LEVERAGING CORPUS LINGUISTICS FOR EFFECTIVE ARABIC VOCABULARY LEARNING

Kamal Yusuf
UIN Sunan Ampel Surabaya
Corresponding Email: kamalyusuf@uinsa.ac.id

Abstract-This paper explores the potential of corpus-based approaches in enhancing Arabic language learning, particularly in the acquisition of vocabulary. By leveraging the power of corpus linguistics, this paper aims to develop a more effective and efficient method for Arabic language instruction. This research focuses on the utilization of a large-scale corpus of Arabic texts to identify patterns and trends in vocabulary usage, shedding light on the most common and relevant words in the language. This corpus-based approach has the potential to revolutionize the way Arabic is taught, moving beyond traditional rote memorization and towards a more nuanced understanding of the language. By analyzing the corpus, we can uncover the underlying structures and relationships between words, providing learners with a more comprehensive and contextualized understanding of the language. The implications of this research are significant, offering a new paradigm for Arabic language instruction that is grounded in empirical evidence and informed by the latest advances in corpus linguistics.

Keywords: Arabic learning, corpus linguistics, corpus based-approach, vocabulary

Introduction

The Arabic language, with its rich cultural heritage and significant role in international communication, has become increasingly important for language learners and scholars alike. However, the acquisition of Arabic, particularly for non-native speakers, remains a challenging task due to its unique script, grammar, and vocabulary. Vocabulary learning is a crucial aspect of language acquisition, and Arabic, with its vast vocabulary and complex morphology, presents a significant challenge in this regard. Traditional methods of language instruction, such as rote memorization and grammar rules, have been shown to be ineffective in promoting vocabulary acquisition and language proficiency.

In recent years, corpus linguistics has emerged as a promising approach to language learning, offering a more authentic and contextualized way of learning language. Corpus linguistics involves the analysis of large databases of language texts to identify patterns and trends in language use, providing insights into how language is used in real-life contexts. This approach has been successfully applied to various languages, including English, French, and Spanish, with significant improvements in language learning outcomes.

The application of corpus linguistics to Arabic language learning is particularly timely, given the growing demand for Arabic language instruction and the need for more effective and engaging language learning materials (Arts et al,. 2024). However, despite the potential of corpus linguistics, there is a lack of research on its application to Arabic language learning, particularly in the context of vocabulary acquisition.

This paper aims to address this gap by exploring the potential of a corpus-based approach to enhance Arabic vocabulary acquisition for non-native learners. Two corpus tools are explored: The Quranic Arabic Corpus (Nur et al,. 2020) and Sketch Engine (Baisa et al,. 2015; Zerrin, 2023). By analyzing a large corpus of Arabic texts, we aim to identify the most frequent and contextually relevant vocabulary items and develop interactive online modules that facilitate learner engagement and active learning. The study contributes to the growing

body of research on the application of corpus linguistics in language learning, highlighting the potential of corpus-based approaches to enhance the efficiency and effectiveness of Arabic language instruction.

The potential of this approach is immense. By using a corpus-based approach, we can revolutionize the way Arabic is taught, moving beyond traditional rote memorization and towards a more nuanced understanding of the language. By analyzing the corpus, we can uncover the underlying structures and relationships between words, providing learners with a more comprehensive and contextualized understanding of the language.

The implications of this research are significant, offering a new paradigm for Arabic language instruction that is grounded in empirical evidence and informed by the latest advances in corpus linguistics. As we delve deeper into this research, we hope to provide valuable insights that can transform the field of Arabic language instruction and contribute to the broader field of language learning and teaching.

Literature Review

The use of corpus linguistics in language learning has been gaining popularity in recent years, and for good reason. Corpus linguistics offers a unique approach to language learning, one that is grounded in the analysis of real language use. By analyzing large databases of language texts, corpus linguistics provides insights into how language is used in real-life contexts, allowing learners to develop a more nuanced understanding of language. One of the key benefits of corpus linguistics is its ability to provide learners with authentic language use. Traditional language learning materials often rely on artificial examples and exercises, which can be misleading and unrepresentative of real language use.

