Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 3, 2020, pp 920-927
https://doi.org/10.18510/hssr.2020.8396

INVESTIGATING ARABIC CORPUS (KorSA) OF INDONESIAN UNDERGRADUATE THESIS ABSTRACTS

Mohammad Ahsanuddin[1*], Ali Ma'sum[2], Nur Anisah Ridwan[3]
[1*,2,3]Department of Arabic Literature, Faculty of Letters, Universitas Negeri Malang, Indonesia.
Email: [1*]mohammad.ahsanuddin.fs@um.ac.id, [2]ali.masum.fs@um.ac.id, [3]nur.anisah.fs@um.ac.id

Article History: Received on 28th March 2020, Revised on 18th May 2020, Published on 17th June 2020

Abstract

Purpose: This study was designed to unveil Arabic corpus written in the Indonesian undergraduate thesis abstracts.

Methodology: Experimental and descriptive approaches were employed to elicit data which were in the forms of *isim* (noun), *fi'il* (verb), concordances, and idioms collected from 59 thesis abstracts written by undergraduate students of Universitas Negeri Malang UIN Maulana Malik Ibrahim Malang, both are state-owned universities based in East Java, Indonesia.

Principal Findings: The results of this study informed that two steps were carried out to craft an Arabic corpus for undergraduate theses, such as enacting need analysis and designing the model of Arabic corpus. Furthermore, this study also uncovered that the main page of the corpus website was concordances and word frequency.

Implications/Applications: If in the context of translating the Qur'an into English only, for example, there is not yet a parallel corpus model of the Qur'an. Furthermore, similar models in other languages, such as the Indonesian context are possible.

Novelty/Originality of this study: This research is the first step towards the formation of the first model of the parallel Qur'an corpus and its translation in the Indonesian language. Besides, this model can later be used as a reference for the preparation of other identical corpus models in the context of bilingual bodies or more for the benefit of translation research.

Keywords: *Corpus, Arabic Language, Undergraduate Thesis, Concordance, Word Frequency, Indonesian Language.*

INTRODUCTION

Corpus linguistics is an applied linguistic approach widely used to analyze languages recently (Anthony, 2013). Biber (1988) contended that corpus encompasses four main distinctiveness, namely: (1) based on an empirical approach (experiment-based) in which the patterns of language use observed in original language texts (oral and written) can be analyzed, (2) the linguistic corpus uses a representative sample of the target language stored as electronic database (corpus) as a basis for analysis, (3) it relies on computer software to calculate linguistic patterns as part of the analysis, and (4) it depends on quantitative and qualitative analysis techniques to interpret findings (Biber, 1988; Alfaifi & Atwell, 2016).

According to Baker (2010), the corpus is a collection of texts, both oral and oral writings stored on a computer (Baker, 2010). Baker (2010) argues that a corpus is found only in electronic media. On the other hand, according to Setiawan, the corpus is a collection of writings written by someone both in the forms of hard copy and soft copy. Corpus in the form of hard copy can be exemplified, such as books, magazines, dictionaries, and newspapers. Examples of the soft copy are in the types of applications, websites, online dictionaries, and so forth (Setiawan, 2018).

From this understanding, it can be concluded that the corpus is a collection of texts, both oral and written, in print and electronic media and can be used as a source of data (Ahsanuddin, 2018). In this case, all types of linguistic units, such as words, phrases, clauses, sentences, and discourses, are part of the corpus which are collected into one unified form. As a result, the corpus is different from the collection. Therefore, the data called corpus is also identical to a large amount of data. One type of corpus is the Arabic thesis abstract.

The abstract is a summary of the contents of a scientific paper intended to help a reader to be able to easily and quickly see the purpose of the writing. In the academic world, this abstract is used by educational institutions and organizations as initial information on a study when it is included in journals, conferences, workshops, etc. In cyberspace (internet) context, an abstract is used as a brief description of a scientific paper.

An abstract is a concise and precise representation of the contents of a document that includes the original text, and usually follows the style and arrangement as in the original text. The purpose of the abstract is to capture the contents of an essential document so that in a short time, the reader can find out the information contained in the report.

Conciseness and significance are two essential concepts in abstracting. The abstract must be written, precisely, comprehensively, not independent (independent), and not intended to provide a critique of the document. The writer is the person who knows the contents of the text best so that he is expected to be able to choose the most essential parts of the text that must be written in the abstract. An extractor is a person who has adequate knowledge of the subject of the text to be abstracted and understands the method of making abstract.

