CHAPTER 1
INTRODUCTION

1.1 Rationale

Understanding the meaning of the text entirely or in part which is written in a document is an essential activity in the process of understanding the document as a whole. Misunderstanding the meaning of a text of a document, especially texts that contain the important words, will lead produce an invalid overall perception of a document.

In this world, millions or more documents are produced and published every day. Not all documents are relevant to read, but a lot of documents that are important for us to understand the correct meaning and they are significant as references to guide our lives. For example documents related to laws and administration and the one which is related to Holy Scripture.

The Mushaf of The Quran is an important document in the lives of Muslims in the world. Al-Quran was revealed by God to the Prophet Muhammad 1,400 years ago in the city of Mecca and then made into mushaf at the time of Caliph Uthman ibn Affan, is a document which was written in classical Arabic. It consists of 114 chapters (surah in Arabic), which are then divided into verses (ayah in Arabic).

Muslims in Indonesia with a population of 85% of the total population of Indonesia generally use Indonesian and regional languages in daily conversations. Although the majority, according to the BPS (Central Bureau of Statistics) data in 2015, there are 54% of the total population of Muslims in Indonesia are still unable to read the Quran, meaning, less than half that can read Al-Quran. Of which that can read Al-Quran, it is estimated that only a few understands the meaning of the reading of the Quran directly. This is due the fact that Arabic is not the mother tongue, and education system in Indonesia does not require Arabic lessons to be taught in school, therefore, they rely heavily on translation document of the Quran in the Indonesian language as a means to understand the meaning of the contents of the document in the Quran itself.

There are some variations of Quran translation in Indonesian, for example : word أَوْلِيَاء [*Aw-liyaa*] can be translated as *pemimpin* [leader], *pelindung* [protector], *penolong* [helper], *wali* [guardian] or *teman setia* [devoted friend]. How can we have a relatively equal level of understanding of the Quran just by using the translation version as if we read it in its original language? The more variations of meaning in translation will increase our understanding of

the words of the Quran. Translation version consists of sentences separated into surahs and subdivided into verses. How do we find the words in the translation is correlated with the words in the Quran?

Some of words in Quran eventhough there are the same words but they can be interpreted differently even by the same translator. In a sentence إِنَّكَ لَمِنَ الْمُرْسَلِينَ [*Innaka laminal mursaliin*] in Quran Surah Yasin (QS 36:3) it is translated into "*Sesungguhnya kamu salah seorang dari rasul-rasul*" [Indeed you, are from among the messengers], but in Quran Surah Al-Baqarah (QS 2:252) it is translated into "*Sesungguhnya kamu benar-benar salah seorang di antara nabi-nabi yang diutus*" [Indeed, you really one of the prophets who were sent]. So الْمُرْسَلِينَ has various meanings : *rasul-rasul* and *nabi-nabi yang diutus*.

Another example, word : رَبِّي in Surah Yusuf (QS 12: 23) means *Tuanku* [my master], but in Surah Ali Imran (QS 3:51) means *Tuhanku* [my Lord].

How we can separate the words of a verse in translation of the Quran with regard to the origin of the word in the Quran? In the Quran translated documents, in order to separate the phrase we should refer to the original document i.e Mushaf Al-the Quran in Arabic. In the context of the Al-Quran in Arabic, it is known that one word in Arabic can be translated into two or more words in Indonesian. For example Figure 1, the word فَجَعَلْنٰهُمْ (faja'alnahum) found in verse (23:41) is one word. When it is translated, it means: *dan Kami jadikan mereka* (And We made them) consists of four words in Indonesian.

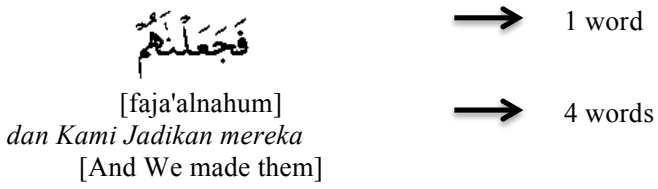

Figure 1 : Translation Process from Arabic to Indonesian

In this case we have to parse the words in the document into words or word-combinations which refer to its origin in Arabic. The process of decomposition and the determination of the words in these segments are called phrase segmentation.

A phrase in Indonesian grammar is a language unit consisting of two or more words that do not exceed the limits of functions. That is, the phrase as a whole is always inside a specific function, namely Subject, Predicate, Object etc.[2]

The meaning of the phrase in this case is a combination of more than one word in Indonesian formed from the translation of the Arabic used in the Quran.

