Conference Paper

Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia

Nur Hizbullah[1], Zaqiatul Mardiah[1], Yoke Suryadarma[2], Luthfi Muhyiddin[2], Oyong Sofyan[3], and Ferry Hidayat[3]

[1]Universitas Al Azhar Indonesia (UAI), Jakarta, Indonesia
[2]Universitas Darussalam Gontor (UNIDA), Ponorogo, Indonesia
[3]Tazakka Modern Islamic Boarding School, Pekalongan, Indonesia

Corresponding Author:
Nur Hizbullah
hizbi77@gmail.com

Received: 1 July 2019
Accepted: 18 July 2019
Published: 31 July 2019

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the AICLL 2019 Conference Committee.

Abstract

Arabic corpora in the Middle Eastern countries have showed a considerable increase in availability and quantity. Unfortunately, Arabic corpora outside Arabian peninsula have still been deemed something new. Existence of a rich variety of linguistic products of Arabic language in Indonesia has the potential for the birth of, among others, Arabic learners' corpora. The corpora will provide concrete evidence of abundance and continuity of the Arabic learners' corpora and will lay a sound foundation for future Arabic language research and teaching. This study aims at identifying existence of Arabic learning and teaching products as useful raw materials for creating Arabic corpora in Indonesia. Methods employed in this study are conducting surveys to three *pesantrens* (modern Islamic boarding schools) located in Jakarta, Central Java, and East Java, distributing questionnaires, and carrying out interviews. As for interviews and questionnaires, they are carried out in order to mine substantial data out of the teachers as respondents chosen at random. The study concludes that activities that produce the Arabic learning and teaching products can be categorised into three categories; formal activities in the form of class instructions, non-formal and informal activities in the form of trainings of Arabic mastery. The products of Arabic learning activities vary accordingly, such as the teachers' works, the students' works, and popular works. The Arabic linguistic products are commonly identified as hand-written texts, which still must be processed digitally in order to be corpora materials, due to lackness of Arabic linguistic products that had been digitalized.

Keywords: Arabic learners' corpora, Arabic corpora in Indonesia, Arabic research and teaching, Arabic language in *pesantren*

1. Introduction

Existence and quantity of Arabic corpora in Middle Eastern region, despite relative novelty, have been increasing. The increase is followed by development of group variety as well as corpus data volume. Nevertheless, this phenomenon occurs limitedly within Arab countries (Zaghouani: 2014). Outside the Arab region, varied Arabic corpora have

How to cite this article: Nur Hizbullah, Zaqiatul Mardiah, Yoke Suryadarma, Luthfi Muhyiddin, Oyong Sofyan, and Ferry Hidayat, (2019), "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia" in *The Second Annual International Conference on Language and Literature*, KnE Social Sciences, pages 980–989. DOI 10.18502/kss.v3i19.4922

not yet been discovered so far. However, an Arabic corpus-collection work published by arabicconcordancer.com is an exception. The Arabic corpus compiled by Hassan and Galib is an academic Arabic corpus ranging from academic final theses, seminar proceedings, and journal articles issued internally by International Islamic University of Malaysia (Hassan & Galib: 2010).

Except this corpus engine, out of the Arab vicinity, there have been no other Arabic corpora representing practice and use of Arabic language in any region on the globe. The absence of Arabic corpora particularly in Southeast Asian zone creates problem and, at the same time, challenge in context of multidisciplinary as well as modern linguistic research development. It necessarily stresses the importance of Arabic corpora as real linguistic data documentation which will become valuable research object.

Considering the great width of practice and of use of Arabic language in Indonesia, the problem of Arabic corpora absence could have been solved. A great variety of Arabic written works are discovered and found in many a formal and non-formal educational institutions as well as in their continuous teaching practices, not to mention individual manuscripts written in Arabic. All the sources only emphasize that Indonesia is rich in Arabic corpus raw materials. Uniquely, the corpus materials are written and created by Indonesian native speakers living far away from Arab peninsula. They must be seen as very valuable components in order to make a path of compilation of Arabic corpora in Southeast Asian territory, especially in Indonesia. The path, however, takes big challenge in terms of corpus collection mechanism and corpus documentation system.