As a miniature document, the abstract functions as a guide to the contents of the document. By reading the abstract the reader can find out the scope of the contents of the document in a relatively short time, so the reader can later decide whether the document is relevant or not as desired, and can decide whether the reader needs or not to read the full paper. Also, abstracts allow readers to read large amounts of literature. This is very useful to avoid duplication in research and development.

Previous works underpinning this study is Sasongko's (2010) research entitled application to build a corpus from crawling results with various data formats automatically. The results of the study unfold that the application can display the results of searching documents and sort them according to the order of the discovery of the data file sought, in the sense that the document data found will first be placed while the data documents found will be placed in the lowest order. The application is also able to convert the text of the documents with various data formats into *txt* text documents, as well as to convert all image file formats into *BMP* format. Conversion is done to equalize the format to facilitate the storage in the database (Atwell, Al-Sulaiti, Al-Osaimi, & Abu Shawar, 2004; Sasongko, 2010; Evert, 2009A).

Previous research carried out by Ahsanuddin (2018) entitled *Tashmim Al-mudawwanah Al-Mutawaziyah Li mustakhlash Al-buhuts Al-Ilmiah Al-Indunisiya Al-Arabiya 'AlaDhauiNadzariyah Mona Baker Li Al-Takafu' Al-Lughawi Fi Al-Tarjamah* is also a reference for this study. The results of the study showed that (1) Indonesian-Arabic abstract parallel research corpus products were made. Before making a parallel corpus, several steps were done: (a) equivalence analysis of the dissertation abstract translation in the Mona Baker perspective with three levels, namely, the word, grammatical, and text levels. The level equivalence of the words studied is the translation of abstract words and keywords. Grammatical level equivalents include numbers (*adad*), pronouns (*dhomir*), personal (*syakhsyiah*), and verbs (*af'al*). The types of *adad* in the dissertation abstract consist of *mutsana*, *mufrad-jama*, *jama-mufrad* and *jama'-jama*. *Dhomir* in the abstract is a third-person pronoun (*dhomirghiyab*). Equivalence level translation of texts in the dissertation abstract covers references (*ihalah*), conjunctions (*adawatrabth*), cohesion (*al-ittisaq al-mu'jami*). The references found in abstract texts are personal references and gesture references, while the most commonly used conjunctions are *wawu* conjunctions and the cohesion that appears is repetition (*tikrar*), (b) parallel corpus design construction. To determine the parallel corpus, researchers examined several previous studies related to the corpus. After the preliminary study, the researchers then designed a parallel corpus using the Dreamweaver program in PHP.

The design contained a corpus and also functions as a search engine, vocabulary list, Concordance, and word frequency. When expert validation was done, the researcher obtained 85% validation value, which means that the parallel corpus is valid for use. Besides, the utilization of the parallel corpus abstract of Indonesian-Arabic research. The benefits obtained from this parallel corpus are as a search engine, vocabulary list, Concordance, and word frequency. Third, the satisfaction of parallel corpus users. This corpus is effective for finding previous research and for translating material (Ahsanuddin, 2018; Evert, 2009B).

Purpose of the study

The main objective of this study is to design a corpus for Arabic to Indonesian language translation. In this paper, researchers developed an abstract corpus in Arabic thesis as a part of a larger research project. The corpus model developed is related to word frequency and Concordance.

LITERATURE REVIEW

Corpus

The corpus is a collection of several texts as sources of language and literary research (McEnery & Wilson, 1996). A collection of texts is called a corpus provided that the collection of texts is used as an object of literary and literary research (Kilgarriff & Grefenstette, 2003; Bunchutrakun, Lieungnapar, Wangsomchok, &Aeka, 2016; Veerachaisantikul & Chansin, 2018).

The corpus was built to require considerable energy. The corpus has a large number of text document members, up to millions of documents. The collection of millions of documents takes a long time. The texts in the corpus are arranged systematically to facilitate management (Wagner& Nesselhauf, 2006). Several aspects to consider in building a corpus are as follows: a) planning and design of the corpus, b) selection of data sources, c) permission from the data owner, d) data collection and coding, and e) handling of the corpus.

The purpose of the stages in building a corpus is the design of the corpus following linguistics and project and administrative costs. Some things related to linguistics that need to be considered in building a corpus are a) the size of the text to be sampled and b) the range of linguistic diversity (synchronous) and the period of the text (diachronic) for the sample material (Biber & Jones,2009).

