Al-Quran Translation Parallel Corpus for the Development of Al-Quran Translation Studies in Indonesia

1st N. Hizbullah[1], 2nd Z. Mardiah[1]
{nurhz@uai.ac.id[1], zaqiapurwono@gmail.com[1]}

Universitas Al Azhar Indonesia, Jakarta, Indonesia[1]

Abstract. In the field of language studies and research, Al-Quran is often used as an object and source of primary data. In line with the dynamics in technology, data and format of conventional Al-Quran writing is converted and developed to become a digital data. Further development brings Al-Quran digital data becoming a corpus with all its variants, including parallel corpus which pairs Al-Quran texts in Arabic with its translations in non-Arab languages. This makes use and utilization of this type of data easier for the development of broader studies on Al-Quran. This study aims to describe the urgency of Al-Quran parallel corpus and its Translation in Indonesian language and local languages in the Indonesian archipelago. This study applies the descriptive method to discuss the availability of materials and data of Al-Quran translation in several local languages in the Archipelago. Furthermore, we will also describe the experimental process in preparing an initial model for the Al-Quran parallel corps and its translation in Indonesian language. The result of this study reveals that the availability of products containing materials and data of Al-Quran and its Translation in Indonesian language and local languages can be an important capital and main basis for preparation of an Al-Quran parallel corpus and its Translation for the development of Al-Quran translation studies in Indonesia. Furthermore, the result of study can be utilized to evaluate or review the methodology used for translation of Al-Quran into local languages in Indonesia. It is also important to introduce this step and continued with strategic efforts to encourage Al-Quran translation studies into local languages. This prototype of model Al-Quran parallel corpus and its Translation into Indonesian language can at least be tested through corpus processing applications available in the website of Sketch Engine.

Keywords: parallel corpus; Al-Quran and translation; Indonesian Translation; local language translation

1 Introduction

In Indonesia, translation of Al-Quran has been performed since around the 17th century. In reference to Izzan, Baihaki mentioned that the first person, in the record, translating Al-Quran into Malay language is Syekh Abdul Rauf bin Ali AlFansuri (died in 1693). He is considered as the first person to use Arab-Malay letters in his work [1, p.46]. This work is assumed as the initial point of the translation history of Al-Quran into Indonesian language. Officially, the Government of the Republic of Indonesia pioneered translation of Al-Quran

into Indonesian language and published it in 1965 through *Lembaga Penyelenggara Penterjemah Kitab Suci Al-Qur'an*, or the Institution of Al-Quran Translation of the Ministry of Religious Affairs [2, p.229]. In addition to Indonesian language, in the past, Al-Quran had also been translated into local languages in the Archipelago. Baihaki recorded from several literatures that at least since the 1930s, Al-Quran was found to have been translated into Javanese language in Solo, Central Java [1, p.53]. This dynamic describes activities of Al Quran translation vigorously done since long time ago. This also reflects the dynamic of development of Islamic studies in this country. In line therewith, there are also many Islamic studies and linguistic studies based on or referring to Al-Quran.

In the past, Islamic and linguistic studies still relied on printed Al-Quran Mushaf, and now in line with the time and development of computer technology, studies require conversion of printed Mushaf to become digital Mushaf. This digital Mushaf contains various formats, both images and texts. Particularly, in the context of linguistic studies, for example, digital texts are required in a certain format referred to as corpus. This corpus can be processed with a certain application and the corpus must be specially formatted to be compatible with the relevant application.

Digital data model or Al-Quran corpus are now widely available and varied to be used as a data object. However, specifically for the purpose of translation studies, a special corpus is needed called parallel corpus containing two texts in two languages namely the source language and target language.For that purpose, this study aims to reveal the urgency of Al-Quran parallel corpus and its translation in Indonesian language and local languages by utilizing material and data readily available for development of Al-Quran translation studies in Indonesia. Furthermore, an experiment preparing the prototype of Al-Quran parallel corpus and its translation to Indonesian language will be elaborated.