It is interesting to study how many kinds of raw material there are in the context of Arabic teaching and learning that are able to be documented out of abundant Arabic corpus raw materials. According to corpus classification, this kind of corpus is a learner corpus, produced by learners, in this case Indonesian language speakers, as learners of Arabic language as a foreign language through the process of learning and teaching in educational institutions. A kind of the institutes providing Arabic education and instruction is *pondok pesantren*. The institution is widely known for its creatively, innovatively, and continuously teaching of Arabic language. In *pondok pesantren*, students learn Arabic in their classrooms and speak as well as write it actively in their daily life; they are also encouraged to be creatively holding lots of activities supporting and sharpening Arabic language skills. Besides the activities, the *pesantren* students are mentored by their teachers to innovate in outside-classroom learning models. This process is carried out systematically and continually in their study time until they acquire qualification and get certain proficiency level. Thus, it can be assumed that *pondok pesantrens* have plentiful Arabic learning products that can be documented as corpora.

Referring to the problem, the situation, and the background, this study aims to elaborate on the types of activities in Islamic boarding schools relating to learning Arabic and to identify the variety of learning materials produced through these activities that can be utilized as corpus materials. In addition, the identification is also directed at the format of the material, whether it is still in the form of conventional handwritings or printed material that should first be processed into digital text, or whether it is already in the form of digital text written and processed with a text processing application (word processor). The real condition of the corpus materials greatly determines the process and mechanism of digitization through scanning and/or rewriting. The better the condition of the materials, the easier it is to scan and it only takes a little time to edit it. Conversely, the corpus material difficult to read requires researchers to rewrite into digital text. This is not the case when the material is already in the form of a digital document. The formatted material only needs converting into a special format and then needs editing before it is made into a corpus.

The identification is expected to be used as a reference map on the variety and format of learner corpus raw material obtained from *pondok pesantrens*. The reference map in turn can be used as a guideline to carry out similar identification more broadly in *other pondok pesantrens*. It is possible to modify the reference map for research on educational institutions such as *madrasas*, public schools, or non-formal educational institutions providing Arabic language instruction, depending on the specificity of their respective institutions. In relation to the purpose and sustainability of this research, this reference map can reflect the novelty of this research which can be developed innovatively in future studies.

2. Literature Review

The existence of Arabic language teaching and learning in Indonesia is inseparable from its relation to the history of Islamic education in *pondok pesantrens* and *madrasas* since several centuries ago. Baharuddin noted, *pondok pesantrens* in the archipelago were erected in the context of the socio-religious dynamics of traditional communities in Java, Sumatra, and other regions in the archipelago. Learning Arabic in *pondok pesantrens* is also closely related to teaching Islamic values and Islamic principles to the people of the archipelago in the context of *da'wa.* When the first scholars taught Islam by referring to Arabic-language literature, such as the Koran, *Hadith*, and other books, there was language contact and cultural interaction of the Muslim community in the archipelago with Arabic language in the form of assimilation and acculturation of Islamic values through a number of Arabic terms and concepts with their meanings and explanations. As a result, a number of Arabo-Islamic terms were absorbed, assimilated, adopted, and adapted into the archipelagic languages (Maulana: 2018). Over time, in the process of *da'wa* and teaching, several Islamic *pesantrens* compiled curriculum, teaching materials, even written works in Arabic. This creationis not only limited to Arabic language material but also broader Islamic studies materials (Baharuddin: 2014). The Arabic linguistic works had emerged along the development of the methodology of teaching Arabic language in the archipelago. In general, the methodology of Arabic teaching in pondok pesantrens is divided into two approaches, namely the traditional-classical approach and the modern approach. Both approaches, with all their advantages and disadvantages, have been tested for a long time and have produced works that are locally, nationally, and even internationally known (Nurkholis: 2018).

In the modern period of Indonesian history, most of the Arabic works had been documented, both conventionally and digitally, and become the object of multidisciplinary research conducted by experts. In particular, in the context of this study, documentation of linguistic products was directed at the creation of corpus, namely a special digital format of written and oral linguistic products that are transcribed and could be processed by means of certain computer applications. Corpus itself has many varieties. The various corpora are mentioned by Nesselhauf (Nesselhauf:2011) and Sketch Engine (Engine: 2018) and summarized by Hizbullah and Rachman into seven categories (Hizbullah & Rachman: 2017).