This research is in the realm of studies in the field of corpus linguistics. The definition of corpus linguistics is an empirical method carried out in the analysis and description of linguistics to examine the real linguistic phenomena used by speakers of languages. The language is arranged systematically based on certain categories and is researched to

elaborate on the real meaning of the use of the language (Cheng, 2012; Veerachaisantikul&Chootarut, 2016; Madina, Sholpan, Zhanar, Bekzhan, &Kuandy, 2017; O'Mahony, 2018).

Many recent Arabic studies use this approach. In particular, there are already several digital Al-Quran corpus models, both available in text form and can be downloaded for processing by researchers and corpus available on the web that also provide certain data search facilities. One of the most popular models of the corpus of the Koran is www.corpus.quran.com. This site provides digital Al-Quran data with systematic mapping or taxonomy of core concepts in the Qur'an. Also, this site provides grammatical analysis facilities of the Qur'an along with the syntactical taxonomy of the Arabic Qur'an. However, this site does not yet specifically provide digital Al-Quran files which can be utilized for further linguistic research and analysis.

Different from the general type of corpus, the parallel corpus contains data in two or more languages arranged side by side or side by side. When referring to the Sketch Engine as one of the providers of Arabic corpus material as well as the Koran, there has not been found any parallel Qur'an corpus. Only a new annotated corpus of the Qur'an without translating the text. According to Eddakrouri (2016), there are also five corpora of the Qur'an but all of them still contain material for the approach of analysis of co-conduct and the concept map of the Koran, there is no parallel corpus of the Koran at all (Eddakrouri, 2016).

If in the context of translating the Qur'an into English only, for example, there is not yet a parallel corpus model of the Qur'an. Furthermore, similar models in other languages such as the Indonesian context are possible. Therefore, this research is the first step towards the formation of the first model of the parallel Qur'an corpus and its translation in the Indonesian language. Besides, this model can later be used as a reference for the preparation of other parallel corpus models in the context of bilingual bodies or more for the benefit of translation research.

RESEARCH METHOD

This study employed a combination of experimental and descriptive methods. The experimental method was used to test the creation of an abstract Arabic corpus using a procedure provided by the corpus processing application. The description was used to explain the stages, procedures, and mechanisms carried out in the series of corpus manufacturing processes from beginning to end.

The research data were in the form of Arabic words in the form of *isim* (noun), *will* (verb), and letters in the form of concordances and idioms. The Source of the data was a thesis abstract, amounting to 59 theses from Universitas Negeri Malang and UIN Maulana Malik Ibrahim Malang, both are Indonesian state-owned universities. The instrument used in this study was a thesis abstract document. The data were, as the analysis method, included in the web and will be analyzed by the web corpus based on word frequency and Concordance.

RESULTS AND DISCUSSION

Results

To produce a product in the form of a web corpus, researchers carried out certain stages. The stages in question were (1) reviewing various literature or reference books related to the linguistic corpus, (2) collecting thesis from Malang State University and UIN Maulana Malik Ibrahim Malang, (3) designing and developing corpus web using *PHP* program, (4) inserting thesis material (abstract) into the corpus website, (5), and (6) running the website. The product of this research is a corpus web that has been online as KorSA (Arabic Thesis Corpus).

Before developing a corpus model in Arabic, the researchers first collected data in the form of an undergraduate thesis from Arabic Education Study Program, Universitas Negeri Malang, and UIN Maulana Malik Ibrahim Malang.

Table 1: Undergraduate Thesis Abstracts from Universitas Negeri Malang and UIN Maulana Malik Ibrahim Malang

No	Year	Number
1.	2019	6
2.	2018	22
3.	2017	15
4.	206	16
	Total	59

From the 59 abstracts, the researchers then entered the data into the corpus website. The intended web address is http://m-ahsanuddin.net/korsa. In making the corpus, researchers used various programs namely xampp server as apache server and Dreamweaver.

Apache HTTP Server or Apache Web Server/WWW is a web server that can run on many operating systems that are useful for serving and functioning web sites. The protocol used to service the web/www facility uses HTTP. The most widely used languages are PHP, HTML, asp, and others.

