QuranMorph: Morphologically Annotated Quranic Corpus

Diyam Akra
Birzeit University
dakra@birzeit.edu

Tymaa Hammouda
Birzeit University
thammouda@birzeit.edu

Mustafa Jarrar
Birzeit University
mjarrar@birzeit.edu

Abstract

We present the *QuranMorph* corpus, a morphologically annotated corpus for the Quran (77,429 tokens). Each token in the *QuranMorph* was manually lemmatized and tagged with its part-of-speech by three expert linguists. The lemmatization process utilized lemmas from Qabas, an Arabic lexicographic database linked with 110 lexicons and corpora of 2 million tokens. The part-of-speech tagging was performed using the fine-grained SAMA/Qabas tagset, which encompasses 40 tags. As shown in this paper, this rich lemmatization and POS tagset enabled the *QuranMorph* corpus to be inter-linked with many linguistic resources. The corpus is open-source and publicly available as part of the SinaLab resources at (https://sina.birzeit.edu/quran).

1 Introduction

The Arabic language exists in three forms (Jarrar, 2021): Classical Arabic is the Arabic used between the 6th and the 15th centuries, including the Quran and the prophetic traditions; Modern Standard Arabic (MSA), is the language used in official and formal contexts nowadays; and the Dialectal Arabic includes the diverse spoken varieties across Arab countries. The development of computational resources for Classical Arabic aligns with the growing interest in advancing resources for both MSA and dialects (Darwish et al., 2021).

This paper introduces a lemmatized and POS-tagged Quranic corpus called *QuranMorph*. The lemmatization and POS-tagging were carried out manually by three expert linguistic annotators. The final annotated corpus contains 77,429 words.

2 Related Work

The Quranic Arabic Corpus is the only available morphologically annotated corpora for Classical Arabic (Dukes and Habash, 2010). The POS tagset used to annotate this corpus consists of 44 tags, which were designed to cover traditional Arabic grammar, and the Quran in particular. The challenge is that these POS annotations cannot be easily integrated with the available corpora. Additionally, the lemmatization of the Quranic Corpus did not also utilize lemmas from known lexicographic resources, such as the SAMA database (Graff et al., 2009). Consequently, using the Quranic Corpus alongside other resources as a training dataset is challenging due to its entirely different lemmas and POS tagset compared to other resources.

There are other related morphologically annotated resources available for Modern Standard Arabic (MSA). These include the Prague Arabic Dependency Treebank (Ar-PADT) (Smrž et al., 2008) and LDC's Penn Arabic Treebank (PATB) (Maamouri et al., 2009).

At the dialectal level, several morphologically annotated resources are available for various Arabic dialects. These include the Egyptian Arabic Treebank (ARZATB) (Maamouri et al., 2014), a morphologically annotated treebank for Egyptian Arabic; Curras (Jarrar et al., 2017), a morphologically annotated corpus for Palestinian Arabic; Baladi (Haff et al., 2022), a corpus for Lebanese Arabic; Nabra (Nayouf et al., 2023), a corpus for Syrian Arabic; Gumar (Khalifa et al., 2018), a corpus for Emirati Arabic; and Lisan (Jarrar et al., 2023b), which provides morphologically annotated corpora for Libyan, Yemeni, Iraqi, and Sudanese Arabic.

All of these corpora are annotated using the lemmas and tagsets in either SAMA or Qabas. The LDC's SAMA (Graff et al., 2009) is a stem database containing stems, their lemmas, and compatible affixes. Qabas (Jarrar and Hammouda, 2024) is an Arabic lexicographic database that consists of about $60K$ lemmas, linked with 110 Arabic lexicons and 2 million tokens corpora (Jarrar, 2021; Jarrar and Amayreh, 2019; Jarrar et al., 2018). Qabas is also linked with other resources for synonymy (Ghanem et al., 2023; Jarrar et al.,

2021), Word Sense Disambiguation (Jarrar et al., 2023a; Al-Hajj and Jarrar, 2021), among other resources. The idea is that corpora annotated with similar or closely related lemmas and tagsets are easier to unify and utilize for joint training.

3 The Quran Corpus

As mentioned earlier, the only valuable corpus for the Quran was developed in 2011 by Dukes and Habash (2010). This corpus consists of 128, 219 forms (clitics) extracted from the Quran, each is annotated with a set of morphological features. To ensure compatibility with this corpus, we decided to adopt the same segmentation and tokenization methodology. The corpus includes a location feature, which specifies the position of each clitic in terms of chapter, verse, word, and its position within the word. For instance, a location of (1:1:1:1) refers to the first clitic in the first word of the first verse in the first chapter. In alignment with this, we adopted the same tokenization but focused solely on words rather than clitics. In other words, we combined clitics to form complete words. As a result, each chapter consists of verses, and each verse consists of words, represented by the location format (1:1:1). As shown in Table 1 the total number of words in the Quran is 77, 429 (which are 19, 009 unique words) in 6, 235 verses. We then annotated these words with their corresponding part-of-speech tags and lemmas, as explained in the following sections.

