [a, b], Faisal Al-Shargi [c], Sane Yagi [d, e], Abdallah T. AlShdaifat [f], Bassam Hammo [b, g]

[a] King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan.
[b] College of Engineering, Al-Ain University, Abu Dhabi, UAE.
[c] Amazon Robotics, New York, USA.
[d] Department of Foreign Languages, University of Sharjah, Sharjah, UAE.
[e] English Department, The University of Jordan, Amman, Jordan.
[f] College of Arts and Languages, Mohamed bin Zayed University for Humanities, Abu Dhabi, UAE.
[g] School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan.

sawalha.majdi@ju.edu.jo, falsharg@amazon.com, saneyagi@yahoo.com, abdallah.alshdaifat@mbzuh.ac.ae, b.hammo@ju.edu.jo

Abstract

This paper introduces the Morphological and Syntactical analysis for the Quran text. In this research we have constructed the MASAQ dataset, a comprehensive resource designed to address the scarcity of annotated Quranic Arabic corpora and facilitate the development of advanced Natural Language Processing (NLP) models. The Quran, being a cornerstone of classical Arabic, presents unique challenges for NLP due to its sacred nature and complex linguistic features. MASAQ provides a detailed syntactic and morphological annotation of the entire Quranic text that includes more than 131K morphological entries and 123K instances of syntactic functions, covering a wide range of grammatical roles and relationships. MASAQ's unique features include a comprehensive tagset of 72 syntactic roles, detailed morphological analysis, and context-specific annotations. This dataset is particularly valuable for tasks such as dependency parsing, grammar checking, machine translation, and text summarization. The potential applications of MASAQ are vast, ranging from pedagogical uses in teaching Arabic grammar to developing sophisticated NLP tools. By providing a high-quality, syntactically annotated dataset, MASAQ aims to advance the field of Arabic NLP, enabling more accurate and more efficient language processing tools. The dataset is made available under the Creative Commons Attribution 3.0 License, ensuring compliance with ethical guidelines and respecting the integrity of the Quranic text.

1 Introduction

This paper introduces MASAQ, a new dataset for Arabic Natural Language Processing (NLP). MASAQ focuses on the Quran, which presents unique challenges for NLP due to its sacred nature and complex linguistic features. While the Quran is a cornerstone of Classical Arabic and offers rich opportunities for NLP research, no existing resource provides the structured data necessary for computational analysis.

The MASAQ dataset addresses this gap by providing a detailed syntactic and morphological annotation of the entire Quranic text, using over 131,000 morphological entries and 123,000 instances of syntactic functions. These annotations cover a wide range of grammatical roles and relationships. MASAQ is particularly valuable for tasks such as: dependency parsing, grammar checking, machine translation, and text summarization.

Previous work on Arabic dependency parsing has resulted in several treebanks. Some of them, e.g., the Arabic Treebank by the University of Pennsylvania, the Prague Arabic Dependency Treebank (PADT), and the Columbia Arabic Treebank, have focused on Modern Standard Arabic (MSA) and news texts. However, these treebanks often use different annotation schemes, and access to some is restricted.

The Quranic Arabic Corpus, developed at the University of Leeds, is one example of a project that uses Quranic Arabic. However, no existing resource faithfully reflects the detail and nuances of traditional Arabic i'rab in a computationally friendly format. MASAQ seeks to address this need, making the sophisticated syntactic analysis found in the Arabic heritage accessible for NLP research and applications.

MASAQ's creation involved the following:

Proceedings of the New Horizons in Computational Linguistics for Religious Texts (Coling-Rel), pages 67–75
January 19, 2025.

- Source Text: The raw data for MASAQ came from the Tanzil Quran text, known for its high accuracy and authenticity.
- Annotation: A team of Arabic native speakers with expertise in traditional Arabic syntactic analysis annotated the text. They used a comprehensive tagset of 71 syntactic roles to capture the grammatical information found in classical grammar treatises.
- Annotation Tool: A custom annotation application was developed to streamline the process, presenting verses for context and breaking down words into their constituent morphemes.

