Quranic Corpus Models for Corpus-Based Studies

Nur Hizbullah*, Abdul Mutaali
Department of Linguistics
University of Indonesia
Depok, Indonesia
*hizbi77@gmail.com, moetaalingua@gmail.com

Abstract—**The advanced development of computing technology covers various religious texts, such as Quran. Through computing technology, the Quran is formatted digitally in the forms of images and texts. The Quranic format in the form of digital texts, furthermore, is transformed into a Quranic corpus used widely as the material and object of Quranic studies in various fields by using the corpus linguistic approach. Although Quranic corpora are extensively available and widely used, introduction to corpora and its extensive use need to be done in Indonesia. This article will descriptively discuss various Quranic corpus models that can be utilized for the interest of corpus-based research about the Al-Quran. This study found that there are several models and formats of monolingual Quranic corpora that can be processed using certain applications. However, bilingual or even multilingual content of Quranic corpora in a parallel corpus format have not found yet specifically. This study concludes that on one hand the existing monolingual Quranic corpora have already been adequate to be applied in the Quranic corpus studies. However, on the other hand, the absence of parallel Quranic corpus models gives an opportunity for the research and trials in preparing the corpus so that it can be applied to develop Quranic corpus studies.**

Keywords—corpus-based studies; corpus model; Quranic corpus; Quran parallel corpus; Quranic studies

I. Introduction

The development of modern computing technology has a huge impact on various texts in the form of conversion of text forms from conventional to digital. Such a change also applies to religious texts, for example the Quran, Islam’s holy book. The appearance of new formats of Quran texts from conventional printed text to digital text serves as the basis for appearance of various modifications of text forms that are now more specific, particularly in relation to the purpose of utilizing the texts for further studies or other purposes.

The digital format of Quran texts is in general divided into two forms, namely image and text formats. Image format generally derives from the master design of pre-printed pages or Quran printed pages. These pages, after that, are scanned and processed in such a way to be subsequently used as the materials for producing digital Quran application. Such an application can be largely found in various internet sources to be later installed directly onto the compatible smart phone or on computer. Another quite dynamically growing digital Quran format is text format. The very flexible process of multi-format conversion enables the transformation of Quranic digital texts into various derivative forms for several purposes. One of the purposes that is now rapidly developing is to build Quranic corpus. Digital Quranic text corpus is now made with various modifications and has been widely utilized in the development of Quranic studies using the multidisciplinary approaches.

The usage of Quranic corpus in numerous multidisciplinary studies has actually been widespread. The presence of corpus in the content of research actually creates a trend or new model of studies, namely corpus-based studies. In Indonesia, however, the sub-discipline of corpus linguistics is still a new phenomenon that is not yet widely known, and so are corpus-based studies. This means that attention and interest in the development of digital data, including digital Quranic data, still need to be developed and strengthened, especially in digital Quranic corpora studies. The usage of digital Quranic corpus, however, is not limited only to those who study Islam or the Quran. As a language text, the Quran can also be studied from the perspectives of linguistics, literature and translation. This means that the studies in those fields can be developed in the scope of corpus. This kind of research has been lately supported with various corpus-processing applications providing a number of corpus processing facilities that facilitate searches, tracking up to analysis of certain parts or the entire data in a corpus. Various Quranic corpora actually turn out to be very interesting to be studied and developed.

If a researcher is smart enough in observing, studying and using an Arabic corpus, they would surely realize that there are challenges and also opportunities in the development of existing corpus models. Therefore, it is significant to encourage Quran researchers to study existing Quranic corpus models that will, afterwards, trigger new ideas to make various Quranic corpus improvements. The development of corpus model will depend on to what extent Quran researchers are able to evaluate existing models and whether there are definite needs of Quran research in Indonesia. Nevertheless, the Quran corpus models development will require collaborations with the information technology and computing. This becomes challenge and opportunity for the development of multidisciplinary fields of studies in Indonesia.

This study refers to several literatures on corpora and their utilization for research and several Quranic corpora as the model used as object of the data in Quranic corpus studies.

In the bilingual contexts of English and Arabic, ideas about parallel corpus have been discussed by Al-Ajmi [1]. Al-Ajmi maintains the importance of a parallel corpus for two languages, English and Arabic, as the foundation for preparing a modern dictionary and covering the creation of lexicographical application that needs to be developed in future. Al-Ajmi proposes that this parallel corpus contains texts in English as the source text covering various topics, such as the environment, globalization, psychology, history, politics, drama, and so forth. Its translated text in Arabic can be collected from several similar scientific literatures [1]. In line with the idea, the development of digital Quranic data is recorded by Sharaf who identified initial preparation of Quranic digital data model in several websites: tanzil.info, islamicity.com/QuranSearch/, and a digital Quran translation model in quod.lib.umich.edu/k/koran [2]. Dukes, furthermore, elaborated a new source of Quranic Arabic Corpus as the initial works of annotated Quran model in corpus.quran.com [3]. In the Middle East, a large Arabic corpus model also includes the Al-Quran as one of its categories, namely the King Saud Classical Corpus of Arabic (KSUCCA) [4]. Alrabia (et.al) subsequently made the corpus as the object of data for an empirical analysis on modeling of word meaning that describes distributional semantical model from certain concepts in the Quran and Classical Arabic in general [5]. An equally important contribution is also shown by Imad and Abdelhak who prepared a Quranic corpus model enriched with linguistic annotation based on the Arabic grammatical theory of Al-Khalil. This corpus is dedicated for the development of studies in the field of Natural Language Processing [6].