2   Methodology

This study applies the descriptive method to present the finding as a result of literature studies on Al-Quran translation into Indonesian language and local languages along with findings from the result of survey on the availability of Al-Quran corpus. Furthermore, an experiment will also be carried out on preparation a prototype of Al-Quran parallel corpus and its translation into Indonesian language. To prepare this corpus, the material and data of Quranic Arabic texts file should be in the format of *.txt (extension of text files) without diacritics and downloaded from tanzil.net. The Indonesian translation texts are obtained from the *Lajnah Pentashihan Mushaf Al-Quran* of the Ministry of Religious Affairs of the Republic of Indonesia. The preparation of parallel corpus prototype refers to the mechanism stipulated by the Sketch Engine.

3   Literature review

According to the *Lajnah Pentashihan Mushaf Al-Qur'an* of the Ministry of Religious Affairs of the Republic of Indonesia, Al-Quran translation published by the Ministry of Religious Affairs of the Republic of Indonesia is still more popular and mostly used than Al-Quran Indonesian translation, for example by Hassan, Shihab, Thalib, and so forth, or the work of Al-Quran translation, using local languages [2, p.230-231]. There are no many research done on translation into Indonesian language and local languages other than the research carried out

by Istianah on criticism against Al-Quran translation by the Ministry of Religious Affairs of the Republic of Indonesia [3], research by Kiptiyah on the tradition of writing of Al-Quran exegesis in Javanese language [4], research by Saifuddin on Al-Quran translation into Javanese language in the philological review [5], and research by Haris on his records on Al-Quran translation in Sasak language (Lombok) [6].

Related to Al-Quran computation, the research by Safeena and Kammani mention that the computational linguistic approach expanding to Al-Quran digital text thus far uses the techniques of *Natural Language Processing*, *Treebank*, and *Syntactic Annotation*. This is concluded from their observation on a number of literatures on Al-Quran computation and several models of digital data and Al-Quran data processing application [7]. In the context of digital data product in the form of Arabic and Al-Quran corpus, Eddakrouri [8] identified that to this date there are 18 Arabic language corpus in the form of digital text materials, but there is no information on Al-Quran corpus. However, there are also 19 website-based Arabic language corpus, four of which is Al-Quran corpus. The existing Al-QuranCorpus is the Al-Quran corpus with annotation without translation texts and there is no at all Al-Quran parallel corpus.

There is a source which presents special Al-Quran corpus namely the website of tanzil.net [9]. In this website, Al-Quran translation corpus is available in various languages, including Indonesian language, but only in the form of translation texts, and no Al-Quran translation parallel corpus [10]. Another source on Al-Quran corpus and its translation in English is in the websitecorpus.quran.com. English language translation for each verse is taken from seven works of Al-Quran translation in English and positioned in parallel [11]. There is an interesting finding on digital Al-Quran application for cell phones available in the Google Play Store. This application provides Mushaf translation in Indonesian language, international languages and local languages. There are 11 versions of local language translation for image-format Mushaf translation, namely Jawa Banyumasan, Mongondow, Kaili, Toraja, Makassar, Dayak Kanayatn, Minang, Batak Angkola, Sasak, Aceh, and Sundanese [12].

From a number of literatures and existing corpus models, there is no as yet a model of Al-Quran paralel corpus model and its translation as desired in this research plan.

4 Result and discussion

The latest dynamic in the revision process of Al-Quran translation into Indonesian language commenced as from year 2002. The result of revision has widely circulated and is used in the printing of Al-quran Translation circulating in Indonesia. However, based on inputs from various parties, the subsequent revision process was done since 2016 more comprehensively covering the aspect of language, translation consistency, substance, and format of preparation systematics [2, p.230]. [2, p.230]. In the prototype of Al-Quran parallel corpus that we have made, the text of language translation that we use are still from the text in the revised edition of year 2002. Based on the above situation, we consider making further prototype using Indonesia translation texts as the result of the latest revision. This is important in the context of presenting a number of corpus variants for a wider interests. These two corpus can be used as the object of research on Al-Quran translation and its development diachronically based on the material from cross-time data.