MASAQ will be a valuable resource for advancing both academic and practical applications in Arabic NLP, including:

- Pedagogical Uses: Simplifying the teaching of Arabic grammar.
- NLP Tool Development: Enhancing tools like part-of-speech taggers and parsers.
- Linguistic Research: Providing valuable syntactic analysis for research purposes.
- Cross-linguistic Research: Supporting efforts like Universal Dependencies and facilitating multilingual NLP tool development.

We released MASAQ under the Creative Commons Attribution 3.0 License, and ensured compliance with ethical guidelines and respecting the integrity of the Quranic text.

2 Background Review

The Quran, as the central religious text of Islam, holds immense significance in the Arabic language. It has long been a cornerstone for linguistic studies and, more recently, has emerged as a valuable resource for NLP tasks. Its rich linguistic structure and extensive historical analysis make it an ideal candidate for supervised learning in NLP.

While traditional Quranic parsing resources offer a wealth of linguistic insights, their format and style present significant challenges for computational analysis. These resources often lack a formal, structured representation, relying instead on prose descriptions that assume a deep understanding of Arabic grammar. This reliance on implicit knowledge makes it difficult to automate the extraction of syntactic information. Moreover, inconsistencies in terminology and varying levels of detail further hinder the development of standardized annotation schemes. To address these limitations and enable the application of NLP techniques to Quranic text, there is a pressing need for a structured and consistent dataset like MASAQ.

2.1 Previous Arabic Dependency Datasets

There are several valuable Arabic dependency treebanks that preceded our work and demonstrated considerable efforts in Arabic syntactic analysis. However, these treebanks have left a gap. They exhibit different annotation schemes, restricted access, and a lack of focus on traditional Arabic i'rab. MASAQ addresses these limitations, providing a unique and open resource for the computational analysis of Quranic Arabic.

Many worthy initiatives in Arabic dependency parsing, such as the Arabic Treebank by the University of Pennsylvania (Maamouri, et al., 2006), the Prague Arabic Dependency Treebank (Hajic, et al., 2004; Smrz, Bielicky, and Hajic, 2008), and the Columbia Arabic Treebank (Habash & Roth, 2009), focused on Modern Standard Arabic (MSA) and news texts. These treebanks often used different annotation schemes, making it difficult to compare and integrate their data. Access to some of these treebanks, such as the University of Pennsylvania's, is restricted, requiring a subscription or institutional affiliation.

Other projects focused on Classical Arabic, specifically the language of the Quran. The Quranic Arabic Corpus (Dukes, 2011), developed at the University of Leeds, pioneered the incorporation of i'rab into its framework. The Arabic Poetry Treebank (Al-Ghamdi, Al-Khalifa, & Al-Salman, 2021), focusing on poetic Classical Arabic, faced challenges with errors in tokenization, POS tagging, and dependency relation labeling due to using a parser designed for MSA.

The Camel Lab Treebank (Taji & Habash, 2020) stands out for its coverage of a diverse range of text types, including classical literature, modern news articles, and user-generated content, and its open-access policy. However, it uses a simplified set of syntactic categories and relations based on traditional Arabic grammar but not identical to it.

The I'rab Dependency Treebank (Halabi, Fayyoumi, & Awajan, 2021), while adopting a traditional Arabic grammatical theory approach, is limited in size, containing only 601 sentences

sampled from the Prague Arabic Dependency Treebank (Hajic, et al., 2004; Smrz, Bielicky, and Hajic, 2008).

The limitations of existing Arabic dependency datasets underscore the need for a resource like MASAQ, which focuses specifically on traditional Arabic i'rab and provides a comprehensive and consistent annotation scheme in an open-access format.

The Quran, as a cornerstone of Classical Arabic, offers a rich resource for NLP research due to its complex syntax and semantics. The intricate nature of Quranic Arabic necessitates utmost accuracy in NLP tasks related to it. MASAQ addresses this challenge by providing a detailed and carefully annotated dataset specifically designed for dependency parsing of the Quran (Sawalha, et al., 2024).

2.2 Significance of Quranic Arabic for NLP

The Quran's language, Classical Arabic (CA), is vital for datasets aiming to capture the nuances of this Arabic variety. Its complex syntax, semantics, morphology, and phonology provide an excellent resource for testing and refining NLP algorithms. This is especially true given the Quran's meticulous preservation and standardization, ensuring a consistent and reliable corpus for machine learning and linguistic research.