With regard to the models of Quranic corpus used as reference in this review, there are three sources that provide Quranic corpus selected as the representation of similar models. The first corpus is digital Quranic text corpus without annotations found in the website tanzil.net. This corpus has two styles of Arabic orthography, namely the Uthmani orthography and Imla’i orthography with their respective variants [7]. The next corpus is annotated Quranic corpus available in the website sketchengine.co.uk. This corpus has two kinds of orthography, namely Arabic and Latin. Each has version with diacritical marks (vowelled) and version without diacritical marks (unvowelled [8]. The other corpus is the Quranic corpus available in the website arabicorpus.byu.edu. [9].

II. Objectives

This study aims to provide an overview of the Al-Quran corpus model along with corpus processing content and available features in the accessible applications. After corpus exploration and its application, it is expected that there will be findings or opportunities for the development of an existing corpus model or the preparation of a new model of the Al-Quran and its Indonesian translation that can be used in the development of corpus-based Al-Quran studies in Indonesia context.

III. Methodology

This study is library research that uses explorative approach with descriptive method. The descriptive method is used to describe the characteristics of the three models of Quranic corpus mentioned in the previous review. This description relates to the characteristics of each corpus along with the availability of facilities to search content and process the corpus found through observation to the website providing the corpus. The description is continued with projected utilization of the corpus in the context of Quranic corpus studies.

IV. Results and Discussion

Based on the characteristics of its contents and sources, there are two types of corpus:

- Pure text data corpus without annotation from a website without corpus processing application, like in tanzil.net; it needs a separate application to process the corpus, such as WordSmith, etc.
- Annotated text data corpus available in the websites completed with its processing application, for example: arabiCorpus (arabicorpus.byu.edu), Quran annotated corpus (sketchengine.eu/quran-annotated-corpus), and Quranic Arabic Corpus (corpus.quran.com).

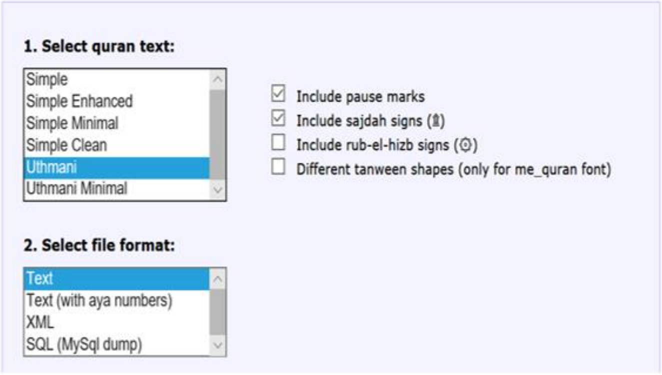
The following is the results of observation and analysis of the corpus models used as samples.

A. Quran Corpus From Tanzil.net [7]

This corpus is included in the first type of corpus based on the above division. The corpus that only contains materials of digital Quranic texts without annotations is divided into two formats based on their orthography, namely:

- Uthmani script, namely a classic model script, is used as standard orthography for Al Quran inscriptions since the era of Caliph Uthman. At present this script is maintained in the Quranic Mushaf known as the Medina Mushaf.
- Simple script, also known as the Imla’I script, that refers to modern Arabic orthography based on the principles of standard Arabic grammar.

Each of the text has sub-type that is distinguished from completion of diacritical marks on the relevant text. Zadeh divides it under the term “plain” for the texts equipped with diacritical marks, “minimum” for the texts only contain basic diacritic marks, and “clean” for the texts which contain no diacritical marks at all [7]. In addition to the type of its orthography, the text of Quranic corpus in this website is also distinguished from its formats. There are specific text formats that are commonly used for the purpose of basic research and a file format, especially made for programming purposes. This division can be viewed in the excerpt from the said corpus source web page as depicted in figure 1.



