Utilizing The Quranic Corpus to Enhance Arabic Language Learning: Opportunities and Challenges

Kamal Yusuf
UIN Sunan Ampel Surabaya, Indonesia
kamalyusuf@uinsa.ac.id

Abstract:

The Quranic Arabic Corpus is a rich linguistic resource, offering morphological, syntactic, and semantic analysis for every word in the Quran. Developed by the University of Leeds, this corpus facilitates an in-depth understanding of the sentence structure and vocabulary of classical Arabic. Utilizing this corpus in Arabic language learning can improve linguistic analysis skills, deepen grammatical understanding, and enrich appreciation of the beauty of the Quranic language. However, despite its great potential, the application of this corpus in Arabic language teaching is still limited. This study aims to explore the benefits of the Quranic Corpus in enhancing Arabic language teaching, as well as identify the challenges and opportunities this technology presents in Arabic language education in educational institutions.

Keyword: Quranic corpus; Arabic language learning; educational technology; Arabic linguistics

INTRODUCTION

Arabic language learning has undergone significant developments with the advancement of educational technology (Ritonga et al., 2024). One tool that can be utilized to enrich the understanding of Arabic is the Quranic Arabic Corpus. This corpus, developed by the University of Leeds (Dukes, 2017), provides an in-depth analysis of the morphology, syntax, and semantics of each word in the Quran, making it a very valuable linguistic resource. By integrating this corpus into Arabic language learning, teachers can improve students' linguistic analysis skills, deepen their understanding of sentence structure, and enrich their appreciation of the beauty and depth of meaning contained in the Quran (Arifianto, 2021).

However, despite the great potential of the Quranic Corpus, its use in the context of Arabic language education is still limited. Many educational institutions have not fully utilized this technology in teaching Arabic. Technical constraints, limited resources, and lack of training for teachers are some of the factors that hinder the wider implementation of this corpus in the learning process.

This study aims to explore the benefits that can be obtained from the use of the Quranic Corpus in teaching Arabic. In addition, this paper will also identify the challenges and opportunities presented by this technology in the context of Arabic language education, with the hope of making a meaningful contribution to the development of a more effective Arabic language curriculum in the future (Alrabiah et al., 2014; Al-Maadeed et al,. 2014; Moser, 2021.

METHOD

This study uses a descriptive qualitative approach with primary data collection sourced from the Arabic Quranic Corpus available on the site https://corpus.quran.com/. This data includes morphological, syntactic, and semantic analysis for each word in the Quran used in Arabic language learning. Other relevant sources, such as journal articles and related books, are also used to explore the context and application of the corpus in Arabic language (Yusuf & Puspita, 2020; Yusuf, 2020; Puspita & Yusuf, 2020).

Data analysis was carried out using thematic analysis, where the main themes that emerged from the literature review and data taken from the corpus will be compiled and analyzed. The findings of this study will be presented in narrative form that describes the benefits of using the corpus, challenges in its implementation, and opportunities for further development.