In order to enrich the corpus, it should be noted that the translation process of Al-Quran into local languages in the Achipelago has actually been quite active. Thus far, the translation process and publishing of 12 Mushaf of Al-Quran in local languages have completed, namely

translation in the languages of Makassar, Bolaang Mongondow, Sasak, Ambon, Bali, Batak Angkola, Padang, (Javanese) Banyumas, Banjarmasin, Toraja, Kaili, and Dayak Kayanatn. There are two translated Mushafs in two other languages that are currently being processed namely in Acehnese, Madurese, and Bugis languages, while in two other languages are still under the planning stage, namely in Palembang and Sundanese languages(based on personal telephone contact with Salim, Staff member at the Center for Research and Development of Lectures on Religious Richness and Organizational Management, the Research, Development, Education and Training Agency of the Ministry of Religious Affairs of the Republic of Indonesia, June 8, 2018). In view of large number of languages throughout Indonesia, the translation process of Al-Quran into local languages will continue and reach other local languages, particularly major ethnic languages which are the main root of more varied sub ethnical languages. Hence, Indonesia potencially have and present unique cultural products in the form of Al-Quran translation in various ethnical languages not the national language in the country with the most languages. This is highly possible and will certainly become the country's richness with a very valuable culture and religion tradition.

The availability of Al-Quran Mushaf in translation in the above various local languages opens a room for further research on the richness of comparative Arabic language and local languages in the context of dual languages or multi languages if the reseach to be conducted involves more than two languages in the context of inter-language contrast analysis. This must be seen as huge potency and at the same time an interesting challenge for research and study on language, translation, and also culture in a very wide coverage. Although there has been available the printed version of the Mushaf and digital version in the form of image with per page display of Mushaf, to support research on computational linguistics and corpus linguistics, the existing format still must be changed in this matter to become a pure text.

Further, slightly different from the format of regular text corpus, for the purpose of translation studies, the corpus used is parallel corpus which contains texts in the source language and the target language in two columns side by side and made using the application of spreadsheet processing such as Microsoft Excel and the like. The file format is in the form of *.xlsx, *.jmx, *.xml, *.xliff, and so forth. In this initial stage of experiment, we attempt to prepare a corpus containing data in two languages and is arranged in a two-column table. This is the simplest model intended for novice users.

This corpus is made using the format of two-column table side by side, on the first line of each column of which contains the name of language to become the source language and the target language side by side, in this case Arabic on the left column and Indonesian language on the right column. The more the variant of the target language, the more number of column it will have. Next, we fill in the left column with the Al-Quran text in Arabic that we took from the tanzil.net [13]. We use the text "Quran Simple Clean" containing Al-Quran texts in standard Arabic writing, not the Utsmani version, and without diacritic signs. This selection is taken to anticipate technical constraints in the process of processing and searching data at the time using the menu available in the application found in the Sketch Engine website. The text of Al-Quran translation in Indonesian language is directly obtained from *Lajnah Pentashihan Mushaf Al-Quran* of the Ministry of Religious Affairs of the Republic ofIndonesia through direct correspondence. The following is the result of application of the mechanism in the corpus prototype that we make as seen in figure 1.

Figure 1. Example of composition of table on content of Al-Quran
parallel corpus and its translation in Indonesian language

As seen in figure 1, the numbering of each verse in Arabic and its translation is made with two types of number. The first number ended with the punctuation mark of period (.) marks the sequence number of Surah in the Al-Quran and the second number also ended with the period mark (.)— is to separate it from the text of the verse or translation — marking the sequence number of the verse in the relevant Surah.

1. Users must Sign in to register their account or Log in if they already have an account in the Sketch Engine.
2. Users download the corpus file they made under the format *.tmx or *.xls. In this case we download the file under the format *.xls made using the spreadsheet processing application of Microsoft Excel.
3. Give a name to the corpus to be uploaded to the user's personal account.
4. Make sure the language of corpus content uploaded is accurately recognized by the application, then click "Create".
5. The process of uploading and making corpus in the account completes in this step.