1. Select quran text:

- Simple
- Simple Enhanced
- Simple Minimal
- Simple Clean
- Uthmani
- Uthmani Minimal

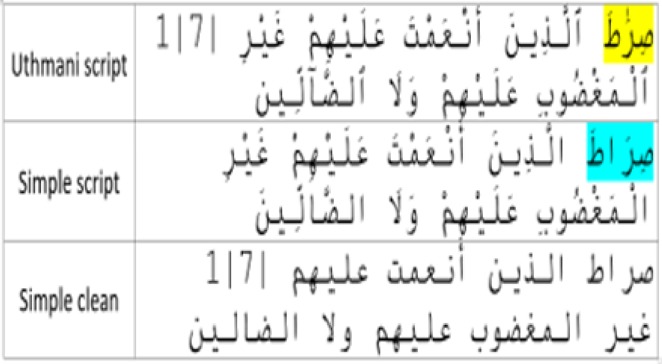
☑ Include pause marks
☑ Include sajdah signs (۩)
☐ Include rub-el-hizb signs (۞)
☐ Different tanween shapes (only for me_quran font)

2. Select file format:

- Text
- Text (with aya numbers)
- XML
- SQL (MySql dump)

Fig. 1. Types of Quranic corpus text in the website tanzil.net.

Furthermore, figure 2 below shows the samples of the differences between the Uthmani orthography and Imla’i orthography and differences in texts that contain diacritical marks or texts without diacritical marks.



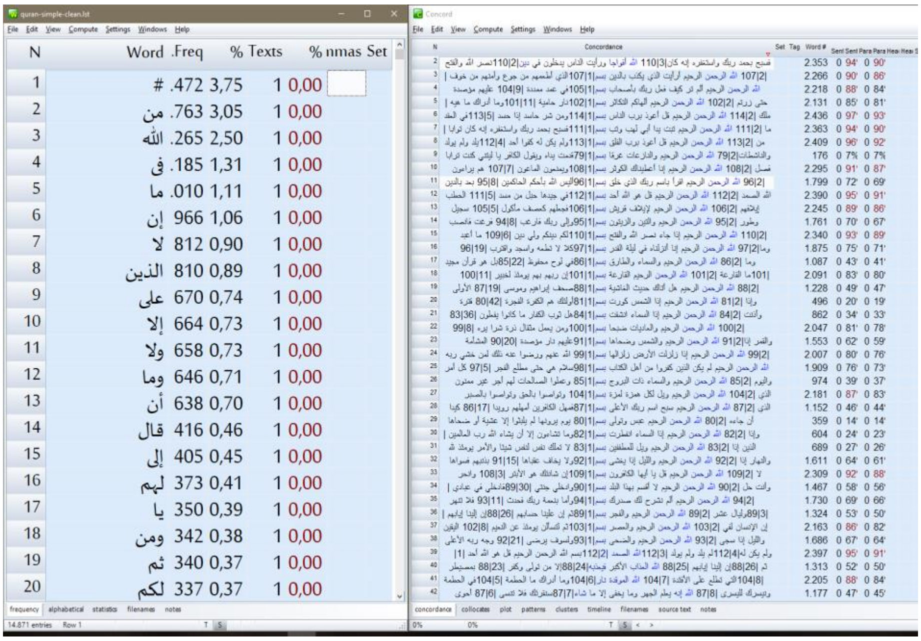
Uthmani script	صِرَٰطَ ٱلَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ 1|7| ٱلْمَغْضُوبِ عَلَيْهِمْ وَلَا ٱلضَّآلِّينَ
Simple script	صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ الْمَغْضُوبِ عَلَيْهِمْ وَلَا الضَّالِّينَ
Simple clean	صراط الذين أنعمت عليهم 1|7| غير المغضوب عليهم ولا الضالين

Fig. 2. Sample difference of orthographical variants of Uthmani script and Imla’i script in the case of word صراط /ṣirāṭ/ ‘road’.

One of differences seen in the two texts in figure 2 is, for example in the word alif, that marks long vocal appearing in the word صراط /ṣirāṭ/ ‘road’ of the Imla’i orthographic version that appears in the form of diacritical mark of upright fatḥa on the orthography of the Uthmani version. The text of third line (simple clean) shows the examples of text of Quranic verses without any diacritical marks at all.

The available digital Quranic corpus in this website can be downloaded and utilized by users independently. However, as the website tanzil.net does not provide corpus processing application, the users need to use other corpus processing applications that can properly process Arabic texts. One of quite representative processing applications is WordSmith [10].