Various Arabic writing and spoken works produced through the process of learning Arabic can be categorized as a learner corpus. One of the existing learner corpus of Arabic language is Arabic Learner Corpus (Alfaifi & Atwell: 2013a) (Alfaifi, Atwell, & Ibraheem: 2014) (Alfaifi: 2015). This corpus captures data from a number of Arabic native speakers-learners and some Arabic-as-a-foreign language learners. Other corpus in the context of learning is children's Arabic language corpus (Arabic Children's Corpus) which is a compilation of a number of texts contained in textbooks and stories for early-age students (Al-Sulaiti, Abbas, Brierley, Atwell, & Alghamdi: 2016) . Alfaifi and Atwell project that the existence of this kind of learning corpus can be used for various fields of research, including intra-language contrastive analysis, learner dictionary creation, second language acquisition, design of learning materials, and optical character recognition (OCR) techniques (Alfaifi & Atwell: 2013b) .

3. Research Method

This research is a survey research based on qualitative method. The survey is directed at three modern pondok pesantrens located in Jakarta, Central Java and East Java. The modern pesantrens were chosen on consideration of the large variety of Arabic-language activities and the number of products created through these activities. In the survey, interviews and questionnaires were carried out on a number of randomly selected Arabic language and Islamic studies teachers. In addition, observations were also carried out to identify forms of Arabic language materials produced by involving parties in learning and teaching Arabic language in these pesantrens. The findings of the survey are described descriptively and presented in table form.

4. Result and Discussion

Based on a survey conducted on three pesantrens as the object of this research, the following is a summary of the findings of Arabic language products based on the categories of activities carried out in these institutions.

Table 1: Classification of Arabic-using activities and their possible corpus raw materials.

No.	Activity	Sub-activity	Product	Format
a.	Formal-curricular	1) Classroom Arabic instruction: *Muthola'ah* (Arabic Reading), *Tamrin Lugah* (Arabic Workbook), *Insya'* (Arabic Composition), *Muhadatsah* (Arabic Speaking Practice), *Nahwu* (Arabic Grammar), *Sharaf* (Arabic Conjugation), etc.	a) Textbooks and modules b) Exercises, speech performance, and structured tasks.	Spoken (raw)and written (raw as well as digitalized)
		2) Classroom Islamic Studies instruction: *Fiqih, Ushul Fiqih, Tarikh Islam, Musthalah Hadits, Tarbiyah*, etc.	a) Textbooks and modules b) Exercises, speech performance, and structured tasks.	Spoken (raw/undigitalized) Written (raw as well as digitalized)
		3) Mid-year Examination and Year-End Examination	Examination answer sheets	Written (raw)
		4) Practicum teaching (*Amaliah Tadris*)	Lesson plan notebooks	Written (raw)
		5) Classic Islamic Books Reading (*Fathul Kutub*)	a) thematic-descriptive analysis b) Academic paper	Written (raw)

No.	Activity	Sub-activity	Product	Format
		6) Scholastic orientations	Activity Guideline Book	Spoken (raw/undigitalized)and written (raw as well as digitalized)
		7) Curriculum plan	Curriculum Book	Written (raw and digitalized)
		8) Academic achievement report	Diploma certificate and report book	Written (raw and digitalized)
b.	Nonformal/ Extracurricular	1) Weekly conversation practice (*Muhadatsah Usbu'i*)	Structured practice	Spoken (raw)
		2) Weekly public speech training (*Muhadharah*)	Public speech texts	Written (raw)
		3) Daily announcements	Announcement texts	Written (raw)
		4) Language mastery competitions	Competition texts and materials	Spoken and written (raw and digitalized)
		5) Study tour	Language practice	Spoken (raw)
		6) Weekly flag ceremony	Ceremony texts	Spoken and written (raw)
		7) Security section's disciplinary action	a) disciplinary motivation guideline book b) general discipline-breaking court	Spoken and written (raw) Spoken(raw)
		8) Language section's disciplinary action	a) Language disciplinary motivation speech and guideline book b) Language discipline-breaking court	Spoken and written (raw) Spoken (raw)
c.	Informal	1) Vocabulary enrichment	Vocabulary using exercises	Written (raw)
		2) Published books	a) Popular articles b) Annual publications	Written (raw and digitalized)
		3) Language using atmosphere conditioning	Writings, graffitis, announcements, pamphlets, banners, security section documents, language section documents, students organization documents	Written (raw and digitalized)