Beyond the text itself, thousands of authoritative works explore facets of the Quran, including interpretation, translation, morphology, syntax, jurisprudence, and the subject matter of over 30 branches of Islamic sciences. These scholarly resources offer detailed analyses that can inform the development of robust NLP datasets.

Despite its value, the Quran poses unique challenges for Arabic NLP, particularly in morphological analysis and parsing. Due to its sacred nature, Muslims expect utmost accuracy in NLP tasks related to the Quran, often approaching 100%. This necessitates careful attention to detail, including sourcing the raw text from authenticated sources, developing a comprehensive and well-justified annotation scheme, and engaging annotators with high expertise in Arabic morphology, syntax, and Quranic interpretation.

The Quran's value is demonstrated by its use in a range of NLP projects. The Quranic Arabic Corpus, for example, utilizes the text for syntactic and morphological analysis. Other projects have leveraged the Quran to improve machine translation systems, text classification, information retrieval, sentiment analysis, and speech recognition and synthesis systems for various Arabic dialects. Additionally, the Quran has been used for tasks like phrase-break prediction and IPA phonemic transcription.

In conclusion, while the Quran presents challenges for NLP research, its linguistic richness, standardization, and extensive body of scholarly work make it an invaluable resource for developing and refining sophisticated NLP tools and resources.

2.3 Limitations of Traditional Quranic Parsing

Traditional Quranic parsing resources, despite their richness and depth, are not suitable for NLP due to limitations primarily that stem from format and style, which make it difficult to adapt such resources for computational analysis.

Lack of Formal Representation: Traditional resources lack a formal, structured representation suitable for computational modeling. Instead, syntactic analyses are presented in prose, assuming a high level of reader expertise in Arabic grammar. This reliance on assumed knowledge leads to abbreviated descriptions that hinder computational analysis. For example, instead of a detailed breakdown, a phrase might simply be labeled as a "prepositional phrase," which is insufficient for training and testing language models.

Inconsistencies and Ambiguity: Traditional resources often exhibit inconsistencies in terminology when describing grammatical functions. For instance, "na't" and "sifa" are both used for "adjective," while "maf'ul bihi" and "maf'ul" both refer to "object". These varying terms introduce ambiguity and make it difficult to develop a standardized annotation scheme for NLP purposes. Additionally, the level of detail in parsing can vary significantly across resources, further complicating computational analysis.

MASAQ as a Solution: The limitations of traditional resources underscore the need for a systematic and structured approach specifically designed for Quranic Arabic. The MASAQ dataset addresses these challenges by providing a consistent representation of Quranic syntax suitable for NLP research. Its standardized syntactic tagset and detailed annotations enable computational analysis and facilitate the development of automatic parsers for Arabic.

3 Syntactic Parsing

Parsing is the process of analyzing textual content to determine its grammatical structure and the relationships between its components. In syntactic analysis, parsing involves examining the structure of a sentence to identify parts of speech (like nouns, verbs, adjectives) and their roles within the sentence, e.g., subject, predicate, object, etc. Parsing is crucial for such NLP tasks as machine translation, speech recognition, and text-to-speech processing.

I'rab is one type of syntactic parsing. It involves grammatically dissecting verses to understand their structure, nuances, and linguistic implications. It encompasses three key aspects: (1) Identifying the grammatical function of a word within a sentence, e.g., subject (مبتدأ *mubtada'*), predicate (خبر *khabar*), circumstantial qualifier (حال *ḥāl*), etc. (2) Specifying the case of the word (e.g., nominative, accusative, genitive inflection) based on its function. (3) Noting any additional context or information associated with the word's phrasal affiliation. Religious scholars use *i'rab* to understand the exact meaning of verses and to resolve grammatical ambiguities in the Quran. Computer scientists may use it for information retrieval and machine translation. Here, Figure 1, is an example of the *i'rab* of the Quran, verse 1:

		Word		*I'rab*	Phrasal I' rab (function of embedded phrases)
Sentence	1	الْحَمْدُ	*al-ḥamdu*	*mubtada'*	
	2	لِلَّهِ	*li-Allāhi*	Prep. + Noun	Predicate (khabar)
	3	رَبِّ	*rabbi*	Adj.	
	4	الْعَالَمِينَ	*al-'ālamīna*	Poss. Comp.	
	5	الرَّحْمَٰنِ	*al-raḥmāni*	Adj.	
	6	الرَّحِيمِ	*al-raḥīmi*	Adj.	
	7	مَالِكِ	*māliki*	Adj.	
	8	يَوْمِ	*yawmi*	Poss. Comp.	
	9	الدِّينِ	*al-dīni*	Poss. Comp.	