Notably, our annotations can be seamlessly integrated with those in the Quranic Corpus using the location format (1:1:1), allowing for the combined use of both corpora.

4 Annotation Methodology

This section presents the lemmatization and POS tagging of the Quran Corpus.

The corpus was uploaded to the Tawseem Web Annotation tool (See the screenshot in Figure 1). The corpus was then assigned to three expert linguists for annotation, which they completed in approximately 400 working hours at a rate of 5 USD per hour. Tawseem's smart features streamline the annotation process by automating and validating tasks, significantly boosting both quality and productivity.

As shown in the figure, the first column represents the context (i.e., verse) where the target word to be annotated is highlighted. The third and fourth columns are the POS and lemma assigned by the linguists with the assistance of the tool. Tawseem integrates with the Alma morphological tagger (Jarrar et al., 2024) from SinaTools (Hammouda et al., 2024), allowing it to suggest annotations (lemmas and POS tags) for review and confirmation by annotators.

4.1 Lemmatization

To confirm the lemma suggested by the Tawseem tool the annotator has to carefully review the context (i.e., verse) containing the target word to ensure an accurate understanding of its meaning. This step is crucial for selecting the most appropriate lemma, taking into account the word's syntactic role and its semantic relationship with the surrounding text. If the suggested lemma is not appropriate, the annotator searches Qabas for other lemmas using the search bar located at the top left of Figure 1. If no suitable lemma is found, a new Lemma entry is added to Qabas, complete with relevant linguistic and morphological details.

The idea of using Qabas lemmas to annotate the Quran corpus stems from their unique linkage to lemmas in 110 lexicons. Additionally, Qabas lemmas are used to annotate various MSA and dialectal corpora, totaling around 2 million tokens. This integration allows the Quran corpus to be utilized alongside other corpora for training purposes, offering a more comprehensive resource for Classical, MSA, and dialectal Arabic.

The online version of Qabas[1] allows users to search for a lemma and view the contexts in which this lemma appears in the Quran and other corpora.

4.2 POS Tagging

Similarly, to confirm the POS tag suggested by the Tawseem tool, the annotator must carefully review the entire verse containing the target word to ensure an accurate understanding of its context. This step is essential for assigning the most appropriate POS tag. If the suggested POS tag is unsuitable, the annotator can select an alternative from the drop-down Qabas tagset, which includes the same 40 POS tags used in LDC's SAMA (Graff et al., 2009).

To sum up, the Table 1 presents statistics about the corpora.

---

[1]Qabas: https://sina.birzeit.edu/qabas

Search panel:

حمد 🔍

☑ اسم ☑ فعل ☑ وظيفية

⦿	حَمْدٌ (اسم)	2023157200
	حَمْد [مفرد]	

⦾	حَمِدَ (فعل ماضي)	2021029900
	حمِدَ يَحمَد، حَمْدًا، فهو حامد، والمفعول حَميد	

⦾	حَمَّدَ (فعل ماضي)	2021030300
	حمَّدَ يحمِّد، تحميدًا، فهو مُحمِّد، والمفعول مُحمَّد	

Annotation table:

MSA Lemma	POS	Word	Context
حَمْدٌ	اسم	ٱلْحَمْدُ	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ
اللَّه	اسم علم	لَّهِ	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ
رَبٌّ	اسم	رَبِّ	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ
عَالَمٌ	اسم	ٱلْعَٰلَمِينَ	ٱلْحَمْدُ لِلَّهِ رَبِّ ٱلْعَٰلَمِينَ
لَا	أداة نفي	وَلَا	۟عَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ
صِرَاطٌ	اسم	صِرَٰطَ	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُو
ٱلَّذِي	اسم موصول	ٱلَّذِينَ	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْر ٱلْمَغْضُوبِ عَلَيْه
عَلَى 2	حرف جر	عَلَيْهِمْ	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْر ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّين
مُسْتَقِيمٌ	صفة	ٱلْمُسْتَقِيمَ	ٱهْدِنَا ٱلصِّرَٰطَ ٱلْمُسْتَقِيمَ
أَنْعَمَ | أَنْعَمَ عَلَى | أَنْعَمَ فِي | أَنْعَمَ لِ	فعل ماضي	أَنْعَمْتَ	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱ

Figure 1: Screenshot of the Tawseem Annotation Tool with example annotations

Words	77,429
Unique Words	19009
Verses	6,235
Unique Nouns	9,789
Unique Verbs	8,265
Unique Functional Words	1,036
Unique Lemmas	4,616
Unique Nouns Lemmas	3,057
Unique Verbs Lemmas	1,479
Unique Functional Words Lemmas	173