Specifically, the applications to process corpus are actually widely accessible and can be selected depending on the availability of menus in their application. Some can be downloaded for free, such as Nooj, TextStat, MonoconcEsy, Aconcord, and AntCont, and some others are charged, such as WordSmith. However, of such many applications, WordSmith is considered as the most adequate one because of its technical capacity to read and process Arabic texts that have many different characteristics from the texts of other languages. Specifically, this application requires the corpus file that can be processed to be first converted, usually the files written with Microsoft Word applications with the extension of *.doc or *.docx. This format must be converted with the encoding method that the conversion uses UTF-8 and is saved under the extensions *.txt. Technically, only file with the extension of *.txt and the conversion method of UTF-8 can be processed using this application. The example of corpus analysis with the application of WordSmith is illustrated in figure 3.



quran-simple-clean.lst

N	Word	.Freq	%	Texts	% nmas	Set
1	#	.472	3,75	1	0,00	
2	من	.763	3,05	1	0,00	
3	الله	.265	2,50	1	0,00	
4	في	.185	1,31	1	0,00	
5	ما	.010	1,11	1	0,00	
6	إن	966	1,06	1	0,00	
7	لا	812	0,90	1	0,00	
8	الذين	810	0,89	1	0,00	
9	على	670	0,74	1	0,00	
10	إلا	664	0,73	1	0,00	
11	ولا	658	0,73	1	0,00	
12	وما	646	0,71	1	0,00	
13	أن	638	0,70	1	0,00	
14	قال	416	0,46	1	0,00	
15	إلى	405	0,45	1	0,00	
16	لهم	373	0,41	1	0,00	
17	يا	350	0,39	1	0,00	
18	ومن	342	0,38	1	0,00	
19	ثم	340	0,37	1	0,00	
20	لكم	337	0,37	1	0,00	

14.871 entries Row 1

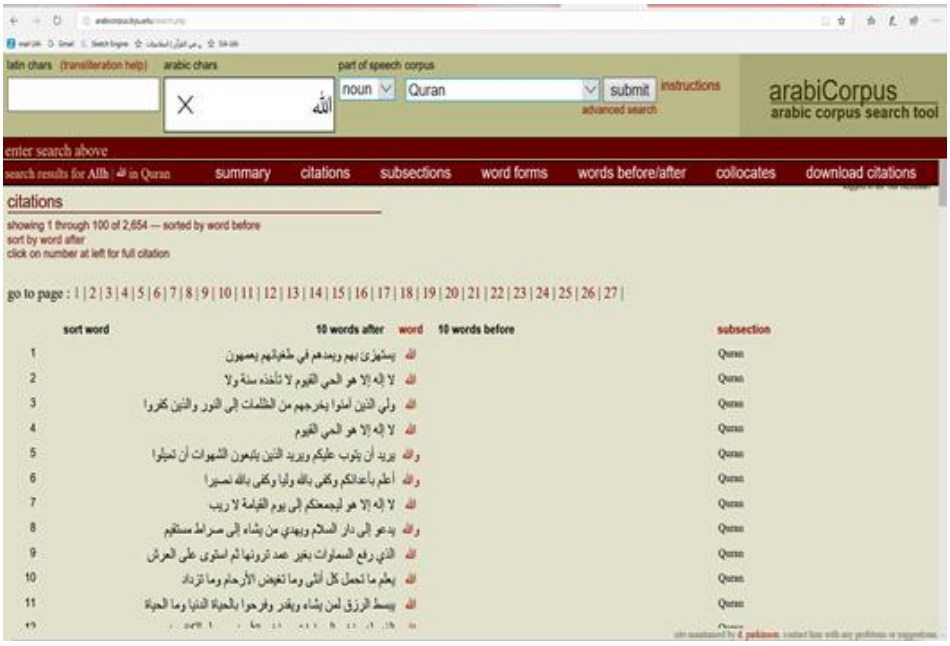
File Edit View Compute Settings Windows Help
frequency alphabetical statistics filenames notes
Concord
Concordance Set Tag Word # Sent Sent Para Para Head Head
concordance collocates plot patterns clusters timeline filenames source text notes

Fig. 3. Example of result of analysis on words, word frequency, and concordance.

In figure 3, the left side shows the samples of results of searching of words and their frequency, while the right side \ shows the result of concordance searched for the word basmalah in Arabic scripts. This mechanism can be used among others as the basis for morphological, syntax and semantic analysis, and a discourse in the field of linguistic studies.

B. *Arabicorpus (arabicorpus.byu.edu) [9]*

Different from the first type of corpus, the corpus found in this website, including Quranic corpus, is placed integrally with the application that provides the feature of corpus search and processing. Corpus search and processing are featured with the method of word input in two types of characters, Latin and Arabic characters, searched based on word class and selected corpus. The search result, then, will be displayed in various menus, namely summary, citations, subsections, word forms, words before/after, and collocations.