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 3, 2020, pp 920-927
https://doi.org/10.18510/hssr.2020.8396

The Dreamweaver program is used to design the corpus web in the form of concordance and word frequency using PHP. After all, the material is made through offline then researchers upload data online. The results of the appearance of the corpus web are as follows:

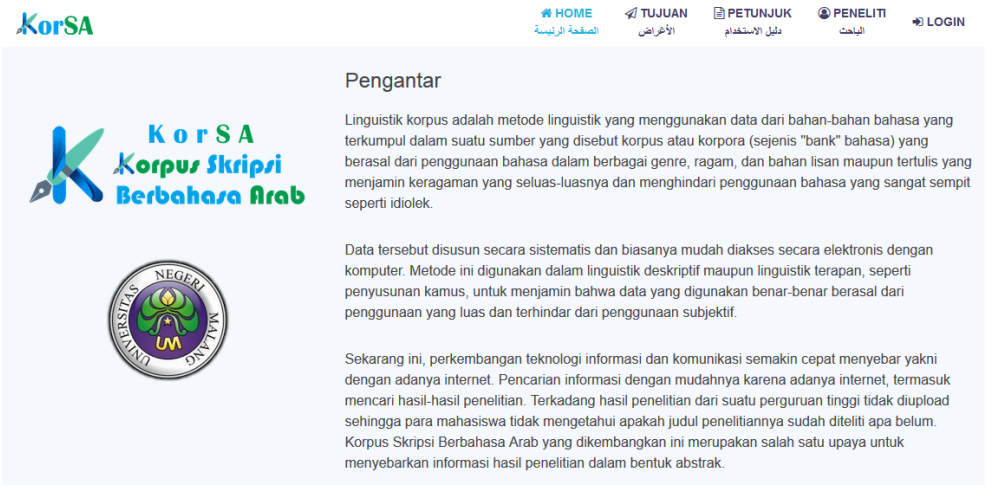

Figure 1: Homepage

Source: Developed by the author on KORSA

In this page, the introduction of the corpus web the goal is presented so that the reader or user of the corpus web understands the corpus. The next page is the destination. It is expected that the corpus web will provide information to users, both students and researchers, to further develop this research.

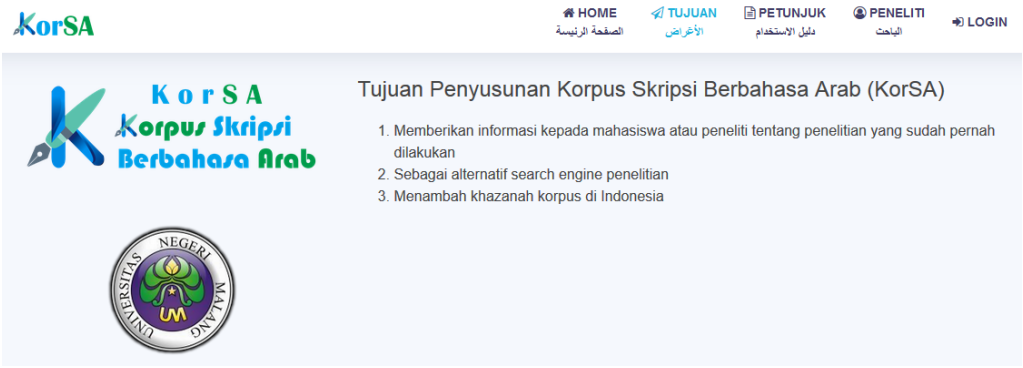

Figure 2: Goals

Source: Developed by the author on KORSA

Section goals on this page are central to inform users about the function of KorSA.

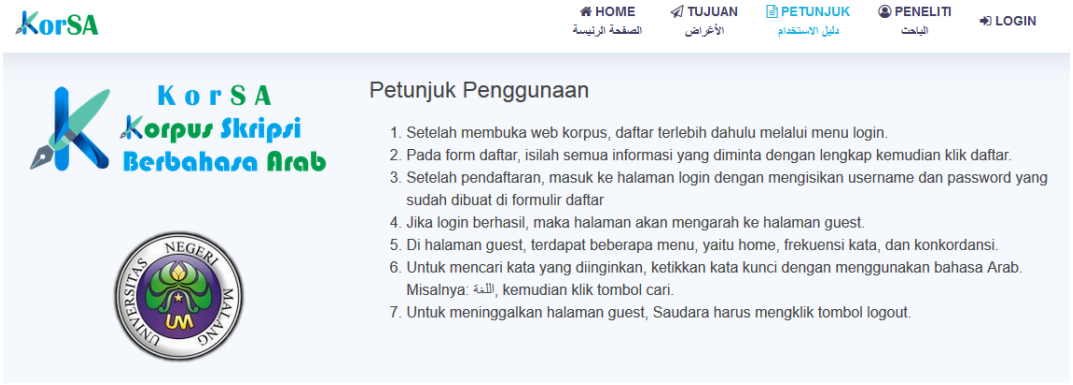

Figure 3: User Instruction

Source: Developed by the author on KORSA

This section provides information for using the corpus.