Corpus linguistics, on the other hand, provides learners with a window into how language is used in real-life contexts, allowing them to develop a more authentic understanding of language. Another benefit of corpus linguistics is its ability to provide learners with a more comprehensive understanding of language. By analyzing large databases of language texts, corpus linguistics can identify patterns and trends in language use that may not be apparent through traditional language learning methods. This can include insights into how language is used in different contexts, how language is used to convey meaning, and how language is used to establish relationships between words. Corpus linguistics has been successfully applied to various languages, including English, French, and Spanish. In the context of Arabic language learning, corpus linguistics has the potential to revolutionize the way Arabic is taught and learned.

Arabic is a language with a rich cultural heritage and a complex grammar and vocabulary system, and corpus linguistics can provide learners with a more nuanced understanding of these complexities. In addition to its benefits for learners, corpus linguistics also has the potential to improve the efficiency and effectiveness of language teaching. By providing teachers with insights into how language is used in real-life contexts, corpus linguistics can help teachers to develop more effective language teaching materials and methods (Yusuf and Puspita, 2020). This can include the development of more authentic language learning materials, the use of more effective language teaching methods, and the provision of more support for language learners. Despite the potential of corpus linguistics, there is a lack of research on its application to Arabic language learning (Puspita and Yusuf,

2023; Siddiq et al,. 2021; Yusuf, 2020). This study aims to address this gap by exploring the potential of a corpus-based approach to enhance Arabic vocabulary acquisition for non-native learners. By analyzing a large corpus of Arabic texts, this paper tries to identify the most frequent and contextually relevant vocabulary items where it can be developed become interactive online modules that facilitate Arabic learner engagement and active learning (Buckwalter and Parkinson, 2014; Jawharah et at,. 2017; Muhammad, 2021; Puspita and Yusuf, 2020).

Method

This study employed a qualitative approach to explore the potential of a corpus-based approach to enhance Arabic vocabulary acquisition for non-native learners. A large corpus of Arabic texts was analyzed to identify the most frequent and contextually relevant vocabulary items.

Result and Discussion

In this section, two corpus will be described. The descriptions will be about the use, advanteges, and examples use of The Quranic Arabic Corpus and Sketch Engine.

The Quranic Arabic Corpus

The Quranic Arabic Corpus is a treasure trove for anyone interested in the linguistic intricacies of the Quran, the holy book of Islam. This digital resource meticulously analyzes the Arabic used in the Quran, providing a deep understanding of its grammar, sentence structure, and word meaning. Encompassing all 77,430 words of the Quran, the Corpus offers a multi-layered breakdown. The core lies in morphological and syntactic annotations. Morphology refers to the individual components that make up words, while syntax delves into how these words are arranged to form sentences. With this granular analysis, researchers can dissect the Quran's language with unparalleled precision.

The Corpus goes beyond basic breakdowns. It incorporates a "syntactic treebank," which visually maps the grammatical relationships between words. Imagine a family tree, but instead of relatives, it shows how each word connects to others within a sentence. This treebank offers a clear picture of the Quran's sentence structure, aiding researchers in grasping the flow and emphasis of the text.

The Corpus acknowledges the significance of meaning as well. While it doesn't provide its own interpretations, it incorporates established English translations, allowing researchers to connect the linguistic analysis to the Quran's message. Developed by the University of Leeds, the Quranic Arabic Corpus serves as a powerful tool for scholars and students alike. By offering a window into the language of the Quran, it fosters a deeper appreciation for this foundational Islamic text.