Upon completion of uploading, the display of the account owner's page that contains the corpus made is as seen in figure 2 below.
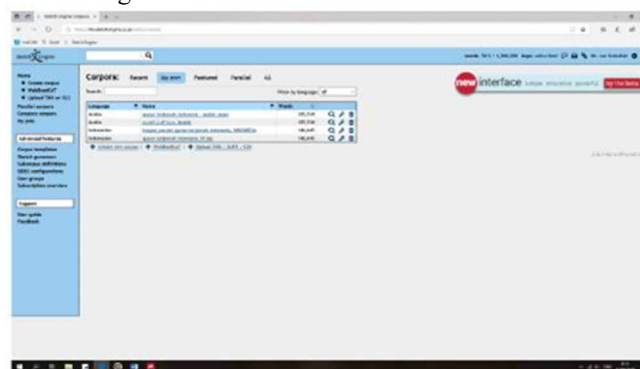
Figure 2. Example of the front page of the user's account
after successful completion of corpus uploading

On the front page of the user's account, there seen several files stored under different names. We try to do experiment on two names each for searching words in Arabic and words in Indonesian language, each of which is marked with the word "Arabic" and "Indonesian" in

accordance with the header of each column in the uploaded corpus data. This is important to ensure that the source of searching against the target language, in this case Indonesian language is from the right column.

After the corpus is uploaded in the account, the user may perform desired searching process. The following is an example of searching wordحسنة/ḥasanat/ and its translation in 'good' Indonesian language from the intended corpus.

Figure 3. Example of application page for presenting corpus
with the menu forsearching the word حسنة/ḥasanat/ and its translation in 'good' Indonesian
Language

Option for searching in the menu as seen in figure 3 is quite varied and covers the projection of unit to be sought, whether a word, phrase, even character and lexical unit of special grammatical code (CQL). In this regard, we try to search the word حسنة/ḥasanat/ in the form of character to anticipate wider coverage against combined emergence of the word in the text of verses in Arabic writing. The option for searching translated word or its equivalence in Indonesian language is directed to the word "good".

After completion of searching, the application will display the result of searches by placing the word sought in the middle of the sentence containing it and mark it with red color. Hence, the user will find it easy to find colocation of words related to the word حسنة/ḥasanat/ and the word "goodness", as seen in figure 4 below.

Figure 4. Example of standard page display for presenting
the result of search of word قنسح/ḥasanat/ and its translation in 'good' Indonesian language

To analyze collocation, the user only needs to determine the extent of coverage of the word to the left or to the right from the word or query being sought then formulate the result of analysis. If the limitation still needs to be expanded, the user only needs to change the format of result display to the mode "Sentence" to display complete sentences which includes the word or query sought as displayed in figure 5 below

Figure 5. Example of display of presentation page of the result of searching by displaying complete sentences which include the sought word

In the display in the "Sentence" mode as in figure 5 above, the user can no longer see the found word in the middle column from the series of searching result, but the position of the word has been random in accordance with the original composition in the relevant sentence. However, complete sentence display as such can help broader analysis at the level of discourse by observing the whole part of the sentence in its relation with the sought word.

The Indonesian language text uses standard characters commonly used, unlike translation of Al-Quran into local languages that have special characters or even non-Latin alphabets. Translations into Banyumasan Javanese, for example, there are a number of special characters that are not commonly used in the applicable standard writing, such as the character "é" and others. Through Digital Al-Quran application for cellular phones [12], seen from the figures 6, there seems to be other problems in the conversion of special character to the format of characters, such as appearance of character "Ø"and the imperfect writing of the name of Surah Al-Fatihah.

Figure 6. Example of display of presentation page of the result of searching
by displaying complete sentences which include the sought word

In common cases, as seen in figure 6 above, loss or change of characters in the conversion process is caused by difference in the inter-computer operating system which processes the making of application or technical problems in the system language available or installed. Consequently, if a certain character from a text with a language is not avaialble in the computer operating system of the data processor, there will occur mis-conversion and then appear a technical problem that may have serious impact on the text display or content as a whole

More complex problems will be found if the text of the target language still uses non-Latin alphabets as in the case of AlQuran translation in Makassar language as in figure 7.