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 3, 2020, pp 920-927
https://doi.org/10.18510/hssr.2020.8396

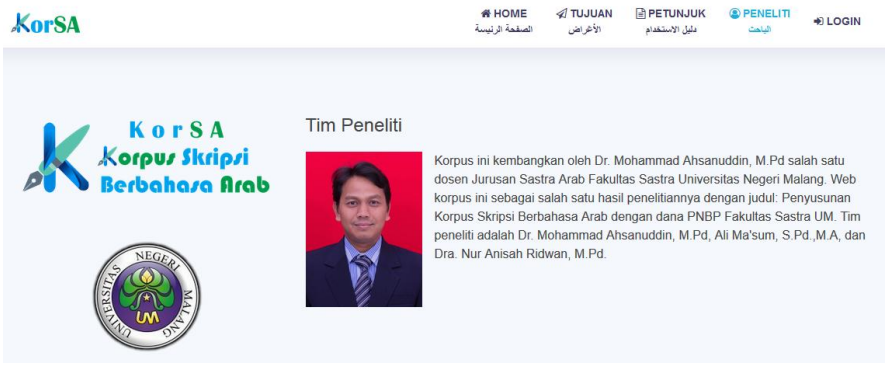

Figure 4: Researchers

Source: Developed by the author on KORSA

Researchers working on developing this corpus are also introduced on the website.

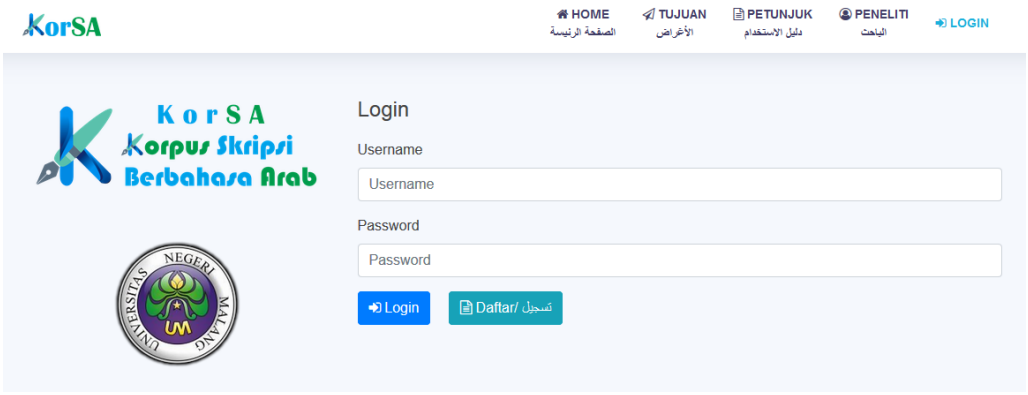

Figure 5: Login

Source: Developed by the author on KORSA

The last page is login. This page has two parts namely login as a guest (guest) and admin. Login as a guest can take advantage of the web corpus in the form of Concordance, and word frequency. The login as admin is used to enter thesis data. The information entered is the name of the researcher, year, and abstract content.

a. Word Frequency

Word frequency is calculated based on the appearance of the word in the corpus. Words in the corpus are then sorted by frequency. An example in this corpus web is the word "المدرسة" (Madrasah) appears 24 times in the corpus.

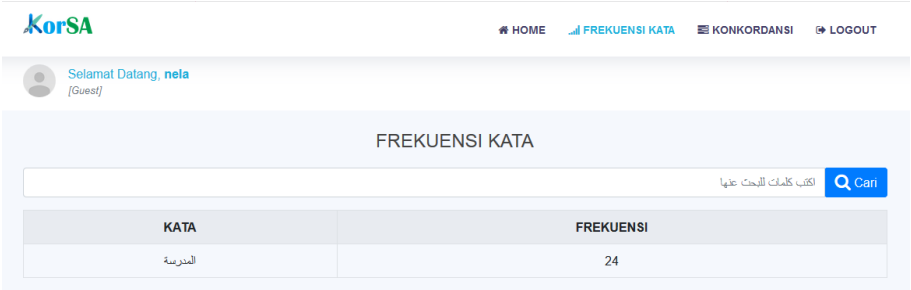

Figure 6: Word Frequency of "المدرسة" (Madrasah)

Source: Developed by the author on KORSA

b. Concordances

A concordance is a collection of the appearance of word form, in their respective textual environment. The simplest form of Concordance is the index. Each word formation is indexed and references refer to the scene in a text. Sinclair's definition is important because it reminds us that, at first Concordance is a list of manually prepared words found in the text or a collection of texts along with references and their location in the text. In the linguistic analysis using computer assistance, Concordance remains in the form of an index but can be generated for various new purposes and in various

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 3, 2020, pp 920-927
https://doi.org/10.18510/hssr.2020.8396

types of texts and texts that continue to develop. Computers that can now be used to make concordances in the blink of an eye and can support a wider range of analytic objectives and not experience the limitations of places like concordances in earlier times.