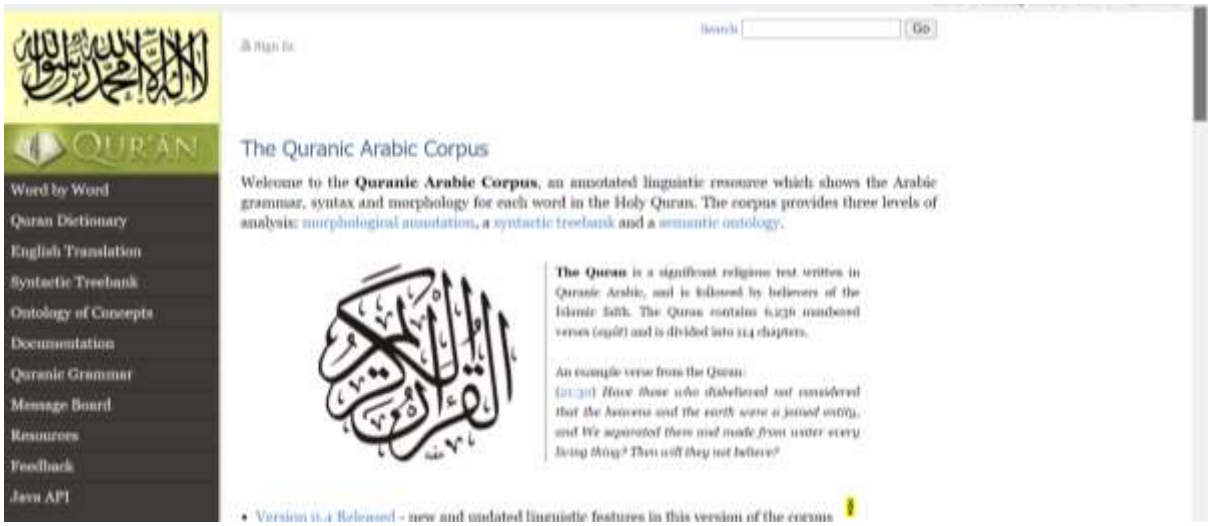

Figure 1 The Quranic Arabic Corpus Interface

Use of The Quranic Arabic Corpus

The Quranic Arabic Corpus transcends being a mere archive of Quranic language. It empowers researchers in various ways:

1. Linguistic Analysis: The detailed morphological annotations allow scholars to dissect the Quranic vocabulary. They can examine how words are formed, their root meanings, and any variations present. This unveils the subtle nuances within the text, potentially revealing deeper interpretations.

2. Syntactic Exploration: The syntactic treebank is a game-changer. By visualizing the grammatical connections between words, researchers can analyze the sentence structure in detail. This helps understand the intended emphasis and flow of the Quranic verses. Imagine identifying a powerful statement not just by its words, but by how those words are intricately woven together.

3. Comparative Studies: The Corpus facilitates comparisons between different parts of the Quran. Scholars can identify grammatical patterns or recurring word choices across chapters. This comparative approach can shed light on the Quran's thematic development or stylistic variations.

4. Unlocking Historical Context: The Quranic language predates modern Arabic. Using the Corpus, researchers can delve into the historical context of the language used. This analysis can provide insights into the cultural and social environment surrounding the Quran's revelation.

5. Machine Learning and Quranic Studies: The Corpus paves the way for utilizing machine learning in Quranic research. By feeding the vast amount of data into algorithms, researchers can discover hidden patterns or explore statistical relationships within the text. This opens doors to entirely new avenues of Quranic analysis.

The Quranic Arabic Corpus is a dynamic resource. As research progresses, the annotations and functionalities may undergo changes. Nevertheless, its present configuration provides a robust set of tools that empower scholars to decipher the intricacies of the Quranic language and acquire a more profound comprehension of this revered literature.

Boosting Arabic Vocabulary with the Quranic Arabic Corpus

The Quranic Arabic Corpus (QAC) isn't just for advanced scholars; it can be a valuable resource for anyone learning Arabic vocabulary. Here's how:

- Rich Vocabulary Pool: The QAC offers access to the entire Quranic vocabulary, a vast collection of words. This exposes you to a wider range of vocabulary than a typical textbook, enriching your understanding of the language.
- Contextual Learning: Unlike a dictionary with isolated definitions, the QAC shows each word in its Quranic context. This allows you to see how the word is used in a sentence, aiding in memorization and grasping its nuances.
- Frequency Analysis: The QAC can display how often a word appears in the Quran. By focusing on frequently used words, you can prioritize learning the building blocks of Quranic Arabic.
- Grammar Reinforcement: Learning vocabulary in isolation isn't enough. The QAC's morphological annotations reveal a word's grammatical structure. This connection between vocabulary and grammar helps you not just memorize words, but understand how they function in sentences.
- Multiple Translations: The QAC incorporates various English translations of the Quran. This allows you to compare different interpretations and solidify your understanding of a word's meaning.