citations

showing 1 through 100 of 2,654 — sorted by word before
sort by word after
click on number at left for full citation

go to page : 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |

	sort word	10 words after	word	10 words before	subsection
1		يستهزئ بهم ويمدهم في طغيانهم يعمهون	الله		Quran
2		لا إله إلا هو الحي القيوم لا تأخذه سنة ولا	الله		Quran
3		ولي الذين آمنوا يخرجهم من الظلمات إلى النور والذين كفروا	الله		Quran
4		لا إله إلا هو الحي القيوم	الله		Quran
5		يريد أن يتوب عليكم ويريد الذين يتبعون الشهوات أن تميلوا	والله		Quran
6		أعلم بأعدائكم وكفى بالله وليا وكفى بالله نصيرا	والله		Quran
7		لا إله إلا هو ليجمعنكم إلى يوم القيامة لا ريب	الله		Quran
8		يدعو إلى دار السلام ويهدي من يشاء إلى صراط مستقيم	والله		Quran
9		الذي رفع السموات بغير عمد ترونها ثم استوى على العرش	الله		Quran
10		يعلم ما تحمل كل أنثى وما تغيض الأرحام وما تزداد	الله		Quran
11		يبسط الرزق لمن يشاء ويقدر وفرحوا بالحياة الدنيا وما الحياة	الله		Quran
12		[illegible]	[illegible]		Quran

site maintained by d. parkinson. contact him with any problems or suggestions.

Fig. 4. Example of result of analysis on the word "Allah" in Quranic corpus in the website Arabicorpus.

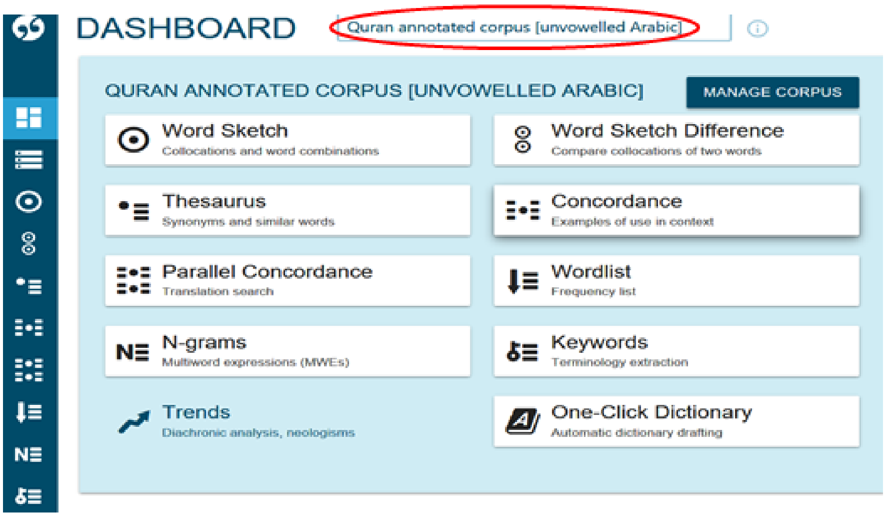
Figure 4 shows the examples of search result of the concordance of the word ﷲ /Allāh/ 'Allah' by displaying a series of ten words after the relevant word. Such a capacity in the application existing in the website is comparable to the capacity in the previous WordSmith application that can also assist researchers in conducting analysis on the level of morphology, syntaxes, semantic, up to discourse in the context of linguistics.

Systematically, a difference is found from the previous website, namely tanzil.net. In this website, all corpora, including Quranic corpus, are placed integrally with the feature of the corpus search and processing. In this website, however, the file of Quranic corpus cannot be accessed separately and can only be used by searching certain lexical units based on word class. Therefore, complete feature of list of words in the Al-Quran is not available. On the other side, there is no special information given by this website on the orthographic version of the Quran texts, whereas to use classical Uthmani orthography or the Imla'i orthography which is in conformity with standard Arabic grammar.

C. *Quran Annotated Corpus (sketchengine.eu) [8]*

Based on the information from its website, this corpus was prepared by Alqassem by using data from the Quranic Arabic Corpus [11] and the QurAna Anaphoric Conference Database [12] that has been lemmatized and tagged with POS tagging [8].

Quranic corpus in this website includes annotated corpus and consists of two types of characters, namely Arabic characters and Latin characters, each completed with the voweled version (with diacritical marks) and unvoweled (without diacritical marks). As in the previous website, Quranic corpus is placed integrally with the application existing in this website. However, the feature for processing corpus in the Sketch Engine website is more complete and adequate. In addition, the registered users of this application will obtain facilities in the form of space for storing data if the users need to make his/her own corpus. Figure 5 below shows the features to process the corpus in the Sketch Engine website.



DASHBOARD

Quran annotated corpus [unvowelled Arabic]

QURAN ANNOTATED CORPUS [UNVOWELLED ARABIC]

MANAGE CORPUS

- Word Sketch – Collocations and word combinations
- Word Sketch Difference – Compare collocations of two words
- Thesaurus – Synonyms and similar words
- Concordance – Examples of use in context
- Parallel Concordance – Translation search
- Wordlist – Frequency list
- N-grams – Multiword expressions (MWEs)
- Keywords – Terminology extraction
- Trends – Diachronic analysis, neologisms
- One-Click Dictionary – Automatic dictionary drafting

