Advances in Social Science, Education and Humanities Research (ASSEHR), volume 154
International Conference on Culture and Language in Southeast Asia (ICCLAS 2017)

Projected Characteristics and Content of Arabic Corpus in Indonesia

Nur Hizbullah
Department of Arabic Literature University of Al Azhar Indonesia, Jakarta, Indonesia
hizbi77@gmail.com

Madian Muhammad Muchlis
Department of Islamic Education Islamic University of Jakarta, Jakarta, Indonesia
madianmuchlis74@gmail.com

Abstract— **Utilization and integration between linguistics and information and communication technology produce a result in the form of a language corpus. Corpus is a collection of data prepared systemically and is developed in such a way to be used as research data. In general, the content of a corpus relates to the purpose preparation of the corpus itself in the context of linguistic researches. In addition, the corpus' content relates to the availability of data materials to be included in the corpus. With its long history and wide coverage of Arabic teaching in Indonesia, there are quite a plenty of materials and data on and in Arabic language that can be documented and compiled to be used as corpus. This ascertains that Arabic Corpus in Indonesia will be filled by various data materials. Under the descriptive-comparative method, this paper will describe various types of Arabic corpus, particularly the aspect of corpus content and compare the content in the corpus and the predicted availability of content materials in the context of the plan to prepare Arabic Corpus in Indonesia. By referring to the existing corpus, it can be projected that the Arabic Corpus to be made in Indonesia is a regional and diachronically corpus. This corpus contains seven distinct classifications in accordance with the availability of data in the field. Hence, the effort of drafting this corpus is important and strategic in order to make a documentation of the Arabic linguistic data that is real produced by the Indonesian speakers and this corpus will be able to showcase the richness of Arabic language in Indonesia for use to develop research in various Arabic studies infuture.**

Keywords*— *Arabic Corpus; corpus characteristic; corpus content; comparative corpus

I. INTRODUCTION

The development and advancement of information and communication technology (ICT) has significant impact on the development and dynamics of all disciplines, including linguistics and its studies. At present, there are already quite a lot of linguistic studies that make use of existing technical devices in TIK and give birth to a new study called corpus linguistics. According to Adolph, corpus linguistics makes real and authentic languages, whether oral language or written language as its study objects [1]. In the context of facilitating and producing an objective study, linguistic data is prepared to be systematically examined to make a corpus. This corpus can later be used in various linguistic studies such as teaching, literature, translation, etc. This relates to the data for each object of study and finally specific data on these studies will be compiled to become a special corpus.

Many languages in the world have already had their corpus, including Arabic language. At present, there are already several Arabic Corpora prepared by Arab researchers, both from the Arab region or beyond. This paper refers to the compilation of Al-Sulaiti which is quite representative. This expert collected data of 29 Arabic corpora and map them based on seven aspects namely content, language period, language speakers, language variation, total languages, material type and writing style [2]. The latest classification was conducted by Eddakrouri who managed to record 35 Arabic corpora [3]. The corpus is classified simply by material type and processing. Eddakouri listed 18 web-based Arabic corpora and the material cannot be downloaded by users. However, the web provides a search facility and word processing in a systematic way. In addition, there are 17 corpora which belong to a textual corpus, in the sense that the corpus owner provides an Arabic data corpus and can be downloaded by the user himself. The processing is done by the user by utilizing a corpus processing application. The classification model on corpus that the author deems adequate is among others classification by Nesselhauf [4] and Sketch Engine [5]. Both classifications are similar but only that Nesselhauf classifies corpora in the context of "pairing" for seven "pairs", while Sketch Engine classify them only descriptively into six groups. Hizbullah and Rachman, proposes classification which integrate the two classifications that become reference, namely Nesselhauf and Sketch Engine, as follows.

1. Based on the corpus content: a. general corpus and b. special corpus.
2. Based on the language period: a. historical corpus and b. modern corpus.
3. Based on the language speakers: a. learner corpus b. corpus for native speakers c. regional corpus.
4. Based on the variation of language content: a. monolingual corpus b. parallel/multi lingual corpus.
5. Based on the language style; a. oral language corpus b. written language corpus c. mixed corpus.
6. Based on its material type: a. text corpus b. text corpus plus audio c. text corpus plus audio visual.
7. Based on its writing: a. Orthographical corpus. Annotation corpus [6].

