OPTIMIZATION OF CORPUS LINGUISTICS IN THE PREPARATION OF ARABIC IDIOM DICTIONARY

Nafisatul Izza R.U., Mohammad Ahsanuddin, and Hanik Mahliatussikah

Universitas Negeri Malang, Indonesia
nafisatul.izza.2302318@students.um.ac.id;
mohammad.ahsanuddin.fs@um.ac.id;
hanik.mahliatussikah.fs@um.ac.id

Abstract: Arabic, as one of the world's international languages, is rich in vocabulary and idiomatic expressions, which often pose challenges for foreign learners in grasping accurate meanings. Observations and interviews with students of the Arabic Language Education Study Program at the State University of Malang revealed a key issue in *maharah kalam* and *kitabah* courses: the lack of learning resources that systematically present idioms with contextual meanings and authentic usage. Existing dictionaries remain largely conventional, focusing solely on lexical translation, thereby creating a gap between vocabulary mastery and the ability to apply idiomatic expressions in a communicative context. To address this, the present study develops an Arabic–Indonesian idiom dictionary by integrating corpus linguistics as the primary approach in identifying, analysing, and systematically presenting idioms. Employing a qualitative-descriptive method, authentic corpus data were processed using Sketch Engine, which features include Wordlist, N-grams, Concordance, and Word Sketch. This analysis revealed the frequency of occurrence, structural variations, and both literal and idiomatic meanings. The study successfully identified 2,100 Arabic idioms considered relevant for learning purposes. This research not only provides a systematic framework for dictionary compilation but also produces an interactive digital product that is practical and learner-oriented. The novelty lies in combining corpus linguistics with digital lexicography, offering significant contributions to Arabic lexicographic studies while advancing technology-based, digital Arabic learning practices responsive to the demands of the digital era.

Keywords: Corpus linguistics, Arabic idiom dictionary, State University of Malang

1. INTRODUCTION

Arabic is one of the world's languages with a vast vocabulary and a rich array of expressions, including idioms or idiomatic expressions that can be challenging for foreign language learners to comprehend (Hidayat et al., 2024). The main problem faced by Arabic as a foreign language learners, especially students of the State University of Malang, based on the results of observations and interviews with students in the Arabic Language Education Study Program, is the limited learning resources that specifically present Arabic idioms with

contextual meanings and their use in real situations in learning *maharah kalam* and *kitabah*. Most dictionaries available are still conventional, focusing solely on lexical translations, which makes it difficult for learners to understand the precise meaning of idioms. As a result, a gap exists between the mastery of basic vocabulary and the ability to use idiomatic expressions effectively in communication.

To overcome these problems, the utilisation of Corpus Linguistics offers a more comprehensive and data-driven solution based on real-world data (Meyer, 2023). Corpus linguistics enables researchers to collect, analyse, and present Arabic idioms based on their frequency of occurrence, variation in form, and context of use in authentic texts (R.U. et al., 2025). Thus, the preparation of idiom dictionaries does not rely only on intuition or limited references, but is based on empirical evidence from actual language data (Holes, 2020). The optimisation of this approach is expected to be able to produce a digital dictionary that is more practical, interactive, and relevant to the needs of Arabic learners in the digital era.

Research Maura Syafa'ah & Hizbullah (2023). It demonstrates that the linguistic method of the corpus has been successfully optimised for the preparation of thematic bilingual dictionaries, such as those in the field of transportation, by systematically collecting hundreds of vocabulary items based on authentic data. However, this approach has not been specifically applied to idioms, so learners' needs in understanding more complex idiomatic expressions have not been addressed. Research by Jarad & Saydeh (2017) found that many Arabic idioms are not systematically covered in bilingual dictionaries, and that the lemmas of existing idioms are often inconsistent and subjective in their determination. This signals the need for more objective and data-driven methods, such as the linguistic corpus, to improve the scope and consistency of idiom entries. Research by Althobaiti (2022) emphasises the importance of validating and collecting idiomatic data from diverse sources, particularly Qatari dialects, to establish a representative idiomatic corpus. However, the application of the corpus for preparing an idiom dictionary that is both applicable and easily accessible to learners remains very limited. Thus, a research gap exists that needs to be bridged, specifically the optimisation of corpus linguistics for the preparation of Arabic idiom dictionaries presented in a systematic, consistent, and applicable manner to support learning needs.