Example Use for Arabic Vocabulary Learning

The Quranic Arabic Corpus can be used to improve Arabic vocabulary through finding word collocations and finding various meanings that fit the vocabulary. Let's say you're a beginner Arabic learner and come across the word "Rahman" (الرحمن) which is often translated as "Most Merciful." Here's how the Quranic Arabic Corpus (QAC) can help you learn it:

1. Look it Up: Enter "الرحمن" in the QAC search bar.
2. Breakdown: The QAC will display the word's breakdown. You'll see it's a definite adjective (the prefix "ال" indicates definiteness).
3. Root and Meaning: The QAC will likely show the root of the word is "رحم" (rahima), which carries the meaning of mercy, compassion, or kinship. This helps you understand "Rahman" as emphasizing the vastness of that mercy.
4. Context: Now, the exciting part! Click on the word in the search result. This will show you every verse in the Quran where "Rahman" appears. See how it's used in different contexts. This reinforces the meaning and provides exposure to its various applications.
5. Frequency: The QAC might show that "Rahman" appears frequently throughout the Quran. This highlights its importance and encourages you to prioritize learning it.
6. Translations: The QAC might display how different translators render "Rahman" in English (Most Merciful, Most Compassionate). This broadens your understanding of the word's meaning.

Sketch Engine

Sketch Engine is a software program designed to crack the code of language. It acts as a powerful search engine and analysis tool for massive collections of text, known as corpora. These corpora can span billions of words, encompassing everything from news articles to historical documents. Imagine a library containing countless books in various languages. Sketch Engine allows you to not just search for specific words within these books, but also delve deeper into how those words are used.

Here's how Sketch Engine empowers researchers and language enthusiasts:

- Unveiling Usage Patterns: Beyond just finding where a word appears, Sketch Engine reveals how it's typically used. It shows surrounding words, grammatical forms, and sentence structures. This helps users understand the nuances of a word's meaning and how it functions in real-world contexts.
- Building Word Sketches: A central feature is the creation of "word sketches." These are one-page summaries that showcase a word's grammatical behavior and typical collocations (words that frequently appear alongside it). This provides a quick snapshot of a word's personality within a language.
- Multilingual Exploration: While some specialize in specific languages, Sketch Engine offers corpora in over 90 languages. This allows researchers to explore the unique way different languages handle vocabulary and grammar.
- Building Your Own Corpus: Not limited to pre-existing collections, Sketch Engine allows users to build their own corpora by uploading specific texts. This can be particularly valuable for researchers studying niche topics or analyzing specialized language.
- Applications Beyond Research: Sketch Engine isn't just for academics. Language learners can use it to see real-world examples of vocabulary and grammar in action. Translators can leverage it to ensure their translations capture the intended meaning and nuance.

In essence, Sketch Engine acts as a bridge between language theory and real-world usage. By analyzing massive amounts of text, it offers a window into the fascinating world of how languages work.