Before the days of digitizing texts with modern computers like today, concordances were made by dedicated individuals or teams over a long period. Team members will read the text, identify words that are important for analysis, and build a hard-working table that allows one to record where each sample of words is found. Examples of concordances on the corpus web such as the word "المدرسة" (Madrasah) are displayed in Figure 7.

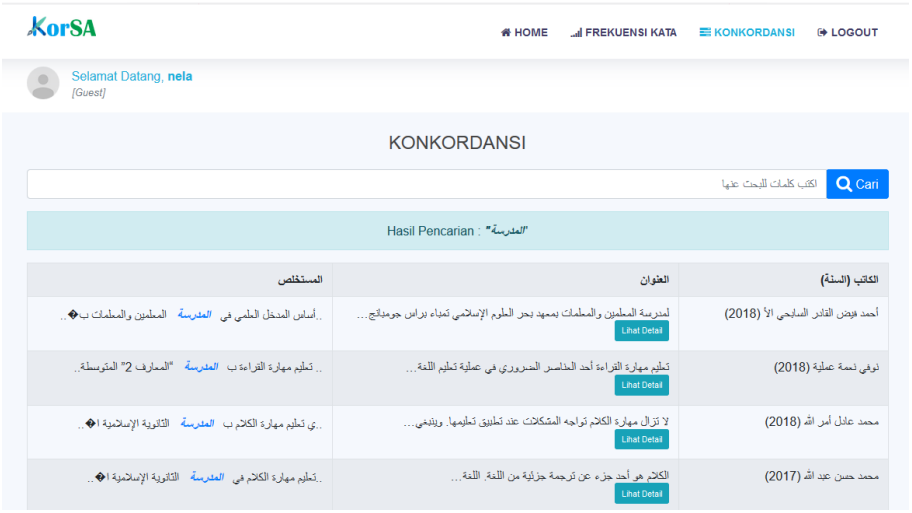

Figure 7: Concordances Page of "المدرسة" (Madrasah)

Source: Developed by the author on KORSA

DISCUSSION

The corpus of linguistics is defined as the study of corporate compilation and analysis (Cheng, 2012). According to McEnery and Hardie(2011) defines the linguistic corpus as a field that focuses on procedures, or methods of studying, or researching language (McEnery & Hardie, 2011). McEnery and Hardie (2011) also alluded to the approach used in corpus linguistics that was also put forward by Tognini-Bonelli (2001). Tognini-Bonelli (2001) states that there are two corpus-based linguistic approaches, namely the corpus-based and the corpus-driven (Tognini-Bonelli, 2001; Wiechmann, 2008). Both have differences in seeing the corpus as evidence that supports the theory. The first uses a deductive approach. Meanwhile, the corpus-driven approach considers the corpus as evidence that must be a theoretical reference, so that it is inductive.

KorSA application development was carried out through several stages. The first stage is gathering requirements (requirements gathering). In the first stage, all the needs, information, and initial data needed for application development are explored and collected. In addition to those directly related to application development, this first stage also determines the Arabic corpus design of the Arabic thesis which will be the content in the KorSA application. In terms of the internal structure of the corpus, in general, Arabic data collected consists of written data. The KorSA application has several features, which are features that can be accessed without the user logging into the application. Examples of features that can be accessed by users without logging in are features for the main view, purpose, usage instructions, and developer profile. While the feature accessed by using the login is to display the search results of words from the corpus in the form of a concordance and the feature to display a list of word frequencies from the corpus.

Concordance is the mainstay in the corpus. Concordance allows strings and related words (McEnery, 1997). According to Setyawan(2018a, 2018b) that Concordance (Concordance) is a collection of the appearance of word formations, in their respective textual environment. The simplest form of Concordance is an index. Each word formation is indexed and references refer to the scene in a text (Setyawan, 2018a, 2018b). The most commonly used format for displaying word formations is Key Word in Context (KWIC), which displays the word you are looking for with several characters before and after the word appears. Like the word 'المدرسة' (Madrasah) in KorSA. The first ten occurrences of word formations are presented in the form of text order in the middle of a context of seventy characters (spaces and punctuation are counted as characters).