The novelty of this research lies in the integrative approach between corpus linguistics and the development of digital idiom dictionaries oriented to learners of Arabic as a Foreign Language (AFL). With these innovations, this research is expected to make a significant contribution to both Arabic lexicography studies and modern Arabic learning practices that are technology-based and in line with contemporary demands. The primary objective of this research is to describe the process of compiling Arabic-Indonesian idiomatic dictionaries, with an emphasis on both literal and idiomatic meanings, and their application in authentic educational communication contexts.

2. METHOD

This research employs a qualitative descriptive approach in field research (Sugiyono, 2018). The primary objective of this study is to describe the preparation of a dictionary of Arabic-Indonesian bilingual idioms, utilising corpus linguistics as a medium to support the

learning of maharah kalam and kitabah for students in the Arabic Language Education Study Program at the State University of Malang. This research does not focus on the presentation of quantitative data, but rather emphasises the role of researchers as the main instrument in the research process. The method used in the preparation of this dictionary (lexicography) refers to (Schierholz, 2015), which is based on theory (Wiegand, 1998). The preparation process is carried out through five systematic stages, namely: (1) preparation phase, (2) data and material collection, (3) data processing or processing, (4) evaluation and assessment of data and materials, and (5) preparation for publication or printing of dictionaries, as shown in the following figure:

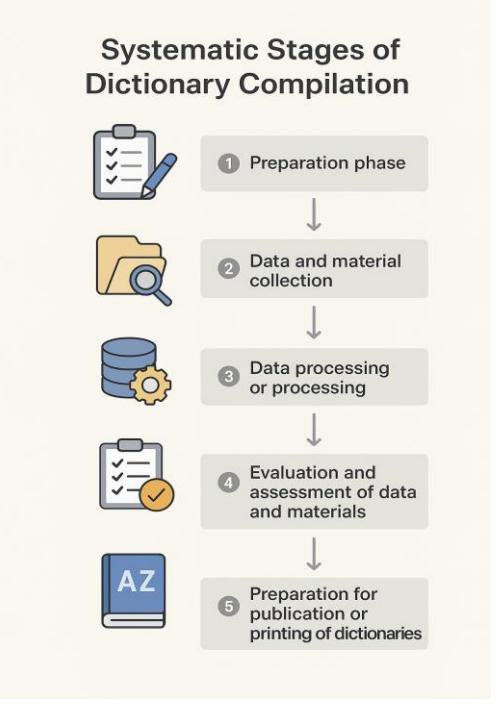

Figure 1.

3. FINDINGS AND RESULTS OF DISCUSSIONS

Data collection was conducted from a variety of sources, including 30 scientific articles on idioms in Arabic, as well as six e-books from contemporary digital sources such as Instagram accounts @kampungarab_pare, @uslubarabi.id, @ashablughah, @lughotunaalarabiyyah, and the Facebook Group "Arabic Idioms." To analyse the data, the researcher utilises the Sketch Engine platform, which is widely used in dictionary preparation due to its ability to facilitate the creation of corpora tailored to research needs (Abdumanapovna, 2019). In this study, the researcher compiled his own corpus, which was then analysed using the Sketch Engine. This platform enables the careful analysis of corpus data, yielding accurate and accountable results (Arruda et al., 2023). The corpus data is systematically selected, organised, and integrated into the product to ensure the accuracy,

completeness, and relevance of the developed idiom material. The types of idioms collected include a combination of isim and *harf, fi'il* and *harf,* as well as a combination of *fi'il, isim*, and *harf,* either in the form of full idioms, partial idioms, or idiomatic word pairs.

The method of compiling the dictionary, as formulated by the researcher, is based on various lexicographic theories proposed by Schierholz (2015) and Wiegand (1998). The main focus lies in the data collection and processing stage, where the researcher designs the process of compiling the dictionary with a corpus linguistic approach. Several steps taken include:

a. The collection of Arabic idioms from various sources, namely 30 scientific articles, six online e-books, as well as various contemporary digital sources such as educational Instagram accounts, linguistic websites, and Facebook platforms.
b. Data input and filtering in (doc) format by removing irrelevant editorial elements.
c. Convert files to Plain Text (*txt) with UTF-8 Unicode.
d. Data analysis based on corpus linguistics is carried out using the Sketch Engine through several stages, as follows:
    1) The initial stage uses the Wordlist feature to identify high-frequency words as the basis for determining idiom candidates.
    2) N-gram analysis is used to find possible hidden idiomatic phrases.
    3) The phrases are manually selected through the Concordance feature to observe the context in which they are used and ensure their idiomatic nature.
    4) The final stage leverages Word Sketch to explore the patterns of colocation and grammatical relationships that support accurate idiom identification. The results of the analysis of these stages produced a formatted file (*xls) that contains an alphabetical list of Arabic idioms.
e. Data cleansing with a focus on Arabic idioms that are more needed in conversation and writing
f. Revise the data and transfer it to the file (*doc) according to the template that has been prepared.
g. The dictionary draft in (*doc) format is ready to be searched.

In the process of data analysis, the researcher utilises four main features available in the Sketch Engine platform, namely:

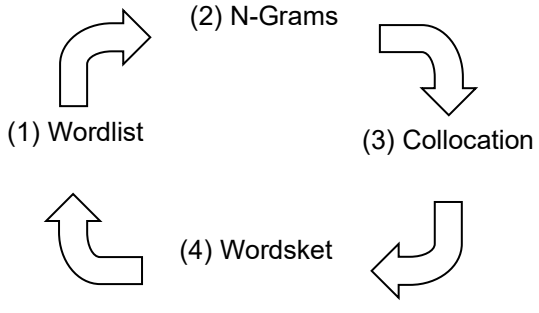

Figure 2.

1) Wordlist: Displays all the words in the text to identify high-frequency words.
2) N-grams: Identify combinations of words that often appear together in a specific order. This process aims to find phrases or word pairs that have the potential to be idioms.
3) Concordance: Verifying the context of usage in a sentence
4) Word Sketch: Strengthen idiom identification by analysing the colocation and grammatical structures that often accompany it.

From the initial process, 2,800 idiom candidates were obtained. The data is then exported in .xls format to undergo the cleanup stage. The focus of the cleansing is directed at idioms commonly used in oral and written contexts, particularly the combination of *fi'il, isim,* and *ḥarf,* with the dominance of partial idiom forms and idiomatic pairs. After the revision process, 2,100 idioms that are suitable for use were screened.
Here are some examples of idioms found :

Idiom	Means	Example	Meaning of Examples
أَثَّرَ فِي	Meninggalkan	كَلِمَاتُ الأُسْتَاذِ أَثَّرَتْ فِي نُفُوسِ الطُّلَّابِ	Kata-kata guru meninggalkan kesan dalam hati para siswa
أَثَّرَ عَلَى	Mempengaruhi	قِلَّةُ النَّوْمِ تُؤَثِّرُ عَلَى تَرْكِيزِ الطُّلَّابِ	Kurang tidur mempengaruhi konsentrasi siswa
أَخَذَ بِـ	Menolong	أَخَذَ الطُّلَّابُ بِأَيْدِي بَعْضِهِمْ لِلتَّعاوُنِ فِي إِنْجَازِ المَشارِيْعِ الجَمَاعِيَّةِ.	Para mahasiswa saling menolong dalam menyelesaikan proyek kelompok
أَخَذَ عَلَى	Mencela, menegur	أَخَذَ المُحَاضِرُ عَلَى الطُّلَّابِ تَأَخُّرَهُمْ المُتَكَرِّرَ عَنِ المُحَاضَرَاتِ	Dosen menegur mahasiswa atas keterlambatan mereka yang berulang dalam kuliah
أَخَذَ عَنْ	Belajar kepada, mencontoh	أَخَذَ الطُّلَّابُ عَنِ المُحَاضِرِ أُسْلوبَهُ فِي الكِتابَةِ العِلْمِيَّةِ.	Mahasiswa mencontoh gaya menulis ilmiah dari dosen mereka
أَخَذَ فِي	Memerhatikan	أَخَذَ المُعَلِّمُ فِي تَقْيِيمِ أَدَاءِ الطُّلَّابِ خِلالَ الحِصَصِ الدِّرَاسِيَّةِ	Guru mulai memerhatikan dan mengevaluasi kinerja siswa selama kelas

4. CONCLUSION

The results of this study demonstrate that corpus linguistics can be employed as an effective approach and methodological basis in the preparation of Arabic-Indonesian bilingual idiomatic dictionaries. The application of this approach has proven to be significant in helping compilers map and group idioms more systematically through the use of the *Sketch Engine*

platform, which features four main components: Wordlist, *N-grams*, *Concordance*, and *Word Sketch*.