The table illustrated in general Arabic-using activities in the modern pesantrens are divided into three groups, namely formal-curricular, non-formal or extracurricular activities, and informal activities. The division is actually also related to the system of distribution of the student activities in general. However, the parallelization of activities with linguistic aspects is unique and special. This was inseparable from the motives and orientation of the pesantrens which strongly emphasized the importance of mastering foreign languages, in this case Arabic and English, by their students and even it was carried out by involving all elements of the educational agents in it, such as teachers, guidance counsellors, daily workers and others. It is also seen that the scope of Arabic-using activities is very broad and reaches not only fundamental aspects of education, but also technical and specific aspects of practical daily life.

Generally, it can be identified that variety of formal-curricular sub-activities emphasizes Arabic aspect in the scientific and academic field. Apparently, not only Arabic-themed lessons were delivered in Arabic. Many other lessons, especially in the field of Islamic studies, were also conveyed in Arabic and using Arabic textbooks. The shape of the products created as corpus raw material is actually very diverse and unique. However, most of the its format is still in the form of conventional handwriting by students or teachers, although there are already some materials in the form of digital data. In this part, the students' Arabic works in the form of filled-exercises, structured assignments, and narrative/essay exam answer sheets are important and interesting and are examined to see how far their Arabic language skills are academically in the context of certain lessons. Likewise, the preparation of teaching practice materials and scientific articles that are written and spoken through book review activities.

Apart from formal-curricular activities, the variety of non-formal/extracurricular activities are complementary in character. However, the weight of the activity is aggravated by a disciplinary approach, in the sense that pesantren students are directed to carry out these activities with commitment and compliance and will face the consequences of disciplinary action for their non-compliance. Despite extracurriculum, the linguistic products of the students produced through these activities can also reflect their language skills more generally and broadly, and no longer limited to the academic realm. In public speaking trainings (*Muhadharah*) and language competitions, for example, the students are accustomed to speaking and writing on various topics. The recording and documentation of materials is also very important as an indicator of the students' Arabic proficiency on a wider scale.

The various informal activities is generally carried out outside the classroom and closer to the daily lives of students, such as in dormitories, sports venues, courses, mosques, and so on. In the dormitory, the students get daily vocabulary enrichment and are strengthened by vocabulary mastery exercises. In addition, in small-scale courses and publishing institutions, the students also learn to write in Arabic and the writings are published in wall magazines or simple print publications. There is also an annual publication in three languages, one of which is Arabic, which is more like an annual report for public consumption. What is quite unique is the use of Arabic in almost all areas of the pesantren life, including tools, equipment, documents, displays and others that come in direct contact with the lives of the students, for instance, billboards, bulletin boards, signs, storage media, licensing cards, certificates and so on. It can all be categorized as a Arabic learner corpus considering the motive for making it is in order to educate students to be familiar and familiar with the use of foreign languages in everyday life, besides being factually made by the students themselves or the teachers. It is interesting to study in this aspect, for example, about language variations, vocabulary, grammaticality, and the frequency of vocabulary use in certain domains. However, the digitalization of the abundant materials that are still in the form of conventional data requires more time and effort.

Considering the depth, breadth and specificity of the existence and availability of Arabic-using activities documents that can be used as corpus materials, it can be said that modern pondok pesantrens are very open areas for corpus-based Arabic research. Although the materials are more of a learning corpus material, however, given the breadth of the educational horizon in the pesantrens, researches on the Arabic corpus can also be carried out in linguistic and translation levels, in addition to the study of teaching Arabic, obviously. Thus, it is important to document linguistic products in pondok pesantrens in a more massive, structured and sustainable manner so that the quality of Arabic language teaching in them can be improved and corpus-based researches on Arabic language in pondok pesantrens can be further developed.

5. Conclusion

This study concludes that raw materials having the potential to be a learner corpus of Arabic in pondok pesantrens can be obtained through three types of activities, namely formal-curricular, non-formal/extracurricular, and informal activities as well as sub-activities included in it. The learner corpus in pesantrens is very broad, diverse and specific, yet only a few Arabic learning products have been digitalized, while most of them are still raw materials, both oral and written. Seeing the breadth and variety of Arabic linguistic data in pondok pesantrens, opportunities and challenges are open to further explore the real situation and conditions of Arabic teaching and learning through corpus-based multidisciplinary research.