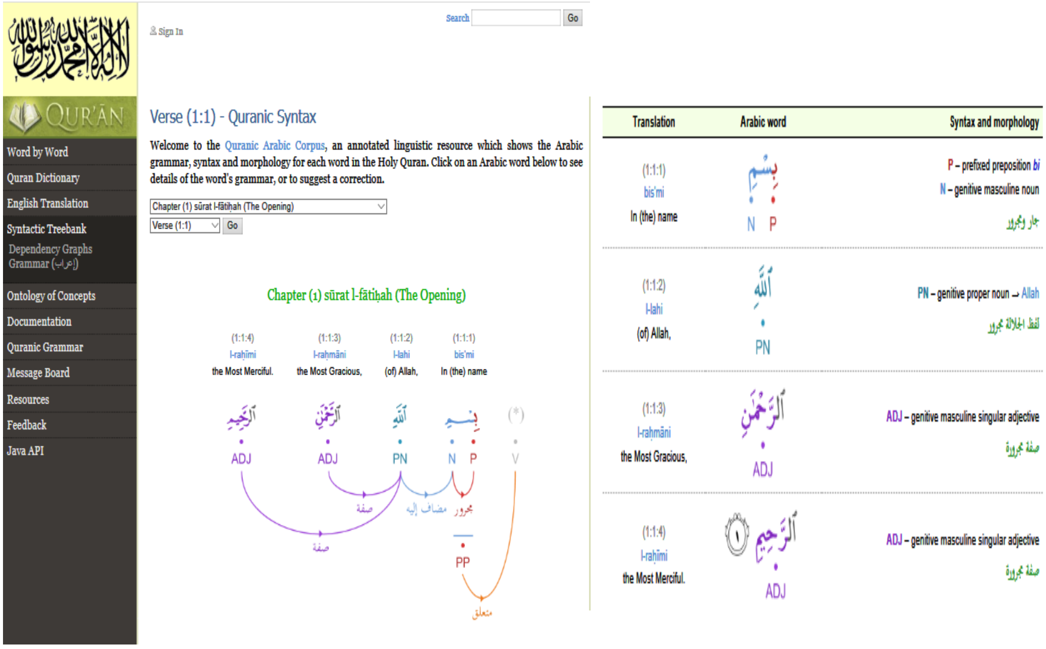
Fig. 5. Front page of charge users of Sketch Engine that provides complete features for processing corpus with the corpus option of "Quran annotated corpus (unvowelled Arabic)".

Although Sketch Engine provides more complete system and features, it does not give access to users to download the files of Quranic corpus contained in the website. Users may only use the corpus files for the purpose of searching in accordance with the available features of corpus processing. There is no information on the relevant orthographic version of Quran texts, either Uthmani orthography or Imla'i orthography.

Completed corpus processing features in the website Sketch Engine enables utilization of the result of data analysis for the purpose of broader linguistic studies. Even, the parallel corpus processing feature owned by Sketch Engine enables the users and researchers to not only perform studies in the field of linguistics in general, but also studies in the field of semantics and translation specifically.

D. Quran Arabic Corpus (corpus.quran.com) [11]

Compared to existing corpus models, Quranic corpus in this website is quite unique with the concept of content and display. From the aspect of content, this website contains translation of the Al-Quran word by word, Quranic dictionary, English translation, syntactic Treebank, ontology of concepts in the Quran, and Quran grammars. Quranic texts in this website are displayed in the form of figures (not texts) equipped with various information in morphology and syntaxes related to each verse along with word by word translation in English. The orthography used in this Quranic text is Uthmani orthography [11]. The following is example of displays of website pages that contain the details of linguistic explanation of the first verse from surah Al-Fatiha (1:1), namely the pronunciation of basmalah as indicated in figure 6.



Verse (1:1) - Quranic Syntax

Welcome to the Quranic Arabic Corpus, an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. Click on an Arabic word below to see details of the word's grammar, or to suggest a correction.

Chapter (1) sūrat l-fātiḥah (The Opening)
Verse (1:1) Go

Chapter (1) sūrat l-fātiḥah (The Opening)

Translation	Arabic word	Syntax and morphology
(1:1:1) bis'mi In (the) name	بِسْمِ N P	P – prefixed preposition *bi* N – genitive masculine noun جار ومجرور
(1:1:2) l-lahi (of) Allah,	ٱللَّهِ PN	PN – genitive proper noun → Allah لفظ الجلالة مجرور
(1:1:3) l-raḥmāni the Most Gracious,	ٱلرَّحْمَٰنِ ADJ	ADJ – genitive masculine singular adjective صفة مجرورة
(1:1:4) l-raḥīmi the Most Merciful.	ٱلرَّحِيمِ ١ ADJ	ADJ – genitive masculine singular adjective صفة مجرورة

Fig. 6. Display of Quranic Syntax page on Surah Al-Fatiha verse 1 (1:1) in the website corpus.quran.com.

Users and researchers who utilize Quranic corpus in this website can only search words, verses, or concepts in accordance with the available content in the website. Hence, the users and researchers cannot download the relevant texts.