The initial research process produced 2,800 idiom candidates. After a rigorous revision and selection stage, 2,100 idioms were identified as worthy of inclusion in the dictionary, taking into account a combination of idioms based on *isim, fi'il,* and *ḥarf* that were relevant to the context of learning Arabic. The research findings also confirm that three main aspects determine the success of the preparation of a corpus linguistics-based idiom dictionary: (1) accuracy in the selection of representative data sources, (2) data processing mechanisms through valid corpus analysis tools, and (3) the competence of researchers in selecting vocabulary or lemmas to be included in the dictionary draft. These three aspects are the primary prerequisites for producing bilingual dictionaries that are not only linguistically accurate but also applicable in supporting the needs of Arabic language learning in the realm of modern education.

5. SUGGESTIONS

Although the preparation of an Arabic idiomatic dictionary based on corpus linguistics in this study has been successfully implemented, the researcher recognises that several limitations still need to be addressed. The dictionary products produced at this stage cannot be considered completely perfect, so they still need further development. Therefore, advanced research is very open to being carried out, either through methodological modifications, the addition of interactive features, or the enhancement of idiomatic content. With this follow-up, it is hoped that corpus-based bilingual idiom dictionaries can be more representative and applicative, as well as make a more significant contribution to the development of modern, adaptive, and technology-based Arabic language learning.

6. BIBLIOGRAPHY

Abdumanapovna, S. A. (2019). The Role of Sketch Engine in Multiple Types of Corpora. *International Journal of Innovative Technology and Exploring Engineering*, *8*(11), 250–254. https://doi.org/10.35940/ijitee.K1307.0981119

Althobaiti, M. J. (2022). A Simple Yet Robust Algorithm for Automatic Extraction of Parallel Sentences: A Case Study on Arabic-English Wikipedia Articles. *IEEE Access*, *10*, 401–420. https://doi.org/10.1109/ACCESS.2021.3137830

Arruda, H. M., Bavaresco, R. S., Kunst, R., Bugs, E. F., Pesenti, G. C., & Barbosa, J. L. V. (2023). Data Science Methods and Tools for Industry 4.0: A Systematic Literature Review and Taxonomy. *Sensors*, *23*(11), 5010. https://doi.org/10.3390/s23115010

Hidayat, R., Sulaimah Saleh, U., Satya, I., & Wargadinata, W. (2024). Idiomatic Phrase Processing in Arabic: A Psycholinguistic Study." *International Journal of Language and Ubiquitous Learning*, *1*(3), 209–221. https://doi.org/10.70177/ijlul.v1i3.668

Holes, C. (2020). Arabic corpus linguistics ed. by Tony McEnery, Andrew Hardie, and Nagwa Younis. *Language*, *96*(1), 202–206. https://doi.org/10.1353/lan.2020.0007

Jarad, N. I., & SSaydeh, A. (2017). Idiom In The Arabic - English Dictionary. *International Journal of Arabic-English Studies (IJAES)*, *7*.

Maura Syafa'ah, D., & Hizbullah, N. (2023). Compiling an Arabic-English Transport Dictionary through Corpus Linguistic Methods. *Scaffolding: Jurnal Pendidikan Islam Dan Multikulturalisme*, *5*(1), 251–270. https://doi.org/10.37680/scaffolding.v5i1.2546

Meyer, C. F. (2023). *English Corpus Linguistics*. Cambridge University Press. https://doi.org/10.1017/9781107298026

R.U., N. I., Mahliatussikah, H., & Ahsanuddin, M. (2025). Students' Perception of the Development of a Digital Dictionary of Arabic Idioms Based on Corpus Linguistics. *Arabiyat : Jurnal Pendidikan Bahasa Arab Dan Kebahasaaraban*, *11*(2), 234–244. https://doi.org/10.15408/a.v11i2.41665

Schierholz, S. J. (2015). Methods in Lexicography and Dictionary Research. *Lexikos*, *25*. https://doi.org/10.5788/25-1-1302

Sugiyono. (2018). *Metodologi Penelitian Kuantitatif, Kualitatif dan R&D*. Alfabeta.

Wiegand, H. E. (1998). *Wörterbuchforschung*. DE GRUYTER. https://doi.org/10.1515/9783110802467