By referring to this classification, this paper has the purpose to identify the existence of various data of Arabic linguistics in Indonesia which conforms to this classification

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

to be used as corpus. Furthermore, the characteristics and potential content in the Arabic Corpus in Indonesia will be projected. The existence of this corpus is very important and strategic in order to compile a comprehensive and systematic documentation of all Arabic language data that produced by the native Indonesian speaker. The data is also important as a foundation for improving the quality and development of innovative Arabic studies and researches in Indonesia. The effort to draft the corpus has become a common concern and commitment to be realized between the Arabic Teachers Association in Indonesia (IMLA) and the Riyadh, Saudi Arabia Center for Research and Intercommunication Knowledge (CFRIK).

II. METHOD

Projection on the characteristics and content of Arabic Corpus in Indonesia is based on the description of existing classification of Arabic Corpus. The content of previous Arabic corpus was projected comparatively with possible availability of data materials for Arabic Corpus in Indonesia. Hence, the existing potential data materials are used to project the type of Arabic Corpus to prepare in Indonesia.

III. RESULT AND DISCUSSION

Referring to the previous two classifications, Nesselhauf and Sketch Engine, in general, Arabic Corpus in Indonesia is classified as regional corpus. This corpus is unique because its linguistic data and fact are produced by non-Arab speakers from or residing in Indonesia. This corpus cannot be found in many parts of the world. If there is, we can mention a work by Hassan and Ghalib from the International Islamic University, Malaysia [7]. This corpus, however, contains only a collection of scientific papers (thesis and dissertation) as the final assignments for master's degree and doctoral degree students from this campus as stated by the author [8]. In addition, Arabic Corpus in Indonesia will be in the form of diachronic corpus which contains linguistic data in the past and at present.

Referring to Hizbullah and Rachman, there are a number of data materials in Arabic language made by Indonesian writers that may potentially be compiled to become an Arabic Corpus [6]. These materials are in the form of all kinds scientific papers presented in scientific forums or published in scientific journals, literary works or popular works, published or not published, works in Arabic writing even though not in Arabic, Arabic language textbooks, from the basic level to university level, Al-Quran and its translation in Indonesian language or local languages in the Indonesian archipelago, special compilation on Indonesian words deriving from Arabic language, and all forms of learning products from Arabic language from learners.

Based on the data materials from the corpus, we can project the content of Arabic Corpus in Indonesia by referring to the classification of corpus language from Nesselhauf and Sketch Engine. In view of variety of data materials existing in Indonesia, this corpus has the potency not to become a general corpus, but a corpus which contains collection and combination of specific corpus dividing into several topics under the following discussion.

1. Scientific works Corpus, containing collection of papers, research reports, scientific papers, and scientific journals related to scientific forum or scientific publishing in Indonesia and overseas. There is also potential existence of scientific works not published by their authors because of certain reasons but are appropriate to be included as the data materials for corpus. In addition to text materials as above, we can also consider inclusion into this classification audio recording or audiovisual from lectures/scientific oration in Arabic by Indonesian speakers.
2. Popular works Corpus, containing periodicals or popular works in Arabic language such as magazines, bulletins, newspaper, and so forth circulating in Indonesia. If there any works in this group written in two languages, then the corpus to be prepared is in the form of bilingual or multilingual parallel corpus.
3. Literary works Corpus, containing literary works in Arabic language or not in Arabic language but are written in Arabic letters. This type of writing is popularly known in the Indonesian archipelago as "Arab Melayu", "Arab Jawi", or "Arab Pegon". All the works published or not published will be classified into this corpus.
4. Religious works Corpus, projected to appear under the reason that there are many works by *ulema* in the archipelago written in Arabic regarding Islamic topics in the field of *fiqh*, *aqidah, tafsir*, and so forth, in the past and at present. It is possible that this type of works is written in two or more languages that will be made into a parallel corpus.
5. Quranic Corpus and its translation, projected in the context of providing a data corpus for linguistic studies on Al- Quran in Indonesia. These studies can be directed at the aspects of translation, grammar and stylistics and so forth. The main data material for this corpus is corpus for Al-Quran texts and its official translation in Indonesian language which is the version of the Ministry of Religious Affairs of the Republic of Indonesia. In addition, it is worth making the corpus for the Al-Quran and its translation written by some Islamic ulemas who have been widely popular among Indonesian Muslim Communities. The Al- Quran and its translations into several local languages in the archipelago can be also included in this group. This can be realized because the Ministry of Religious Affairs of R.I has already had several models of Al-Quran and its Translations into the Javanese language, Sundanese, and so forth. This Corpus will be in the form of parallel corpus parallel because it consists of two texts in two languages, namely text of Al-Quran verses in Arabic language and its translation text in Indonesia language and/or local languages in the archipelago.
6. Corpus of loan words from Arabic-Indonesian, prepared considering there are a large number of loan words in

Indonesian language deriving from Arabic, because of other than the dynamics in the use of Arabic words among the community, particularly Muslims in Indonesia. This corpus can be used to periodically observe potential appearance of new loan words from Arabic language used by Indonesian speakers. Analysis than can be carried out on this kind of data is among others on the similarity or difference of lexical meanings between the two languages, shift in meaning, contrastive analysis and so forth.