1.2 Theoretical Framework

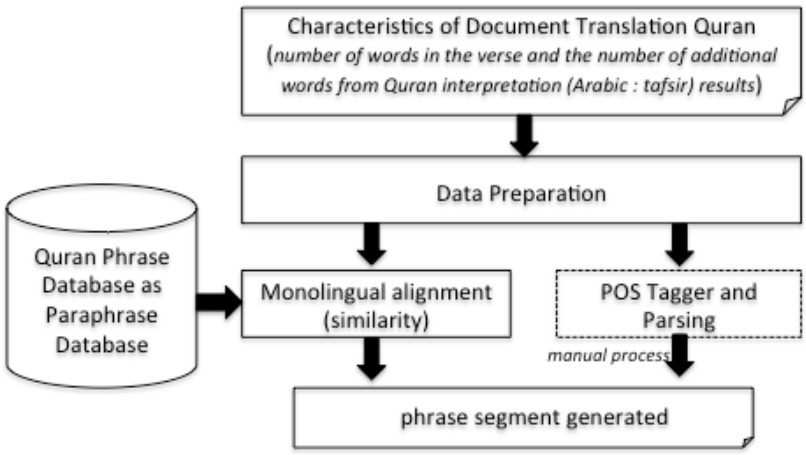

Figure 2 : Theoretical framework

Some of the theory in Figure 2, which can be used as reference for this study, can be described as follows:

a. Data preparation process, in this case tokenization process [3], aims to decipher the sentence in verse translation of the Quran into separate words to support the next process.

For Example: as Input text *Sesungguhnya Kami yang menghidupkan orang-orang mati* [Indeed We give live to the dead] in Surah Yasin verse 12. After tokenization process, it produces six tokens, which are (‘*Sesungguhnya’, ‘Kami’, ‘yang’, ‘menghidupkan’, ‘orang-orang’, ‘mati’*]

b. Quran Phrase Database is built just like Paraphrase Database [1] as the source of the referral process alignment that is built by word-by-word Quran translation documents and their paraphrases.
c. Monolingual Alignment as alignment process aims to align the words tokenization process results with a database of phrases Quran. In this case the alignment process is carried out using the method of similarity [4]

For Example : in Figure 3 , as a target text : tokens from Surah Yasin verse 32, : [‘*dan’, ‘setiap’, ‘mereka’, ‘semuanya’, ‘akan’, ‘dikumpulkan’, ‘lagi’, ‘kepada’, ‘kami’*]. We do alignment process by comparing with Quran Phrase Database as a source text.

Source Text : [‘*dan tidaklah’, ‘tiap-tiap’, ‘kecuali’, ‘semuanya’, ‘di hadapan kami’, ‘mereka dihadirkan’*] and paraphrases pairs *(‘tiap-tiap’,’setiap’*), (*‘mereka*

dihadirkan','mereka dikumpulkan'), *('di hadapan', 'kepada'*)

Target Text: [‘*dan', 'setiap', 'mereka', 'semuanya', 'akan', 'dikumpulkan', 'lagi', 'kepada', 'kami'*]

Figure 3 : Example Monolingual Alignment

d. Part-of-Speech (POS) Tagger [5] and Parsing Indonesian using Indonesian Treebank Bracketing Guidelines [6] as an alternative method for the segmentation of the phrase.. This process is carried out manually. The Example of POS Tagger and Parsing is explained in Page 23.

1.3 Conceptual Framework/Paradigm

The phrase segmentation process to document translation of the Quran conceptually can be done by defining a number of variables involved, namely:

a. Dependent Variable: Suitability of phrase segment generated
b. Independent Variable:
    1) Characteristics of Document Translation Quran (number of words in the verse and the number of additional words Quran interpretation (Arabic: *tafsir*) results)
    2) Performance of word tokenization process
    3) Completeness of Quran Phrase Database built
    4) Performance of alignment (similarity) process
    5) Performance of manual process of pos tags and syntax parser

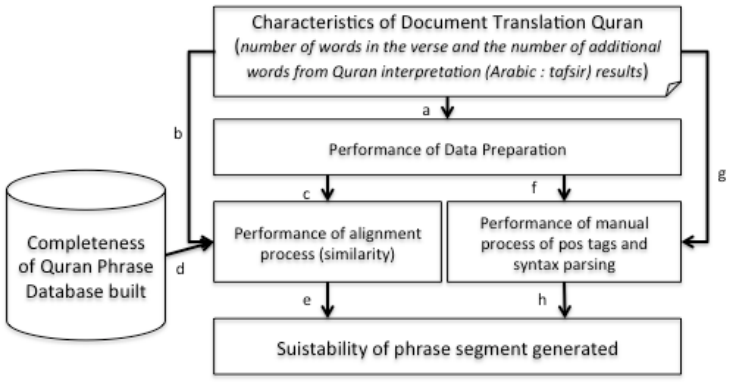

Figure 4 : Linkage between variables in conceptual framework

We describe the relationship between variables based on Figure 4 as follows:

a. *Characteristics of Document Translation Quran* and *Performance of Data Preparation process*

The number of words including punctuation in a sentence (verse) can affect the level of performance of the data preparation process. More words means the data preparation process will be longer and more complex, which led to the lower level of performance in the data preparation process.

b. *Characteristics of Document Translation Quran* and *Performance of alignment (similarity) process*.