Acknowledgements

Gratitude should be expressed to the Universitas Al-Azhar Indonesia's Institute for Research and Community Service (LP2M) which has funded this research and authors' participation in the "2nd Annual International Conference on Language and Literature" (AICLL) North Sumatra Islamic University, Medan, 2019. Warm gratitude should also be addressed to the pesantrens' headmasters where this research was carried out and to fellow researchers of the Universitas Darussalam (UNIDA) Gontor, Ponorogo and Pondok Modern Tazakka Batang who involve in co-writing this article.

[1] Al-Sulaiti, L., Abbas, N., Brierley, C., Atwell, E., & Alghamdi, A. (2016). Compilation of An Arabic Children's Corpus. In *10th Language Resources and Evaluation Conference*. Retrieved from http://eprints.whiterose.ac.uk/100839/1/ArabicChildrensCorpus180915.pdf, April 17th, 2019

[2] Alfaifi, A. (2015). Arabic Learner Corpus. Retrieved from https://www.arabiclearnercorpus.com/, December 8th, 2018

[3] Alfaifi, A., & Atwell, E. (2013a). *Arabic Learner Corpus v1: A New Resource for Arabic Language Research*. *Second Workshop on Arabic Corpus Linguistics*. https://doi.org/10.1039/c3ib40166a, April 17th, 2019

[4] Alfaifi, A., & Atwell, E. (2013b). *Potential Uses of the Arabic Learner Corpus*. Retrieved from http://corpus.leeds.ac.uk/teaching/l3t/l3t2013_submission_7.pdf, April 17th, 2019

[5] Alfaifi, A., Atwell, E., & Ibraheem, H. (2014). Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In *Proceeding of Learner Corpus Studies in Asia and the World (LCSAW) 2014, Kobe University, Japan*. (pp. 77–89). Retrieved from https://catalog.ldc.upenn.edu/docs/LDC2015S10/ALFAIFI_LCSAW2014.pdf, April 16th, 2019

[6] Baharuddin, I. (2014). Pesantren dan Bahasa Arab. *Thariqah Ilmiah: Jurnal Ilmu-Ilmu Kependidikan Dan Bahasa Arab*, *01*(01), 16–30. Retrieved from http://jurnal.iain-padangsidimpuan.ac.id/index.php/TI/article/view/253/234, April 4th, 2019

[7] Engine, S. (2018). Types of Text Corpora. Retrieved from https://www.sketchengine.eu/user-guide/user-manual/corpora/corpus-types/, October 31st, 2018

[8] Hassan, H., & Galib, M. F. M. (2010). Arabic Concordancer المنقب العربي. Retrieved from http://arabicconcordancer.com/#, October 31st, 2018

[9] Hizbullah, N., & Rachman, F. (2017). Beberapa Model dan Karakteristik Korpus Bahasa Arab sebagai Acuan Penyusunan Korpus Bahasa Arab di Indonesia. In مجموعة بحوث مؤتمر دولي واستكتاب "اتجاهات اللغة العربية في العصر الرقمي (تعليميا ، أدبيا ، برمجيا) ". Yogyakarta: LP3M Press.

[10] Maulana, M. F. (2018). Islamization Versus Deislamization of Language a Case of Indonesian Vocabularies. In Purwarno, M. Manugeren, A. Suhendi, P. Siwi, & S. Ekalestari (Eds.), *The 1st Annual International Conference on Language and Literature (AICLL)* (pp. 400–408). Medan: KnE Social Sciences and Humanities. https://doi.org/10.18502/kss.v3i4.1957, April 16th, 2019

[11] Nesselhauf, N. (2011). *Corpus Linguistics: A Practical Introduction*. Retrieved from http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/ Corpus Linguistics Practical Introduction.pdf, April 5th, 2019

[12] Nurkholis. (2018). Metode Pembelajaran Bahasa Arab di Pondok Pesantren Tradisional. *An-Nabighoh: Jurnal Pendidikan Dan Pembelajaran Bahasa Arab*, *19*(2), 249–267. https://doi.org/http://dx.doi.org/10.32332/an-nabighoh.v19i2.1002, April 16th 2019

[13] Zaghouani, W. (2014). *Critical Survey of the Freely Available Arabic Corpora*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1702/1702.07835.pdf, April 15th 2019