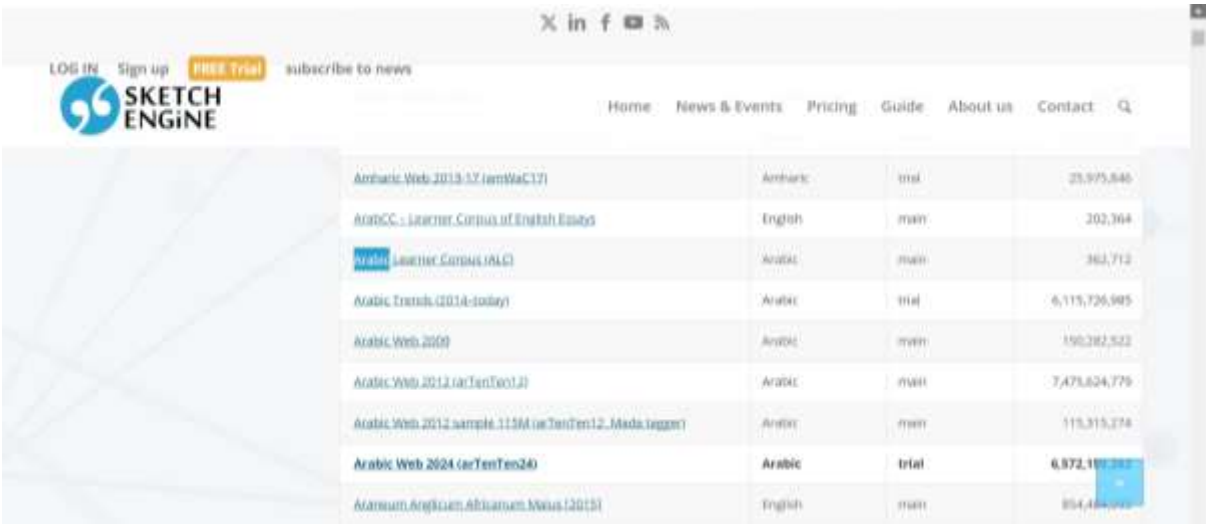

Figure 2 Sketch Engine Corpora

Use of Sketch Engine for Learning Arabic

Sketch Engine can be a valuable tool for anyone on their Arabic learning journey, from beginners to advanced students. Here's how it can supercharge your Arabic knowledge:

- Unveiling Word Nuances: Learning vocabulary through a dictionary alone can be limiting. Sketch Engine allows you to see how words are used in real-world contexts. Look up an Arabic word and see how it appears in sentences from news articles, novels, or even historical documents. This helps you grasp the subtle differences in meaning depending on the context.
- Collocation Mastery: Arabic, like any language, has words that often appear together. Sketch Engine's "collocation" feature shows you which words frequently accompany your target word. By learning these common pairings, you'll sound more natural and expand your vocabulary beyond isolated terms.
- Grammar in Action: Sketch Engine isn't just about vocabulary. You can see how words change their grammatical form depending on their role in a sentence. This visual reinforcement can solidify your understanding of Arabic grammar rules.
- Digging Deeper: Let's say you encounter a grammatical concept in your Arabic course. Use Sketch Engine to explore real-world examples of that concept. See how it's used in different contexts, solidifying your understanding beyond textbook explanations.
- Tailored Learning: Not everyone learns the same way. Sketch Engine allows you to build your own custom Arabic corpus. Upload texts you find interesting, like articles on your favorite topics, and analyze the vocabulary and grammar used within them. This personalized approach can make learning more engaging.

Please note that Sketch Engine provides a vast amount of information, however you may need to spend some time first exploring its features to navigate it effectively. Begin by conducting simple searches and progressively delve into its more advanced features. Although Sketch Engine is highly effective, it should not be used as a substitute for your main learning materials such as textbooks or courses. Utilize it in conjunction with them to acquire a more profound comprehension of the language in practice.

By integrating Sketch Engine into your Arabic learning regimen, you can acquire a more comprehensive comprehension of vocabulary, syntax, and the authentic functioning of

the language in real-life situations. It is an invaluable instrument for enhancing your proficiency in Arabic and developing greater confidence in speaking.