Figure 1: *I'rab* for One Quranic Verse

3.1 Constituency vs. Dependency Parsing

Syntactic analysis may be in the form of constituency parsing or dependency parsing. These approaches to syntactic analysis have distinct methodologies and applications. Constituency parsing utilizes context-free grammars to create hierarchical trees that divide sentences into phrasal constituents, such as noun and verb phrases, whereas dependency parsing represents syntax through directed edges in a graph, capturing dependencies between words with a single root, typically the verb. Parsing natural language presents challenges due to ambiguity, often requiring supervised machine learning models trained on annotated data to resolve multiple valid interpretations. The choice between constituency and dependency parsing depends on the application, with dependency parsing being advantageous for information extraction and free word order languages, and constituency parsing preferable for extracting sub-phrases.

I'rab is a type of Arabic dependency rather than constituency parsing since it is focused on identifying the grammatical relationships between words in a sentence, such as subject-verb relationships, while constituency parsing is focused on identifying the phrase and sub-phrase constituents.

4 MASAQ Dataset: Composition and Annotation

MASAQ can be characterized by the composition and structure of its raw data, as well as by its annotations (Sawalha, et al., 2024).

4.1 Composition and Structure of MASAQ

The MASAQ dataset is built upon the foundation of a verified Quranic text that has been enriched through a layered annotation process.

Raw Data Source: The foundation of MASAQ is the Tanzil Quran text, selected for its high accuracy and verification standards. The Tanzil project uses a three-step verification process to ensure accuracy, including: automatic text extraction, rule-based verification, and manual verification against the Medina Mushaf.

This rigorous process makes the Tanzil version widely acclaimed for its lack of errors. The text is available in both the Uthmani and imla'i scripts, with the Uthmani script being the most authoritative and the imla'i offering a modernized representation.

The Tanzil version adheres to the 1924 Cairo edition of the Quran, endorsed by Al-Azhar University. This edition standardized the *Ḥafṣ 'an 'Āṣim* reading and established the commonly accepted verse numbering and chapter ordering. MASAQ utilizes this text in accordance with the Creative Commons Attribution 3.0 License.

Annotation Structure: MASAQ enhances this raw text by adding a layer of comprehensive syntactic analysis. Table 1 provides a profile of the dataset, showing the number of dataset entries to be 130K morphemes, instances of syntactic functions (*i'rab*) 123K, and Quran words (before morphemic analysis) 77,408 words.

The annotated data in MASAQ follows a specific structure. Each row represents a word or word part (morpheme) from a specific verse and chapter (*sūra*) of the Quran. The data is organized sequentially, starting with the first word of the Quran and proceeding to the last. Each word or segment is aligned with morphological, syntactic, and semantic tags to describe its features and function within the sentence.

MASAQ's Raw and Annotated Data: To understand the structure of MASAQ, it is helpful to consider the distinction between the raw data and the added annotations:

- Raw Data: This refers to the Quranic text itself, taken from the Tanzil project. It is the base layer upon which the annotations are built.
- Annotated Data: This layer adds value to the raw text by providing detailed linguistic information about each word or word part. The annotations include morphological details (e.g., root, prefixes, suffixes and their types), syntactic roles (e.g., subject, object, preposition), and semantic information.

Table 1 gives further details about the composition of the MASAQ dataset, including information about:

- The breakdown of words based on the number of morphemes they contain.
- The number of definite article tokens, reflecting a key feature of Arabic grammar.

Table 1: Profile Summary of MASAQ

Type	Count
Dataset entries (Morphemes)	131,930
Instances of syntactic functions (*i'rab*)	123,565
Quran words (before morphemic analysis)	77,408
Words composed of one morpheme	34,909
Words composed of two morphemes	31,997
Words composed of three morphemes	9,175
Words composed of four morphemes	1,154
Words composed of five morphemes	152
Words composed of six morphemes	21
Definite article tokens	8,365

These details offer a glimpse into the complexity of the Quranic language and the challenges of analyzing it computationally.