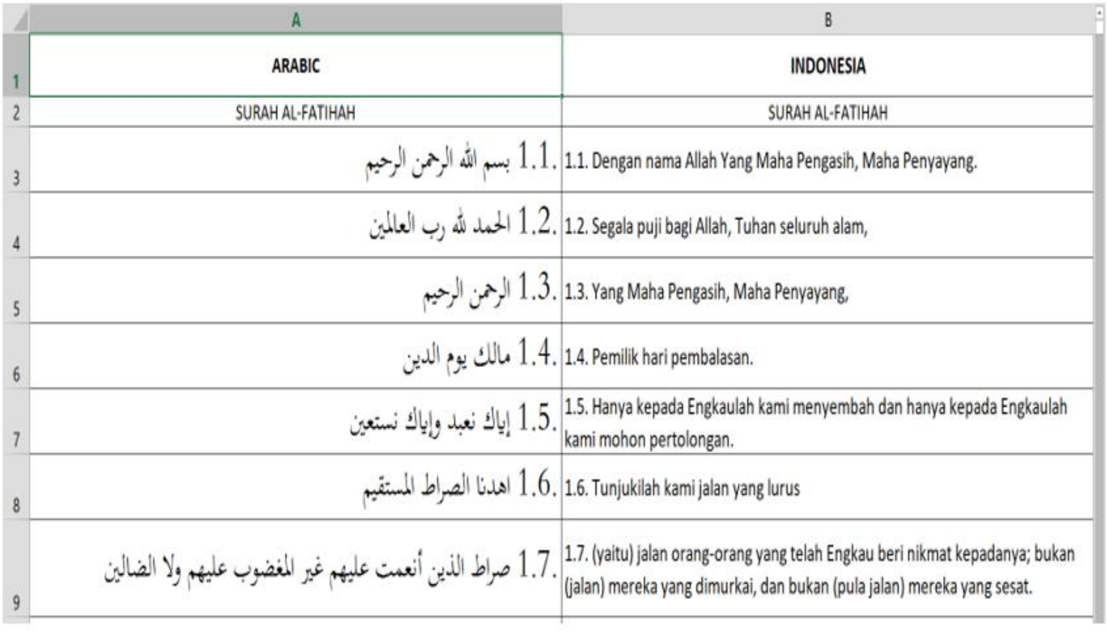
By displaying the content of Quranic corpus as seen in figure 6, this website focuses more on the information about Quranic grammar and translation in English. Therefore, the information is essential for Quran studies, particularly in the fields of grammar, morphology, syntax, and translation.

E. Opportunities for Develoment of Quranic Corpus Model

The selection of four models of Quranic corpus to be discussed and analyzed does not only take into account the factor of representativeness and advantage of each model compared to other models, but also observed that the existing four models give an opportunity for further applied studies to develop the models based on the specific needs of Quranic studies in Indonesia.

Related to the text formatted Quranic corpus downloaded from tanzil.net website, Quran researchers actually have opportunities and changes to do experiments in designing similar corpus from the text materials in the document format (*.doc/docx). If tanzil.net website provides Quranic corpus in two orthographic options, namely Uthmani and Imla’i, then there will be an opportunity in Indonesia, for example, to make Quranic corpus with Indonesian–standard Uthmani orthography serving as a reference for writing and publishing Quran Mushaf nationally. The type of texts and orthography are specific texts have circulated in Indonesia and experienced some adjustments in several aspects of writing so that they can be easily read and used by Indonesian Muslims who are, in general, do not speak Arabic. These texts are now available in the form of files under the format of spreadsheet and can be used for various needs. These texts have been officially released by the Lajnah Pentashihan Mushaf Al-Quran (LPMQ) of the Ministry of Religious Affairs of the Republic of Indonesia (interview with Z. Afif, LPMQ staff, on October 3, 2018 in Bandung, West Java, Indonesia). Furthermore, these Quran texts can be first converted to document format (*.doc/docx) by adjusting the composition of writing verse by verse reconverted to a file under special text format (*.txt) processed with corpus applications, such as WordSmith or Sketch Engine.

With regard to the feature for processing multilingual parallel corpus existing in the application of Sketch Engine website, Quran researchers also have the opportunity to make a new model of Quran parallel corpus together with its translation side by side in Indonesian language or local languages. This opportunity is widely open as the Ministry of Religious Affairs of the Republic of Indonesian is now actively publishing the Al-Quran and its translations into local languages, such as Javanese, Sundanese, Minangkabau, Bugis and so forth. Figure 7 below is example of Quran parallel corpus model and its translation that can be made using the spreadsheet processing application.