7. Learner Corpus, containing productive data as a result of learning by Indonesia community speaking Arabic as a foreign language, in the form of writing, composition, articles, speech recording, conversation recording, debate recording carried out in the context of learning Arabic by foreign speakers (Indonesians). This data is collected from learning process in educational institutions, both at the basic, intermediate or higher levels and those conducted in formal and non-formal education. The moments related to those data can be in the form of lectures, practices, competition, festivals, and so forth. Arabic text books can be included into this category or made into a separate sub- category in future. All kinds of media for learning and teaching Arabic made by Indonesian teachers can also be included, in the form of text or audio and audiovisual media.

IV. Conclusion

By taking into account the availability of data materials for Arabic Corpus in Indonesia, this corpus is very realistic to be materialized. The next step is exploring and identifying corpuses found. Regional approach is important in this matter considering that some regions in the archipelago have Islamic religious practices and history and teaching of Arabic language has the potency to have works written by the local ulema in Arabic. The projection of Arabic language characteristics in Indonesia as regional corpus and diachronic corpus will be enriched by seven content classifications which will later be expanded again to conform to the dynamics of data finding in the field. This corpus will become a new contribution and is important as the new foundation for development of Arabic studies in Indonesia, especially in relation to the use of information technology in language research. Besides, the existence of this corpus can be an addition to showing the world and native Arabic speakers of the Arabic treasures in the archipelago. Furthermore, there is a need for studies and experiments on adequate access systems and in accordance with the needs of Arabic research in Indonesia.

Acknowledgement

We acknowledge and appreciate Universitas Al Azhar Indonesia (UAI) and the Institute for Research and Community Service (LP2M) of UAI for their supports and assistance in our participation representing the Department of Arabic Literature of UAI in the event of International Conference on Culture and Language in Southeast Asia (ICCLAS) titled “Islam and Identity of Southeast Asia Languages and Cultures” in the Syarif Hidayatullah State Islamic University, Jakarta, November 14-15, 2017.

References

[1] S. Adolphs, *Introducing Electronic Text Analysis - A Practical Guide for Language and Literary Studies.* New York: Routledge, 2006.

[2] L. Al-Sulaiti, “Arabic Corpora,” 2010. [Online]. Available: Avalaible:http://www.comp.leeds.ac.uk/eric/latifa/ arabic_corpora.htm [Accessed: 23-Sept-2017].

[3] A. Eddakrouri, “,Arabic Corpora”,” 09-Nov-2017. [Online]. Available: https://sites.google.com/a/aucegypt.edu/ infoguis-tics/directory/Corpus-Linguistics/arabic- corpora.

[4] N. Nesselhauf, “Corpus Linguistics: A PracticalIntroduction,” 2011. [Online]. Available: http://www.as.uni-heidelberg.de/personen/Nesselhauf/ files/Corpus%20Linguistics%20Practical%20 Introduction.pdf. [Accessed: 19-Sep-2017].

[5] S. Engine, “Arabic TenTen Corpus,” 2015. [Online]. Available: https://www.sketchengine.co.uk/ artenten-corpus/. [Accessed: 17-Sep-2017].

[6] N. Hizbullah and F. Rachman, “Beberapa Model dan Karakteristik Korpus Bahasa Arab sebagai Acuan Penyusunan Korpus Bahasa Arab di Indonesia (Several Models and Characteristics of Arabic Corpus as Reference in Preparation of Arabic Corpus in Indonesia),” presented at the International Seminar on Arabic, Universitas Muhammadiyah, Yogyakarta, 2017.

[7] H. Hassan and M. Ghalib, “Arabic Concordancer,” 2010. [Online]. Available: www.arabicconcordancer. com. [Accessed: 17-Aug-2017].

[8] Interview with Assoc. Prof. Haslina Hassan in Yogyakarta, August 12, 2017