4.2 The Full Tagset

The full tagset encompasses a comprehensive list of 71 syntactic roles assigned to words, word parts, and phrases within the dataset. These tags are essential for understanding the relationship between syntactic units in a Quranic sentence. Each tag encapsulates the grammatical information that classical grammar treatises prescribed for the Quranic text; wherever there is difference of opinion on the grammatical function of an item or unit, the simplistic interpretation is favored. What makes this tagset especially valuable is that it fixes how syntactic roles are referred to.

These tags are meant to account for all syntactic functions of significance that are used in classical literature on the i'rab of the Quran. Here are a few notes to comment on the tags that might not be highly frequent in Arabic Computational Linguistics or whose English translation might not be familiar.

EXPLET: Expletive, when a word or phrase is inserted into a sentence without being necessary for the expression of the basic meaning of the sentence. It is a placeholder. Expletives can serve rhetorical, emphatic, or stylistic functions. The term originates from the Latin word expletivus, meaning "serving to fill out or take up space". For example, اسم لا محل له من الإعراب *'ism lā maḥalla lahu min al-ʾiʿrāb.*

DEM_GEN: Demonstrative in the Genitive Case. The demonstratives in Arabic are called *'ism ishāra.*

SUBJ_COP_PART: Subject of a Copula Particle. The first term in a stative sentence may change case as a result of the effect induced by كَانَ *kāna* and إِنّ *inna* and their sisters, which are referred to here as copula particles. There are a few tags that involve the copula in its various forms, as a particle and as a verb: perfect, imperfect, and imperative (كَانَ *kāna*, يَكونُ *yakūnu*, كُنْ *kun*).

An interjection is a word or phrase that is grammatically independent from the words around it and it mainly expresses feeling rather than meaning. It's a part of speech used to express a spontaneous emotion or reaction, such as surprise, excitement, or disgust. Interjections are common in everyday speech and informal writing but they are also present in the most cultivated forms of written language. As a single word or phrase, they may be

used on their own or as part of a sentence. They often take the form of an imperative verb (INTERJ_CV, اسم فعل أمر, as in هَلُمَّ *halumma*), an imperfect verb (INTERJ_IV, اسم فعل مضارع, as in وَيْ *wayy*; أُفٍ *'uffin*), or a perfect verb (INTERJ_PV, اسم فعل ماض, as in هيهاتَ *hayhāta*).

A Comitative Object refers to the use of a grammatical construction that expresses the idea of "accompaniment" or "togetherness" with another entity. In Arabic, the comitative object (مفعول معه) is a noun that indicates an entity accompanying the action of the verb, typically introduced by the conjunction "و" and placed in the accusative case.

Driven by curiosity, let us briefly examine the syntactic nature of the Quran. The most frequent 50 syntactic tags are presented in Figure 4.

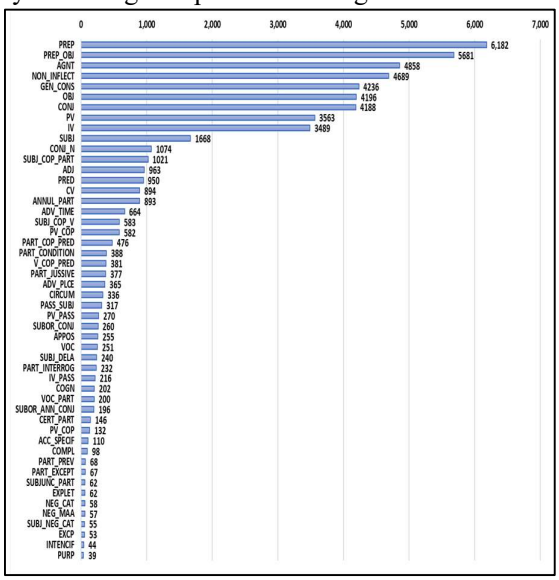

Figure 4: The most frequent 50 tags in the MASAQ Corpus

Figure 4 presents the 50 most frequent syntactic functions in MASAQ. The tag PREP (preposition) appears as the most frequent, followed by PREP_OBJ (prepositional object). One might expect their frequencies to match, since they typically form a governor-governed pair. However, the discrepancy can be explained by syntactic embedding, where a clause can serve as the governed element instead of a single word or phrase; therefore, its surface dependency parsing will not allocate a PREP_OBJ tag to the embedded clause.