Figure 7. Example of display of presentation page of the result of searching
by displaying complete sentences which includes the sought word

Figure 7 shows that the alphabet used in the Makassar language translation is the Lontara alphabet from Bugis region, South Sulawesi. For the interest of preparing Al-Quran parallel corpus and its translation, we see the need for conversion of writing in that alphabet to Latin writing using the transliteration system. This will certainly need further technical handling. However, if the process is completed and successfully applied, similar model can also be applied in the case of other local languages using non-Latin alphabet. Hence, the process of preparing Al-Qurancorpus and its translation into local languages in the Archipelago can be continued and will become an important contribution to the richness of the corpus with its typical and unique characteristics.

4   Result and discussion

It can be concluded that technically it is highly possible to prepare an Al-Quran parallel corpus and its translations in Indonesian language and local languages in the Archipelago. The availability of translated Mushaf as existing is an important capital for the process to the next stage namely digitalization of texts and materials to prepare standard corpus format or parallel corpus. In the future time, the existence of this parallel corpus will become an important contribution to the richness of cross-sciences in Indonesia and is very urgent for the projection

of development of Al-Quran translation studies in particular and multidisciplinary language studies in Indonesia.

References

[1]   E. S. Baihaki, "Penerjemahan Al-Qur'an: Proses Penerjemahan Al-Qur'an di Indonesia," *J. Ushuluddin*, vol. 25, no. 1, p. 44, 2017.

[2]   J. Syatri, A. Akbar, A. Hakim, A. Jaeni, and M. Musadad, "SIKAP DAN PANDANGAN MASYARAKAT TERHADAP TERJEMAHAN ALQUR'AN," *Suhuf J. Pengkaj. Al-Qur'an dan Budaya*, vol. 10, no. 2, 2017.

[3]   Istianah, "Fenomena Alih Bahasa Al-Qur'an," *Suhuf J. Pengkaj. AlQur'an dan Budaya*, vol. 8, no. 2, pp. 203–231, 2015.

[4]   S. M. Kiptiyah, "Tradisi Penulisan Tafsir Al- Qur ' an Bahasa Jawa," *Lekt. Keagamaan*, vol. 15, no. 2, pp. 420–445, 2017.

[5]   Saifuddin, "Tradisi Penerjemahan Al-Qur'an ke dalam Bahasa Jawa Suatu Pendekatan Filologis," *Suhuf J. Pengkaj. Al-Qur'an dan Budaya*, vol. 6, no. 2, pp. 225–248, 2013.

[6]   T. Haris, "Tinjauan Buku AL-QUR'AN DAN TERJEMAHNYA BAHASA SASAK Beberapa Catatan," *Suhuf J. Pengkaj. Al-Qur'an dan Budaya*, vol. 10, no. 1, pp. 211–226, 2017.

[7]   R. Safeena and A. Kammani, "Quranic Computation: A Review of Research and Application," *Proc. - 2013 Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Its Sci. NOORIC 2013*, no. September 2014, pp. 203– 208, 2015.

[8]   A. Eddakrouri, "Arabic corpora." [Online]. Available: https://sites.google.com/a/aucegypt.edu/infoguistics/directory/CorpusLinguistics/arabic-corpora. [Accessed: 11-Jun-2018].

[9]   H. Zarrabi-Zadeh, "Tanzil Documents." [Online]. Available: http://tanzil.net/docs/home. [Accessed: 11-Jun-2018].

[10]   H. Zarrabi-Zadeh, "Quran Translations," 2018. [Online]. Available: http://tanzil.net/trans/. [Accessed: 11-Jun-2018].

[11]   K. Dukes, "Verse (1:1) - English Translation." [Online]. Available: http://corpus.quran.com/translation.jsp?chapter=1&verse=1. [Accessed: 11-Jun-2018].

[12]   PartaiNasDem.id, "Al-Quran Digital." 2018

[13]   H. Zarrabi-Zadeh, "Download Quran Text," 2018. [Online]. Available: http://tanzil.net/download. [Accessed: 11-Jun-2018].

[14]   "Parallel corpus from tabular data (basic user)." [Online]. Available: https://www.sketchengine.eu/user-guide/user-manual/corpora/setting-upparallel-corpora/#tab-id-1. [Accessed: 08-Jun-2018].