FINDINGS AND DISCUSSION

Arabic and the Qur'anic Corpus: Oppotunities

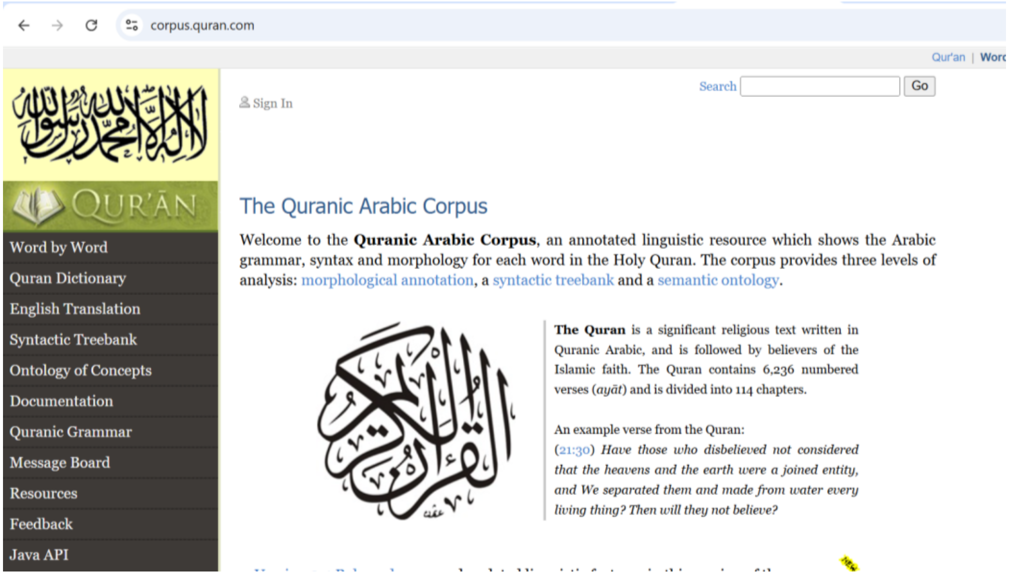
Quranic Arabic is a variant of Classical Arabic used in the holy texts of Muslims. It has a unique morphological and syntactic structure, which distinguishes it from Modern Standard Arabic. This uniqueness includes the use of complex word forms, variations in sentence structure, and the use of a rich and meaningful style of language (Masood & Nousheen, 2025; Mohamed & Shokry, 2022; Zeroual & Lakhouaja, 2016).

To study and understand these uniqueness, the Quranic Arabic Corpus was developed by the University of Leeds. This corpus provides a rich linguistic resource, including morphological, syntactic, and semantic analysis for every word in the Qur'an. With over 77,000 annotated words, this corpus allows researchers to conduct in-depth analysis of the language structure of the Qur'an in a systematic and detailed manner (Dukes & Habash, 2010).

This corpus consists of three main levels of analysis:

1. Morphological Annotation: Each word in the Qur'an is annotated with morphological information, including word types, tenses, and other grammatical features.
2. Syntactic Treebank: Provides a syntactic tree structure that describes the relationships between words in a sentence, allowing for deeper syntactic analysis.
3. Semantic Ontology: Connects entities mentioned in the Qur'anic verses with concepts in the ontology, helping in understanding the semantic meaning of the text.

With this corpus, the process of learning Arabic, especially in the context of the Qur'an, can be carried out with a more scientific and systematic approach.



The Quranic Arabic Corpus

Welcome to the **Quranic Arabic Corpus**, an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The corpus provides three levels of analysis: morphological annotation, a syntactic treebank and a semantic ontology.

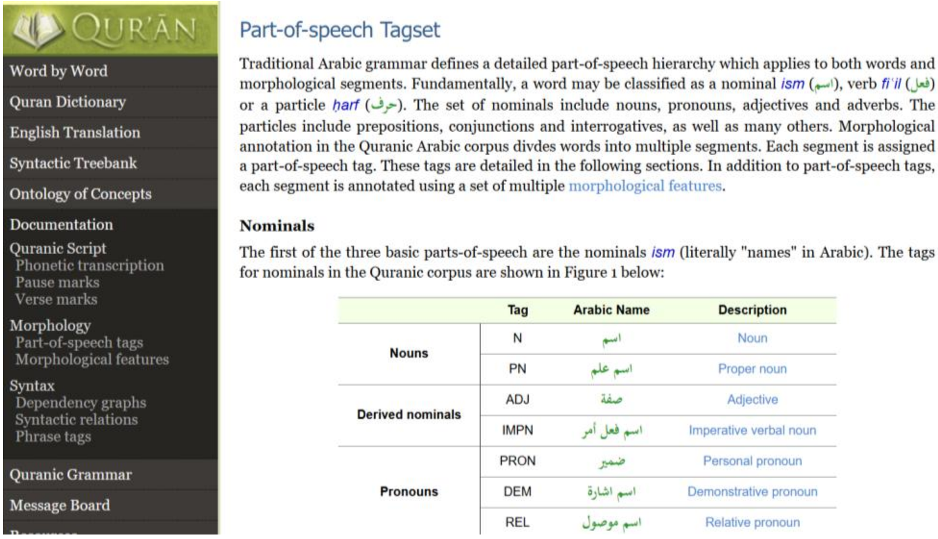
The Quran is a significant religious text written in Quranic Arabic, and is followed by believers of the Islamic faith. The Quran contains 6,236 numbered verses (*ayāt*) and is divided into 114 chapters.