Notably frequent is AGENT, which is assigned case by the verb that governs it. The frequency of verbs governing AGENT appears significantly lower because these verbs are split between the tags PV (Past Verb) and IV (Imperfect Verb).

There is clear prevalence of NON_INFLECTing particles, the GEN_CONS (genitive construct), and OBJ (object) which represents the theme and patient roles.

Among the least frequent grammatical functions are PURP (Purpose Adjunct), INTENSIF (Intensifier), EXCP (Exceptus), NEG_CAT (Categorical Negation), along with its related dependent elements.

Another view of the syntactic nature of the Quran dataset is to study some of its inflectional morphology. Let us consider Table 2.

Table 2: Frequency of Case/Mode Markers

Index	Description		Count
1	*Sukūn*	السكون	17677
2	*Fatḥa*	الفتحة	14492
3	*kasra*	الكسرة	6392
4	*ḍamma*	الضمة	4978
5	*nūn*	ثبوت النون	1008
6	elided *nūn*	حذف النون	788
7	implied *fatḥa*	فتحة مقدرة	707
8	implied *ḍamma*	ضمة مقدرة	627
9	*yā'*	الياء	539
10	*wāw*	الواو	236
11	implied *kasra*	الكسرة المقدرة	153
12	elided vowel	حذف حرف العلة	149
13	diptote *fatḥa*	الفتحة ممنوع من الصرف	114
14	*'alif*	الألف	43

Case and mode are expressed in Arabic either through diacritical marks or through such grammatical features as the vowels and "نون" (*nūn*). The high frequency of "سكون" (*sukūn*) is the result of it being the jussive mode and imperative mode marker, as well as the marker of some particles and indeclinables, and a cliticization necessity. Similarly, "فتحة" (*fatḥa*) is the marker of the subjunctive mode for verbs, the accusative case for nouns, as well as the ending diacritic of several particles and indeclinables. Both are crucial for proper pronunciation of cliticized verbs.

Notice also that the frequencies of the vowel and "نون" (*nūn*) markers are relatively low.

One possible implication for the facts presented in this section is pedagogical. School curriculum developers should avoid deterring pupils from learning Arabic by the complexity of the grammar metalanguage. Clearly, the focus of any grammar instruction, if at all, should be on the teaching of verbal and stative sentences and the basic concepts of the part of speech: noun, verb, particle, and the syntactic roles of agent, object, prepositional phrase, genitive construct, adjective, and adverb.

4.3 Annotation Process

The annotation process in MASAQ involved a multi-faceted approach, considering the complexities of Quranic Arabic and its *i'rab*. Due to Arabic's rich morphology, a single word can be composed of multiple morphemes (meaningful units), each potentially serving a distinct grammatical role. Therefore, the annotation in MASAQ had to address both the word and sub-word levels, capturing both morphological details and syntactic relationships within each sentence.

A team of Arabic native speakers proficient in *i'rab* was assembled. Their expertise comes from holding postgraduate degrees in Arabic, with a focus on morphology or syntax. Annotators applied the centuries-old methodology of *i'rab* using our comprehensive tagset. This approach facilitated efficient annotation, as the team was already familiar with the concepts. It also enhanced accuracy and consistency, minimizing the need for extensive training.

To further streamline the process and reduce potential discrepancies, a dedicated annotation application was developed. This tool presented verses (sentences) for context, allowing annotators to understand the grammatical relationships within the sentence. It also displayed words sequentially and it automatically broke them down into their constituent morphemes to simplify the annotators' task. While the tool presented morpheme-level breakdowns, the annotators assigned grammatical functions to only stems and clitics, i.e., at both word and sub-word levels, reflecting the nuances of traditional i'rab. Table 3 illustrates how one verse was morphologically and syntactically analyzed.