Example Use for Arabic Vocabulary Learning

Sketch Engine is used to improve Arabic vocabulary through finding word collocations and finding various meanings that fit the vocabulary. For example, it can be used to generate frequency lists of Arabic single-word or multi-word expressions of various types. Imagine you're learning Arabic and come across the word "كتابة" (Kitabat), which translates to "writing." Here's how Sketch Engine can help you solidify your understanding:

1. Search it Up: Type "كتابة" (Kitaba) in the Sketch Engine search bar.
2. Concordance Lines: Sketch Engine will display the word in context, showing full sentences from its vast Arabic corpora. Analyze these "concordance lines" to see how "كتابة" is used.
3. Collocations in Action: Look at the words surrounding "كتابة" in the concordance lines. You might see phrases like "عملية كتابة" (amaliyat kitabat - writing process) or " أسلوب كتابة" (usأسلوب كتابة - writing style). This helps you learn common collocations associated with "كتابة."
4. Grammar Check: Notice how the grammatical form of "كتابة" might change depending on its role in the sentence. It could be a noun (مصدر - مصدر) or an adjective (صفة - sifat) depending on the context. Sketch Engine helps you see these grammatical variations.
5. Digging Deeper: Let's say you want to explore formal vs. informal writing styles. Use Sketch Engine to compare how "كتابة" appears in formal news articles versus casual online chats. This can reveal subtle differences in word choice.

Conclusion

In conclusion, the use of The Quranic Arabic Corpus and Sketch Engine in this study demonstrated the potential of corpus-based approaches to enhance Arabic vocabulary acquisition for non-native learners. The corpus provided a rich and diverse range of texts, while Sketch Engine enabled learners to analyze the corpus in depth and identify patterns and trends in language use. The study highlights the importance of incorporating corpus linguistics into language learning materials, providing learners with a more authentic and engaging learning experience.

References

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International journal of corpus linguistics*, *11*(2), 135-171.

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, *49*, 721-751.

Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A., & Suchomel, V. (2014). arTenTen: Arabic corpus and word sketches. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 357-371.

Baisa, V., Suchomel, V., Kilgarriff, A., & Jakubíček, M. (2015). Sketch Engine for English Language Learning. *Corpus Linguistics 2015*, 33.

Buckwalter, T., & Parkinson, D. (2014). *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Hu, C., & Yang, B. (2015). Using Sketch Engine to investigate synonymous verbs. *International Journal of English Linguistics*, *5*(4), 29.

Jawharah, Alasmari., Janet, C.E., Watson., Eric, Atwell. (2017). Using the Quranic Arabic Corpus for comparative analysis of the Arabic and English verb systems.

Ma, Q., & Mei, F. (2021). Review of corpus tools for vocabulary teaching and learning. *Journal of China Computer-Assisted Language Learning*, *1*(1), 177-190.

Muhammad, Lukman, Arifianto. (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, doi: 10.18502/KSS.V5I3.8563

Nur, Ali., Mamluatul, Hasanah., Agung, Prasetyo. (2020). The Integration Of Qur'an And Linguistic Education Based On Ontology Of Qur'anic Concept In Quranic Arabic Corpus. doi: 10.18860/IJAZARABI.V3I2.9769

Paker, T., & Ergül Özcan, Y. (2017). The effectiveness of using corpus-based materials in vocabulary teaching. *International Journal of Language Academy*.

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.

Puspita, D., & Yusuf, K. (2023). Categorizing obsolete, archaic, and classic words in an Indonesian dictionary. *Lexicography*, *10*(1).

Siddiq, M., Arif, I. M. Q., Shafi, S. C., & Masood, M. H. (2021). A survey research analysis of effectiveness of vocabulary learning through English vocabulary corpus. *International Journal of Education and Pedagogy*, *3*(2), 1-13.

Van Mol, M. (2000). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In *Proceedings of the ninth EURALEX International Congress, Stuttgart* (pp. 831-836).

Vojtěch, Kovář., Vít, Baisa., Miloš, Jakubíček. (2016). Sketch Engine for Bilingual Lexicography. International Journal of Lexicography, doi: 10.1093/IJL/ECW029

Yusuf, K. (2020, May). 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.

Zerrin, ERDİNÇ. (2023). The Role of Sketch engine in the Compilation of News English Learning Dictionary. doi: 10.1007/978-981-99-1428-9_228