An example verse from the Quran:
(21:30) *Have those who disbelieved not considered that the heavens and the earth were a joined entity, and We separated them and made from water every living thing? Then will they not believe?*

Figure 1: The Interface of Quranic Arabic Corpus

Morphological Annotation

The Quranic Arabic Corpus provides in-depth morphological annotation for each word in the Quran, covering morphological, syntactic, and semantic analysis. Each word is annotated with part-of-speech tags and other morphological features, such as word form, root, and lemma. For example, the word "kitābuhu" (كتابُهُ) is annotated with segments that each have part-of-speech tags such as CONJ (conjunction), V (verb), and PRON (pronoun), as well as morphological features such as grammatical gender, number, and case.



Part-of-speech Tagset

Traditional Arabic grammar defines a detailed part-of-speech hierarchy which applies to both words and morphological segments. Fundamentally, a word may be classified as a nominal *ism* (اسم), verb *fi'il* (فعل) or a particle *ḥarf* (حرف). The set of nominals include nouns, pronouns, adjectives and adverbs. The particles include prepositions, conjunctions and interrogatives, as well as many others. Morphological annotation in the Quranic Arabic corpus divdes words into multiple segments. Each segment is assigned a part-of-speech tag. These tags are detailed in the following sections. In addition to part-of-speech tags, each segment is annotated using a set of multiple morphological features.

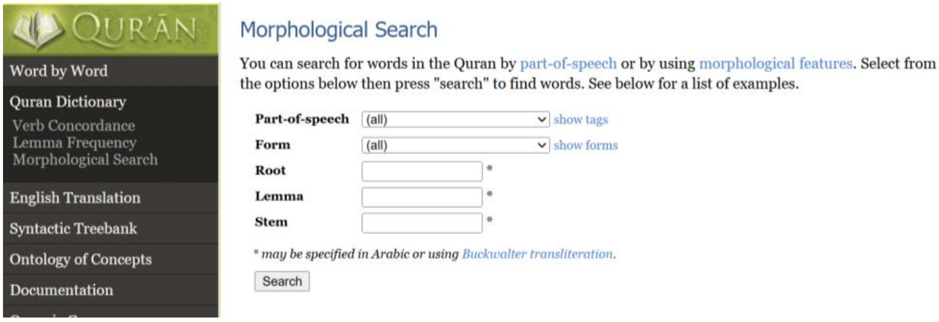
Nominals

The first of the three basic parts-of-speech are the nominals *ism* (literally "names" in Arabic). The tags for nominals in the Quranic corpus are shown in Figure 1 below:

	Tag	Arabic Name	Description
Nouns	N	اسم	Noun
	PN	اسم علم	Proper noun
Derived nominals	ADJ	صفة	Adjective
	IMPN	اسم فعل أمر	Imperative verbal noun
Pronouns	PRON	ضمير	Personal pronoun
	DEM	اسم اشارة	Demonstrative pronoun
	REL	اسم موصول	Relative pronoun

Figure 2: Part of Speech Tagset

The morphological search feature on the Quranic Arabic Corpus website allows users to search for words in the Quran based on sentence parts or other morphological features. Users can select the type of word such as noun, verb, or particle, and specify the form, root, lemma, or stem. For example, to search for all nouns that are proper nouns, users can select "Proper noun" as the sentence part and then press the "search" button to find the words.



Morphological Search

You can search for words in the Quran by part-of-speech or by using morphological features. Select from the options below then press "search" to find words. See below for a list of examples.

Part-of-speech (all) show tags
Form (all) show forms
Root *
Lemma *
Stem *

* *may be specified in Arabic or using Buckwalter transliteration.*

Figure 3: Morphological Search

In addition, this search feature also allows searching based on root words or lemmas. For example, to search for all words derived from the root word "ktb" (كتب), users can enter "ktb" as the root word and press the "search" button. Likewise, to search for all forms of the lemma "kitāb" (كتاب), users can enter "kitāb" as the lemma and press the "search" button. With detailed morphological annotations and advanced search features, the Qur'anic Arabic Corpus is a very useful resource for researchers and teachers in analyzing the language structure of the Qur'an systematically and in depth.