The prevalence of embedded phrases in Quranic Arabic posed a challenge for annotation. While not fully accounted for in MASAQ (due to cost constraints), the significance of embedding was acknowledged for future work. In conclusion, the annotation of MASAQ involved a rigorous process conducted by experts. They used a systematic approach, aided by a custom-built application, to provide a detailed and consistent analysis of the Quran's syntactic structure, laying the groundwork for further NLP research and applications.

Table 3: A Morphologically and Syntactically Analyzed Phrase from one Quranic verse 2:76

Word	Morpheme			Grammatical Function	English Gloss
يحاجونكم	ل	Proc.	SUBJUNC_PART	SUBJUNC_PART	*so*
	ي	Pref.	IMPERF_PREF	-	
	حاج	Stm	IV	IV	*challenge*
	و	Suff.	SUBJ_PRON	AGNT	*they*
	كم	Enc.	OBJ_PRON	OBJ	*you*
به	ب	Stm	PREP	PREP	*with*
	ه	Enc.	PRON_3MS	PREP_OBJ	*it*
عند	عند	Stm	ADV	ADV_PLCE	*before*
ربكم	رب	Stm	NOUN_CONCRETE	GEN_CONS	*Lord*
	كم	Enc.	POSS_PRON	GEN_CONS	*your*

5 Development of the MASAQ Parser

To develop the parser, a comprehensive system was built using NLP techniques and machine learning algorithms. The parser processes Arabic text, by extracting linguistic features and parsing each word into its corresponding grammatical tags. The dataset used for training comprises annotated Quranic words, with each word linked to its verse, chapter, and grammatical tags. The tags include grammatical and syntactic roles such as nouns, verbs, adjectives, and constructs like possessive phrases, nominal sentences, etc.

The parser leverages word-level features, including single letter prefixes, suffixes, positional attributes, and adjacent words, to predict grammatical tags accurately. A pipeline integrating a feature vectorizer and a Linear Support Vector Classifier is employed for classification. Additional preprocessing functions handle linguistic nuances, such as identifying numeric content and checking word shapes.

The model was trained and tested on a dataset consisting of Quranic text. Features were engineered to account for contextual relationships between words, such as preceding and following words, enabling the parser to capture complex syntactic patterns. The trained model achieved a high degree of accuracy and was saved for efficient deployment. This system provides a robust

foundation for parsing Quranic text and could be adapted to other Arabic text corpora for similar linguistic analysis.

Example: A parsed phrase from Quranic verse 20:114:

Sentence: وَقُل رَّبِّ زِدْنِي عِلْمًا

Analysis:

- وَ: حرف عطف
- قُل: فعل_أمر
- رَّبِّ : منادى
- زِدْنِي: فعل_أمر + مفعول_به
- عِلْمًا: مفعول_به

This example illustrates the parser's ability to accurately identify the grammatical roles of words within a Quranic verse.

5.1 Evaluation

The evaluation of three classification algorithms—LinearSVC, Logistic Regression, and Random Forest—was conducted to determine the most accurate model. The results of the experiments revealed distinct performance levels among the tested algorithms. LinearSVC achieved an accuracy of 98.23%, showcasing its capability to handle the classification task effectively. Logistic Regression, while robust, performed comparatively lower with an accuracy of 88.0%. On the other hand, Random Forest outperformed the other models with an accuracy of 99.0%, making it the best-performing algorithm in this evaluation. The results highlight Random Forest as the most suitable model for achieving high accuracy in this task. Figure 5 shows the achieved accuracy of the three models used for syntactically parsing Arabic text.

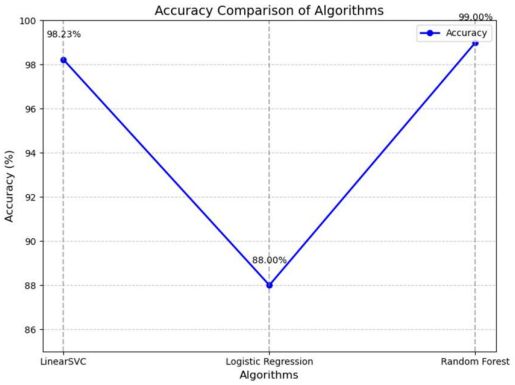

Figure 5: Achieved Accuracy of the Three Models

Example 1: A parsed phrase from Quranic verse 20:114

Sentence	وقل رب زدني علما
Analysis	و: حرف عطف قل: فعل_أمر رب: منادى زدني: فعل_أمر + مفعول_به علما: مفعول_به