Table 1: Statistics about our *QuranMorph* Corpus

5 Discussion

Due to the unique nature of the Quran, annotators encounter distinct challenges that can be grouped into the following categories:

- Difficulties arose due to the Quranic script (الرسم القرآني / *ālrsm ālqrāny*), which is not same as standard Arabic orthography, making the annotation process more complex. These differences include variations in spelling, diacritization, and letter forms, requiring annotators to carefully analyze each word within its contextual and historical linguistic framework. For example, the verse (الذين يؤمنون بالغيب و يقيمون الصلاة و مما رزقناهم ينفقون) appears as (ٱلَّذِينَ يُؤْمِنُونَ بِغيب وَيُقِيمُونَ ٱلصَّلَوٰةَ وَمِمَّا رَزَقَهم يُنفِقُونَ) in Quranic script.

- Annotators often had to consult Quranic exegesis (تفسير القرآن / *tfsyr ālqrān*) to accurately determine the most suitable lemma for a target word. This step is necessary be-

cause Quranic words, while often resembling standard Arabic vocabulary, frequently carry context-dependent meanings that differ from their conventional usage. For instance, in the verse: (ثُمَّ بَدَّلْنَا مَكَانَ السَّيِّئَةِ الْحَسَنَةَ حَتَّى عَفَوا)-Quran 7:95, the term (عَفَوا / *'afawā*) does not mean forgiveness or pardon as it does in Modern Standard Arabic. Instead, within this Quranic context, it signifies an increase or growth.

- The same word can appear in diverse contexts, leading to variations in its meaning and associated lemma. Consequently, annotators must carefully analyze grammatical conjugations and semantic nuances to accurately determine the correct lemma. For example, in the Quran, the word (ضرب / *ḍrb*) carries different meanings depending on the context: In the verse (أَلَمْ تَرَ كَيْفَ ضَرَبَ اللَّهُ مَثَلًا كَلِمَةً طَيِّبَةً كَشَجَرَةٍ طَيِّبَةٍ)-Quran 14:24, the verb (ضَرَبَ / *ḍaraba*) means (to provide an example) or (to illustrate a concept). In contrast, in the verse (وَإِذَا ضَرَبْتُمْ فِي الْأَرْضِ فَلَيْسَ عَلَيْكُمْ جُنَاحٌ)-Quran 4:101, (ضَرَبْتُمْ / *ḍarabtum*) conveys (to travel) or (to journey through the land).

6 Conclusion

This paper presents the *QuranMorph* a morphologically annotated corpus for the Quran (77,429 tokens). Each token in the *QuranMorph* is lemmatized and tagged with its part of speech. The lemmatization process utilized the lemmas from Qabas. The corpus was also tagged with parts of speech using the SAMA/Qabas tagset.

7 Acknowledgment

References

Moustafa Al-Hajj and Mustafa Jarrar. 2021. ArabGloss-BERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A Panoramic survey of Natural Language Processing in the Arab Worlds. *Commun. ACM*, 64(4):72–81.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of quranic arabic. In *Lrec*, pages 2530–2536.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tem Buckwalter. 2009. Standard arabic morphological analyzer(sama) version 3.1. *Linguistic Data Consortium LDC2009E73*, pages 53–56.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + Baladi: Towards a Levantine Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation(LREC 2022)*, Marseille, France.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. SinaTools: Open Source Toolkit for Arabic Natural Language Understanding. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

Mustafa Jarrar. 2021. The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. *Applied Ontology Journal*, 16(1):1–26.

Mustafa Jarrar, Diyam Akra, and Tymaa Hammouda. 2024. ALMA: Fast Lemmatizer and POS Tagger for Arabic. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic-multilingual database with a lexicographic search engine. In *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):2-s2.0-85001544989.

Mustafa Jarrar and Tymaa Hasanain Hammouda. 2024. Qabas: An Open-Source Arabic Lexicographic Database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13363–13370, Torino, Italy. ELRA and ICCL.

Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, and Khaled Shaalan. 2021. Extracting Synonyms from Bilingual Dictionaries. In *Proceedings of the 11th International Global Wordnet Conference (GWC2021)*, pages 215–222. Global Wordnet Association.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023a. SALMA: Arabic Sense-annotated Corpus and WSD Benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Asian and Low-Resource Language Information Processing*, 18(2):10:1–10:21.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, Irqi, Libyan, and Sudanese Arabic Dialect Copora with Morphological Annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati arabic. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 3839–3846.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2009. Creating a methodology for large-scale correction of treebank annotation: The case of the arabic treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt*.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian Arabic Dialects with Morphological Annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 12–23. ACL.

Otakar Smrž, Viktor Bielick´ybielick´y, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague arabic dependency treebank: A word on the million words. In *Proceedings of the workshop on Arabic and local languages (LREC 2008)*, pages 16–23.