Syntax Treebank

The Qur'anic Arabic Corpus provides in-depth syntactic annotations through the Syntax Treebank which is based on the theory of dependent grammar. This approach describes the relationships between words in a sentence using mathematical graph theory, allowing the visualization of the syntactic structure of the Qur'anic verses in the form of a dependent graph. Each word in a sentence is connected to another word that functions as a head, forming a tree structure that describes the syntactic relationships between elements in a sentence.

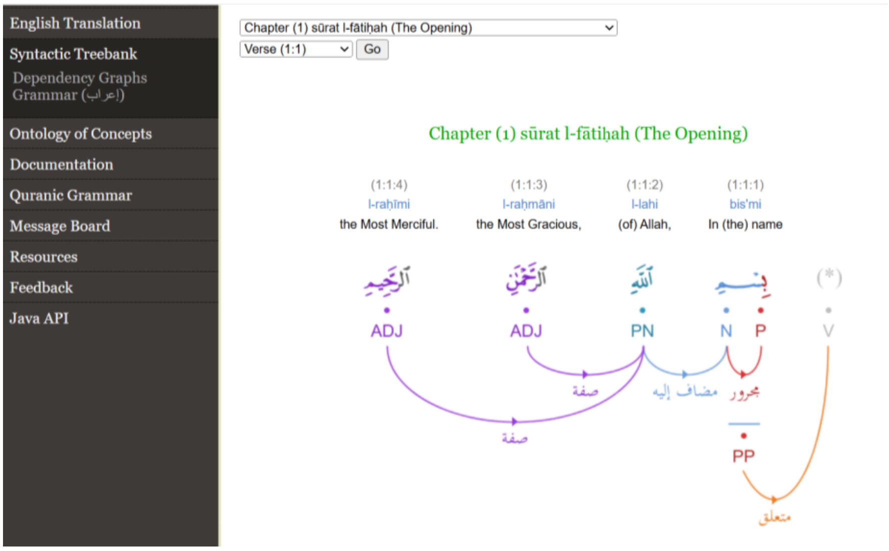

Chapter (1) sūrat l-fātiḥah (The Opening)
Verse (1:1)

Figure 4: Quranic Syntax Parsing

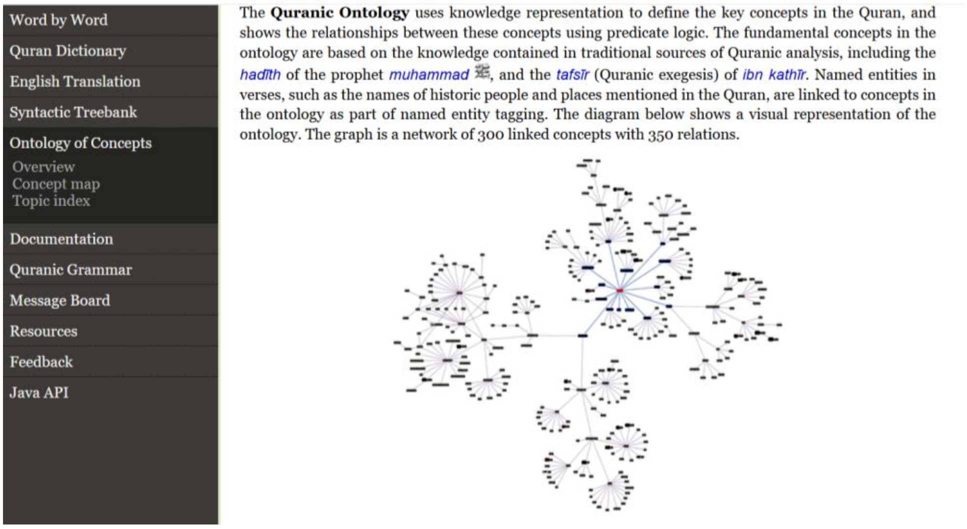
This syntactic annotation covers various important elements in the sentence structure, such as subject, predicate, object, and complement, as well as other grammatical relationships. For example, in verse (2:31), the word "allama" (taught) functions as a predicate that has the object "ādama" (Adam), which in turn has the complement "al-asmāʾ" (the names). This structure illustrates how the words in the sentence relate to each other and form the overall meaning.