6 Ethical Considerations in MASAQ

The development of MASAQ, while guided by rigorous methodology and expert annotation, encountered inherent limitations and required careful navigation of ethical considerations. One key limitation stems from the inherent ambiguity of i'rab, which is tied to the interpretation of meaning in the Quran. This subjectivity introduces potential inconsistency in the annotations. Another challenge lies in the complexity of structural embedding in Quranic Arabic, which MASAQ did not fully address due to cost constraints.

Despite these limitations, MASAQ prioritizes ethical compliance. The use of the verified Tanzil Quran text ensures accuracy and integrity, and adherence to the Creative Commons Attribution 3.0 license guarantees proper usage of resources. The research focuses solely on linguistic analysis, avoiding ethical concerns related to sensitive personal data. The authors emphasize fairness and adopt a rigorous methodology to minimize bias and enhance the dataset's reliability for future NLP research.

Acknowledgements

We are deeply thankful to the Deanship of Scientific Research at the University of Jordan for financially supporting this project through a research grant awarded to the principal investigator, Prof. Bassam Hammo.

References

Habash, N., & Roth, R. (2009). Catib: The columbia arabic treebank. Proceedings of the ACL-IJCNLP 2009 conference short papers,

Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., & Beška, E. (2004). Prague Arabic dependency treebank: Development in data and tools. Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools,

Halabi, D., Fayyoumi, E., & Awajan, A. (2021). I3rab: A new Arabic dependency treebank based on Arabic grammatical theory. Transactions on Asian Low-

Resource Language Information Processing, 21(2), 1-32.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M. T., Habash, N., Rambow, O., & Tabessi, D. (2006). Developing and Using a Pilot Dialectal Arabic Treebank http://www.lrec-conf.org/proceedings/lrec2006/summaries/543.html

Sawalha, Majdi; Al-Shargi, Faisal; Yagi, Sane; AlShdaifat, Abdallah T.; Hammo, Bassam; Belajeed, Mariam; Al-Ogaili, Lubna R. (2025) Morphologically-analyzed and syntactically-annotated Quran dataset, Data in Brief, https://doi.org/10.1016/j.dib.2024.111211.

Sawalha, Majdi; Yagi, Sane; Alshargi, Faisal; Hammo, Bassam; Alshdaifat, Abdallah (2024), "MASAQ: Morphologically-Analyzed and Syntactically-Annotated Quran Dataset", Mendeley Data, V7, doi: 10.17632/9yvrzxktmr.7

Sawalha, Majdi; Brierley, Claire; Atwell, Eric (2012). Prosody Prediction for Arabic via the Open-Source Boundary-Annotated Qur'an Corpus, Journal of Speech Sciences 2 (2), 175-191.

Sawalha, Majdi; Brierley, Claire and Atwell, Eric (2014) "Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0)", in: Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop 31st May 2014, Reykjavik, Iceland

Silveira, N., Dozat, T., De Marneffe, M. C., Bowman, S. R., Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014,

Smrz, O., Bielický, V., Kourilová, I., Krácmar, J., Hajic, J., & Zemánek, P. (2008). Prague Arabic dependency treebank: A word on the million words. Proceedings of the workshop on Arabic and local languages (LREC 2008),

Taji, D., Habash, N., & Zeman, D. (2017). Universal dependencies for Arabic. Proceedings of the Third Arabic Natural Language Processing Workshop,

Taji, D., & Habash, N. (2020). PALMYRA 2.0: A Configurable Multilingual Platform Independent Tool for Morphology and Syntax Annotation. Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020),

Tanzil Project. (n.d.). Tanzil project. Retrieved from https://tanzil.net/docs/tanzil_project

Zhao, Y., Zhou, M., Li, Z., & Zhang, M. (2020). Dependency Parsing with Noisy Multi-annotation Data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),

Zaki, Y., Hajjar, H., Hajjar, M., & Bernard, G. (2016). Survey of syntactic parsers of Arabic language. ACM International Conference Proceeding Series.