Regarding word frequency, one of the steps in text processing in the field of text information retrieval is text cleansing of irrelevant words used as an index. In a text document, there can be many types of words such as prepositions, conjunctions, pronouns, adjectives, and so forth. Some of these words may not potentially be used as index documents because their appearance is not unique or has never been used in a search query. For this reason, the process of filtering out the words is carried out (Luhn, 1960) dan (Flood, 1999). Filtering is done by providing a list of words that are not

important to be indexed (stopword list). Zipf's(1949) Law is sometimes used as a basis for forming stop lists, mainly in analyzing the appearance of words (Zipf, 1949).

The purpose of word frequency in addition to obtaining data on how many times the word appears on the corpus web can also be used to analyze the word. Setyawan (2018a), (2018b) mentioned that the benefit of word frequency is to understand each unit in the text (Setyawan, 2018a, 2018b). Each text has several levels of meaning, and this level tends to be related to physical, structural units, starting from a single word, phrase, clause, sentence, throughout the whole text. One of the fundamental problems faced when processing text is the question of what a word is. We might naively argue that words are entities in text separated by spaces or punctuation. The definition of words like this largely ignores the fact that in practice words do not consist of only one entity that is only limited by spaces or punctuation.

CONCLUSION

The results of this study were twofold. To design an Arabic corpus, several stages were carried out, namely, need analysis and the corpus construction, named as KorSA. Besides, the main page on the KorSA web corpus is the concordance and word frequency. The results of this study informed that two steps were carried out to craft an Arabic corpus for undergraduate theses, such as enacting need analysis and designing the model of Arabic corpus. Furthermore, this study also uncovered that the main page of the corpus website was concordances and word frequency.

If in the context of translating the Qur'an into English only, for example, there is not yet a parallel corpus model of the Qur'an. Furthermore, similar models in other languages such as the Indonesian context are possible. Therefore, this research is the first step towards the formation of the first model of the parallel Qur'an corpus and its translation in the Indonesian language. Besides, this model can later be used as a reference for the preparation of other identical corpus models in the context of bilingual bodies or more for the benefit of translation research.

LIMITATION AND STUDY FORWARD

This study has some limitations which must be addressed in the future to make the findings more reliable. The study used 59 theses from Indonesian state-owned universities to design the corpus. Though this data was enough for generating a corpus, however, more data from different institutes and different regions will provide a more diverse range of data and the resulting corpus design will be more robust. Moreover, in future researchers must utilize this corpus model and identify how well it contributes to translation; highlight its strengths and weaknesses to make it more comprehensive. This research is a single step, researchers are encouraged to take it forward and develop more corpus on the basis of the current model.

ACKNOWLEDGMENT

Authors would like to thank the administration of Universitas Negeri Malang and UIN Maulana Malik Ibrahim Malang for providing access to theses for this study. No financial support is received from any party.

AUTHORS CONTRIBUTION

All authors made useful contributions. Mohammad Ahsanuddin drafted the initial idea, Ali Ma'sum collected data, and Nur Anisah Ridwan worked on the final write-up. All authors worked collectively on data analysis and formulation of the model.

REFERENCES

1. Ahsanuddin, M. (2018). *Tashmim Al-mudawwanah Al-Mutawaziyah Li mustakhlash Al-buhuts Al-Ilmiah Al-Indunisiya Al-Arabiyah 'AlaDhauiNadzariyah Mona Baker Li Al-Takafu' Al-Lughawi Fi Al-Tarjamah*. UIN Maulana Malik Ibrahim Malang.
2. Alfaifi, A. & Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, *9*(2): 347-357. https://doi.org/10.1007/s10772-015-9285-5
3. Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–61. https://doi.org/10.17250/khisli.30.2.201308.001
4. Atwell, E., Al-Sulaiti, L., Al-Osaimi, S. & Abu Shawar, B. (2004). A review of Arabic corpus analysis tools. In B. Bel& I. Marlien (Eds.), *Proceedings of TALN04: XI Conference sur le TraitementAutomatique des LanguesNaturelles* (volume 2, pp. 229–234). ATALA.
5. Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh, Scotland: Edinburgh University Press.
6. Biber, D. & Jones, J. K. (2009). "Quantitative Methods in Corpus Linguistics." In Lüdeling, A. and M. Kytö (eds.) Corpus Linguistics: An International Handbook Vol. 2. Berlin, New York: Mouton de Gruyter, 1286–1304.
7. Biber, D. (1988). *Corpus linguistics.* Cambridge, UK: Cambridge University Press.
8. Bunchutrakun, C., Lieungnapar, A., Wangsomchok, C., &Aeka, A. (2016). A corpus-based approach to learning a tour guide talk. *International Journal of Humanities, Arts and Social Sciences*, *2*(2), 58-63. https://doi.org/10.20469/ijhss.2.20002-2