Through this syntactic annotation, researchers and educators can analyze the sentence structure in the Qur'an more systematically and in depth. This allows for a better understanding of how the elements in the sentence interact and convey meaning, as well as providing insight into the use of language style and syntactic structure in the holy text.

Semantic Ontology

The Semantic Ontology in the Quranic Arabic Corpus uses knowledge representation to define key concepts in the Quran and to show relationships between them using predicate logic. The basic concepts in this ontology are based on knowledge contained in traditional sources of Quranic analysis, including the hadith of the Prophet Muhammad and the commentary of Ibn Kathir. Named entities in the Quranic verses, such as names of historical people and places, are linked to concepts in the ontology as part of the labeling of named entities.

The ontology consists of over 300 concepts that are interconnected through around 350 semantic relationships. For example, the relationship “Satan is a jinn” in the ontology represents the knowledge in the Quran that the individual known as Satan belongs to a group of beings called jinn. Other concepts in the ontology are grouped into logical categories, according to the properties they possess. For example, the Sun, Earth, and Moon are grouped into the category “Astronomical Body.”



The **Quranic Ontology** uses knowledge representation to define the key concepts in the Quran, and shows the relationships between these concepts using predicate logic. The fundamental concepts in the ontology are based on the knowledge contained in traditional sources of Quranic analysis, including the *hadīth* of the prophet *muhammad* ﷺ, and the *tafsīr* (Quranic exegesis) of *ibn kathīr*. Named entities in verses, such as the names of historic people and places mentioned in the Quran, are linked to concepts in the ontology as part of named entity tagging. The diagram below shows a visual representation of the ontology. The graph is a network of 300 linked concepts with 350 relations.

Figure 5: The Quranic ontology

Through this semantic ontology, the Arabic Quranic Corpus enables a deeper analysis of the meaning of the Quranic text. By connecting the words in the verses to the concepts in the ontology, we can understand the broader context and meaning of the text. It also helps in resolving ambiguities, such as in the case of pronoun resolution, where a pronoun attached to a verb or noun can be linked back to the entity referred to in the ontology. In addition, this semantic ontology is also used in the tagging of named entities, where entities such as the “Night of Decree” (Lailatul Qadr) in a particular verse are linked to the relevant concept in the ontology, helping in understanding the historical and theological context of the verse.

Daily Quranic Corpus: Challenges

Although the Qur’anic Arabic Corpus offers many benefits in Arabic language learning, its application in Arabic language education faces several challenges that need to be overcome. These challenges can be divided into several main categories, including technical, pedagogical, and institutional challenges.

Technical Challenges

One of the biggest challenges in utilizing the Qur’anic Corpus is the limited access and technical knowledge related to its use. Although the corpus is accessible online, many teachers and researchers do not have the technical skills needed to use it to its full potential. This includes the ability to use morphological search tools, understand syntactic annotations, and analyze semantic data in the corpus. To overcome these challenges, adequate technical training is needed for teachers and researchers, as well as support in the form of learning materials that can help them understand how the corpus works and its benefits.

Pedagogical Challenges

Pedagogically, the integration of the Qur’anic Corpus in Arabic language teaching requires a significant shift in teaching methods that are more data- and technology-based. Many Arabic language teachers still rely on traditional methods that do not pay attention to in-depth linguistic analysis. Therefore, to use this corpus effectively, teaching must be more oriented towards developing language analysis skills, which can change the way students understand sentence structure, morphology, and semantics in the Qur'anic text. This change requires a more dynamic and technology-based curriculum adaptation, as well as the development of teaching materials that are relevant to the use of the corpus.