	A	B
1	ARABIC	INDONESIA
2	SURAH AL-FATIHAH	SURAH AL-FATIHAH
3	1.1. بسم الله الرحمن الرحيم	1.1. Dengan nama Allah Yang Maha Pengasih, Maha Penyayang.
4	1.2. الحمد لله رب العالمين	1.2. Segala puji bagi Allah, Tuhan seluruh alam,
5	1.3. الرحمن الرحيم	1.3. Yang Maha Pengasih, Maha Penyayang,
6	1.4. مالك يوم الدين	1.4. Pemilik hari pembalasan.
7	1.5. إياك نعبد وإياك نستعين	1.5. Hanya kepada Engkaulah kami menyembah dan hanya kepada Engkaulah kami mohon pertolongan.
8	1.6. اهدنا الصراط المستقيم	1.6. Tunjukilah kami jalan yang lurus
9	1.7. صراط الذين أنعمت عليهم غير المغضوب عليهم ولا الضالين	1.7. (yaitu) jalan orang-orang yang telah Engkau beri nikmat kepadanya; bukan (jalan) mereka yang dimurkai, dan bukan (pula jalan) mereka yang sesat.

Fig. 7. Display of sample page of Quranic parallel corpus and its translation in Indonesian language made with the spreadsheet processing application.

Specifically, this format is standardized by the application of Sketch Engine to process multilingual parallel corpus. Hence, preparation of this kind of corpus model must follow the application provisions. However, this opportunity needs to be taken into consideration since this is new and only few Quran researchers in Indonesia who pay attention to it and give an opportunity for the development in this field. The existence of this kind of corpus model is expected to offer new paradigm and methodology for Quranic corpus studies in Indonesia.

The Quranic corpus model found in corpus.quran.com website can further be developed by making an Indonesian language version content. Although Al-Quran and its translation in Indonesian language with the format of word by word translation have been widely circulated, this digital format needs to be constructed by considering inclusion of other contents as existing in the original website. This will facilitate Quran researchers and language researchers that specify Quran as the object of their studies, in terms of the method and technique for searching, processing and analyzing the data.

Although basically proposal for development of Quranic corpus model is not a new thing, the need for the model is real in order to give a new paradigm and orientation toward Quranic corpus studies.

V. Conclusion

This study concludes that there are a number of certain models and formats in Quranic corpus that are quite representative, ready-to-use, and can be processed by using certain applications. Some applications are in the form of independent application and some others are website-based. These corpora and applications are important to know and studied in developing and strengthening corpus-based Quranic studies, especially among Quran researchers in Indonesia. Even though the existing corpus models are quite representative, there are still opportunities widely opened for Quran researchers to further develop the model in accordance with the orientation and their research needs. It is expected that development of the model will give a significant contribution to Quran research and studies in Indonesia.

Acknowledgment

I thank the Directorate or Research and Community Services (DRPM) of Universitas Indonesia that has sponsored this research and publication through the grant scheme of Indexed International Publication for Final Assignments (PITTA) for students of Universitas Indonesia year 2018.

References

[1] H. Al-Ajmi, “A New English ? Arabic Parallel Text Corpus for Lexicographic Applications A New English – Arabic Parallel Text Corpus for Lexicographic Applications,” Lexicos, vol. 14, no. 1, 2004.

[2] A.M. Sharaf, The Qur’an Annotation for Text Mining, 2009. [Online]. Retrieved from: http://textminingthequran.com/papers/firstyear.pdf

[3] K. Dukes, E. Atwell, and N. Habash, “Supervised Collaboration for Syntactic Annotation of Quranic Arabic,” Language Resources and Evaluation, vol. 47, no. 1, pp. 33–62, 2013.

[4] M. Alrabiah, A. Al-Salman, and E. Atwell, “The Design and Construction of the 50 Million Words KSUCCA,” In Proceedings of WACL’2 Second Workshop on Arabic Corpus Linguistics, pp. 5–8, 2013.

[5] M. Alrabiah, N. Alhelewh, A. Al-Salman, and E. Atwell, “An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus,” International Journal of Computational Linguistics (IJCL), vol. 5, no. 1, 2014.

[6] I. Zeroual and A. Lakhouaja, “A new Quranic Corpus rich in morphosyntactical information,” International Journal of Speech Technology, vol. 19, no. 2, pp. 339–346, 2016.

[7] H. Zarrabi-Zadeh, Download Quran Text. Retrieved June 11, 2018, from http://tanzil.net/download

[8] Z. Alqassem, Arabic Corpus of the Quran, 2013. [Online]. Retrieved from https://www.sketchengine.eu/quran-annotated-corpus/.

[9] D. Parkinson, arabiCorpus. [Online]. Retrieved from: http://arabicorpus.byu.edu/.

[10] M. Scott, (n.d.). WordSmith, Lexical Analysis Software and Oxford University Press. [Online]. Retrieved from: https://lexically.net/wordsmith/

[11] K. Dukes, (n.d.) The Quranic Arabic Corpus. [Online]. Retrieved From: http://corpus.quran.com/

[12] A.B.M. Sharaf and E. Atwell, QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In LREC pp. 130-137, 2012.