9. Cheng, W. (2012). *Exploring corpus linguistics:Language in action.* Oxford, UK: Routledge. https://doi.org/10.4324/9780203802632
10. Eddakrouri, A. (2016). Web-based (Searchable) corpora. Infoguistics. Retrieved from https://bit.ly/2Z5kv2r
11. Evert, S. (2009b). "Corpora and Collocations." In Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics: An International Handbook*. Vol. 2. Berlin, New York: Mouton de Gruyter, 1212–1248.
12. Evert, S. (2009a). "Rethinking Corpus Frequencies." Paper presented at the ICAME 30 Conference, Lancaster, May, 27-31.
13. Flood, B. J. (1999). Historical note: The start of a stop list at biological abstracts. *JASIS*, *50*(12). https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1066::AID-ASI5>3.0.CO;2-A
14. Kilgarriff, A. &Grefenstette, D. G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*(3), 333–347. https://doi.org/10.1162/089120103322711569
15. Luhn, H. P. (1960). Key word-in-context index for technical literature (kwic index). *American Documentation*, *11*(4), 288-295. https://doi.org/10.1002/asi.5090110403
16. Madina, T., Sholpan, Z., Zhanar, K., Bekzhan, A., &Kuandy, K. (2017). Implementation of the official language policy and the linguistic reality in Astana, Kazakhstan. *International Journal of Humanities, Arts and Social Sciences*, *3*(6), 264-274. https://doi.org/10.20469/ijhss.3.20003-6
17. McEnery, T. (1997). Multilingual corpora–current practice and future trends. In *13th ASLIB Machine Translation Conference* (pp. 75-86), Helsinki, Finland.
18. McEnery, T., &Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Oxford, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395
19. McEnery, T., & Wilson, D. A. (1996). *Corpus linguistics*. Edinburgh, Scotland: Edinburgh University Press.
20. O'Mahony, C. T. (2018). An analysis of dialects and how they are neither linguistically superior nor inferior to one another. *International Journal of Humanities, Arts and Social Sciences*, 4(5), 221-226. https://doi.org/10.20469/ijhss.4.10004-5
21. Sasongko, J. (2010). Application to Build corpus from crawling data with various data formats automatically. *Dynamic*, *15*(1).
22. Setiawan, T. (n.d.). Corpus linguistics in language teaching. *Seminar NasionalPerspektifBaruPenelitianLinguistikTerapan*, UniversitasNegeri Yogyakarta, Yogyakarta, Indonesia.
23. Setyawan, A. (2018a). *Benefits of the frequency list in language corps*. Retrieved from https://bit.ly/34xIQzd
24. Setyawan, A. (2018b). *Understanding concordance and how to use it in a linguistic corpus.* Retrieved from https://bit.ly/2Q9SwL9
25. Tognini-Bonelli, E. (2001). C*orpus linguistics at work*. Amsterdam, Netherlands: J. Benjamins Publishing. https://doi.org/10.1075/scl.6
26. Veerachaisantikul, A. &Chansin, W. (2018). A corpus-based approach to lessons development for EFL reading course. *Journal of Advances in Humanities and Social Sciences*, *4*(5), 197-205. https://doi.org/10.20474/jahss-4.5.1
27. Veerachaisantikul, A., &Chootarut, S. (2016). General vocabulary in Thai EFL university students' writing: A corpus-based lexical study. *Journal of Advanced Research in Social Sciences and Humanities*, *1*(1), 52-57. https://doi.org/10.26500/JARSSH-01-2016-0107
28. Wagner, J. &Nesselhauf, N. (2006). *Collocations in a learner corpus*. Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/scl.14
29. Wiechmann, D. (2008). "On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias." *Corpus Linguistics and Linguistic Theory* 4(2): 253-290. https://doi.org/10.1515/CLLT.2008.011
30. Zipf, H. (1949). *Human behaviours and the principle of least effort*. Boston, MA: Addison-Wesley.