Institutional Challenges

At the institutional level, one of the biggest challenges is the limited infrastructure and support from educational institutions. Many educational institutions, especially in developing countries, are still limited in providing the technological facilities needed to optimally access and utilize the Qur'anic Corpus. In addition, budget constraints are often an obstacle in providing adequate hardware and software to access and use this corpus. To overcome this problem, there needs to be a policy from the government and educational institutions to prioritize investment in educational technology and the provision of facilities that support the use of linguistic analysis tools such as the Qur'anic Corpus.

Challenges in Disseminating Knowledge

Although the Qur'anic Corpus provides a lot of useful linguistic information, not many Arabic language teaching has integrated the use of this corpus in their teaching. The dissemination of knowledge about the importance of utilizing this corpus in Arabic language learning is still limited. Therefore, the next challenge is how to disseminate information about the benefits and potential offered by the Qur'anic Corpus, and motivate more teachers and institutions to integrate this technology into their curriculum.

CONCLUSION

The Qur'anic Arabic Corpus offers great potential to enrich Arabic language learning by providing in-depth morphological, syntactic, and semantic analysis of the Qur'anic text, allowing for a better understanding of sentence structure and meaning. However, its utilization is still limited by technical, pedagogical, and institutional challenges, such as limited technical knowledge of teachers, the need for a technology-based curriculum, and inadequate educational infrastructure. To maximize the potential of this corpus, teacher training, curriculum development that integrates technology, and educational policies that support the provision of adequate facilities are needed. Thus, despite these challenges, with the right efforts, the Qur'anic Corpus can be a very effective tool in improving the quality of Arabic language learning in the future.

REFERENCES

Alasmari, J. S. N. (2020). *A Comparative Analysis of The Arabic and English Verb Systems Using the Qur'an Arabic Corpus [A corpus-based study]* (Doctoral dissertation, University of Leeds).

Al-Maadeed, S., AlJa'am, J., Khalifa, B., & Abou Elsaud, S. (2021, April). MOALLEMCorpus: A large-scale multimedia corpus for children education of Arabic vocabularies. In *2021 IEEE Global Engineering Education Conference (EDUCON)* (pp. 885-890). IEEE.

Alrabiah, M., Alhelewh, N., Al-Salman, A., & Atwell, E. S. (2014). An empirical study on the Holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics (IJCL)*, *5*(1), 1-13.

Arifianto, M. L. (2021). Utilizing the Quranic Arabic Corpus as a supplementary teaching and learning material for Arabic syntax: An overview of a web-based Arabic linguistics corpus. *KnE Social Sciences*, 403-412.

Dukes, K., & Habash, N. (2010). Morphological Annotation of Quranic Arabic. In *Lrec* (pp. 2530-2536).

Dukes, Keis. 2017.corpus.quran.com. Leeds: University of Leeds

Masood, S., & Nousheen, S. (2025). Lexical Patterns and Their Semantic Implications in Surah-Al-Hujrat: A Corpus-Based Approach. *Journal of Asian Development Studies*, *14*(1), 508-518.

Mohamed, E. H., & Shokry, E. M. (2022). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*, *34*(3), 934-945.

Moser, J. (2021). Evaluating Arabic textbooks: A corpus-based lexical frequency study. *International Journal of Applied Linguistics*, *31*(2), 248-263.

Puspita, D., & Yusuf, K. (2020). Sketching the semantic change of Jahanam and Hijrah: A corpus based approach to manuscripts of Arabic-Indonesian Lexicon. *Arabi: Journal of Arabic Studies*, *5*(1), 1-10.

Ritonga, M., Mudinillah, A., Wasehudin, W., Julhadi, J., Amrina, A., & Shidqi, M. H. (2024). The effect of technology on Arabic language learning in higher education. *Journal of Education and Learning (EduLearn)*, *18*(1), 116-127.

Yusuf, K. (2020). Data driven learning by discovering lexical bundles using corpus resources. In *International Conference on English Language Teaching (ICONELT 2019)* (pp. 47-50). Atlantis Press.

Yusuf, K., & Puspita, D. (2020). Diachronic corpora as a tool for tracing etymological information of Indonesian-Malay lexicon. *Register Journal*, *13*(1), 153-182.

Zeroual, I., & Lakhouaja, A. (2016). A new Quranic Corpus rich in morphosyntactical information. *International Journal of Speech Technology*, *19*, 339-346.