The number of words in a verse can affect the iteration process for checking the similarity. More words can cause longer processing time.

The number of words additional results the Quran interpretation (usually made a mark in parentheses) are usually not accommodated in lateral translation of the word (the dictionary), thus reducing the number of words that can be aligned in the alignment process.

c. *Performance of data preparation process* and *Performance of alignment (similarity) process*

The number of words as a result of word tokenization as a part of data preparation process determines the number of iterations required for the alignment process.

d. *Completeness of Quran Phrase Database built* and *Performance of alignment (similarity) process*.

Completeness of the Quran built a database of phrases that serve as a reference for determining the alignment process the number of phrases that can be generated.

e. *Performance of alignment (similarity) process* and *Suistability of phrase segment generated*

   The number and shape of phrases that can be generated in the alignment process determine the suitability level of segmentation phrase when compared with the gold standard

f. *Characteristics of Document Translation Quran* and *Performance of manual process of pos tags and syntax parser*.

   The number of words in the verse in Quran translated documents affects the processing time and the accuracy of post tagging and compilation of format treebank syntax parsing.

g. *Performance of manual process of pos tags and syntax parser* and *Suistability of phrase segment generated*

   Results from Treebank format determine the process of identifying the appropriate phrase segment.

1.4 Statement of the Problem

According to the Rationale, Muslims in Indonesia rely heavily on document translation of the Qu'ran, so How can we have a relatively equal level of understanding of the Quran just by using the translation version as if we read it in its original language ?

Specific problem can be stated as follows:

a. How to do phrase-based segmentations based on Indonesian documents (in this case the document translations of the Quran)?
b. How to measure the degree of conformity of phrases segmentation?

1.5 Objective and Hypotheses

According to the problem statement, the objective of the research is to perform phrase-based segmentation in the Indonesian Documents, in this case the document is a translation of the Quran with case studies of Surah Yasin, by applying the method of Monolingual Alignment and Syntax Parsing in order to obtain the best level of conformity when compared with the gold standard.

This research is aimed to Indonesians who do not speak Arabic or understand the meaning the Quran directly so that they will utilize Indonesian translation of the document as a medium to explore and understand the Quran Document. They can take advantages of the resulting of phases segments to learn more about the meaning of the phrase and linkages with similar words or phrases / synonyms in other verses.

As a reference for the study, we present some hypotheses of this thesis as follows:

a. Phrase-based segmentation to Indonesian documents can be done through monolingual alignment with the level of suitability depends on the variations of the word in the verse and the completeness of the phrase Quran database with the level of suitability in F1Score almost more than 60% if it is compared with gold standard. For benchmark, here are F1Scores in other method and other case: Thadani et al. [7] : 84.1 %, Yao et al. [8]: 79.6 %, Yao et al. [9] : 79.9 %, Sultan [4] : 81.2 %, J. Ma [10]: 89.7%
b. Syntax parsing method can be an alternative of the phrase segmentation because it has considered Indonesian grammars and does not depend on completeness of the database of phrases Quran

1.6 Assumption

In this thesis, we make some assumptions as follows:

a. Manual process (POS tagging and document parsing Indonesian) has been carried out according the correct method, so the results can be used as input to the next phrase segmentation process.
b. Gold standard that we have created is assumed to be correct and can be used as a comparison to the suitability of the process results system.
c. Selection of surah Yasin as a sample case study is randomized, assuming the number of texts and variations of characteristic verse translation are sufficient.

1.7 Scope and limitation

The scopes of our study are as follows:

a. In accordance with the title of the thesis, we focus on phrase-based segmentation of the documentation Indonesian research, in this case the document translation of the Quran, with a case study on the Surah Yasin from verse 1-83.
b. Due to the manual process, sintax parsing method is limited only to fifteen verses in Surah Yasin with randomize verses.
c. Document translation of the Quran which is used, as a sample input document, is a

translation the Quran by Ministry of Religious Affairs downloaded through the website Tanzil – Quran Navigator[1].

1.8 Significance of the Study

Nowadays, study on phrases segmentation in Indonesian as one of the methods of natural language processing is still very rare, so we hope this research can make a positive contribution in the process of developing NLP (Natural Language Processing) in Indonesian generally and especially in the study of the Quran.

[1] Tanzil.net/trans (update date June 4, 2010)